Guía general de aprendizaje de

ADYN Herramienta de Normalización

Normalización de:

Direcciones postales, Nombres de personas e Identificadores de personas físicas y jurídicas

Instituto de Estadística de Andalucía
CONSEJERÍA DE ECONOMÍA, INNOVACIÓN Y CIENCIA

Unión Europea
FEDER

Índice de Contenidos:

1. Objetivo de la Guía.

El objetivo de esta Guía General es proporcionar al usuario de *ADYN Herramienta de Normalización* un manual completo para realizar de forma eficiente el proceso de normalización de direcciones postales, de nombres de personas e identificadores de personas físicas o jurídicas (NIF, DNI y NIE) e incidir en los diferentes aspectos que conlleva un proceso de normalización de datos probabilístico.

En esta Guía General se propondrá un ejemplo de normalización de direcciones postales basado en el fichero '*Ejemplo.csv*' que se adjunta con la aplicación en la carpeta 'C:\adyn\ejemplos\direcciones', de tal forma que cada paso del proceso de normalización quede lo suficientemente detallado. Este fichero consta de un único campo llamado 'Direcciones' que contiene las direcciones postales que deseamos normalizar y el número de registros que componen el fichero es 11.

A continuación mostramos el contenido de dicho fichero:

Ejemplo

	A
1	Direcciones
2	pl/ la solera nº 3- 1º d
3	pl/ del cabeceo nº 4- 4º c
4	pl/ grazalema bl-12 - 3º d
5	pl/ grazalema bl-12- 2º c
6	c/ atalaya nº 16 - 1º c
7	espiritu santo 2 bajo c
8	parque atlantico nº 4 8º b
9	calle la cartuja bq 12 piso 1 pta a
10	camino de vista alegre nº 2
11	paseo de las delicias nº 5 11º r
12	plza. andromeda bq 9-2º b

Imagen 1: Vista de los 11 registros del fichero '*Ejemplo.csv*'.

La guía también se acompaña de una serie de ejemplos relativos a la normalización de nombres de persona, incluyéndose en el **Anexo I** un caso práctico sobre cómo se lleva a cabo ésta utilizando el fichero '*Ejemplo_nombres.csv*'. Por otro lado, en lo que respecta a la normalización del campo identificador de persona física o jurídica, en el **Anexo II** se muestra detalladamente el objetivo y la forma de realizar dicho proceso haciendo uso del fichero '*Ejemplo_nif.csv*'. Tanto éste último como el fichero referido a nombres de personas se adjuntan en la aplicación junto con el de direcciones postales.

Notar que cuando trabajemos con ficheros de datos reales, el campo que deseamos normalizar, direcciones postales, nombres de personas o identificadores de personas físicas o jurídicas, no será el único campo existente en el fichero, sino que será un campo entre otros tantos que lo componen.

El objetivo de aplicar la normalización es obtener un nuevo fichero de datos que muestre el campo a normalizar segmentado en tantos campos como partes se han considerado que componen una dirección postal, un nombre de persona o un identificador de persona física o jurídica, con la particularidad de que estos nuevos campos ya no presentarán errores ni inconsistencias.

Por lo tanto, considerando nuestro ejemplo de direcciones postales, el resultado de normalizar el campo 'Direcciones' ha de ser similar a la siguiente imagen:

Ejemplo

	A	B	C	D	E	F	G	H	I	J
1	tipo_de_via	nombre_de_via	id_de_numero	numero	id_de_bloque	bloque	id_de_planta	planta	id_de_puerta	puerta
2	plaza	la solera	numero	3				1		d
3	plaza	del cabeceo	numero	4				4		c
4	plaza	grazalema			bloque	12		3		d
5	plaza	grazalema			bloque	12		2		c
6	calle	atalaya	numero	16				1		c
7		espiritu santo		2				bajo		c
8	parque	atlantico	numero	4				8		b
9	calle	la cartuja			bloque	12	piso	1	puerta	a
10	camino	de vista alegre	numero	2						
11	paseo	de las delicias	numero	5				11		r
12	plaza	andromeda			bloque	9		2		b

Imagen 2: Vista de los 11 registros normalizados del fichero '*Ejemplo.csv*'.

Para llegar a un nivel de normalización como el que se muestra en imagen hemos de realizar un proceso que resumido consiste en tomar una muestra de los registros del fichero '*Ejemplo.csv*' y a través de ella conocer la estructura o patrones de los datos contenidos en la muestra para construir un Modelo Oculto de Markov que nos permitirá extrapolar ese conocimiento al fichero original de datos. Finalmente usaremos este modelo para normalizar el campo en cuestión del fichero original de datos.

Por último hemos de indicar que en estas primeras versiones de *ADYN Herramienta de Normalización* hemos decidido trabajar con ficheros de texto separados por comas (CSV) ya que es uno de los formatos estándar más utilizados para el almacenamiento de datos. Para conocer más características de este formato el usuario puede consultar el apartado 'a)' y 'b)' del **Anexo III** de esta Guía General.

2. Proceso de normalización de un fichero de datos.

Normalmente, la mayoría de la información con la que nos encontramos en el mundo real contiene errores, está incompleta o incorrectamente formateada. Es por ello por lo que nos planteamos como objetivo transformar los datos originales brutos en otros datos con formatos consistentes y bien definidos, así como resolver las posibles inconsistencias sobre la forma en la que se representa y codifica la información.

El conjunto de técnicas encaminadas a la obtención de datos consistentes se engloban en el llamado *proceso de normalización de datos* y redundará en una mejor calidad y fiabilidad en posteriores análisis de esos datos.

En el proceso de normalización se establecen dos fases principales. Una primera fase de limpieza donde no importa el contenido semántico del fichero de datos, y se realizan tareas de codificación del fichero así como de eliminación de abreviaturas y signos de puntuación en los datos contenidos en él. La segunda fase es la de estandarización del conjunto de datos, en este caso se analiza el contenido semántico del fichero y se clasifica el contenido de este según el valor de sus componentes. Debido a esa clasificación se realizará la segmentación de los datos en cada una de las componentes que los forman.

El objeto de la aplicación informática *ADYN Herramienta de Normalización* es la limpieza, estandarización y segmentación de nombres de personas, de direcciones postales y de identificadores de personas físicas y jurídicas. Por ejemplo, en el caso de los nombres de personas, la normalización consistirá en limpiar, estandarizar y segmentar esos datos en nombres propios, apellidos y partículas auxiliares asociadas a ambos. De igual forma conseguiremos limpiar, estandarizar y segmentar las direcciones postales y los NIF, DNI, etc.

Para normalizar alguno de estos tres campos haremos uso de tres herramientas:

- Las listas de corrección: permiten limpiar el campo a normalizar del fichero de datos, es decir, contienen los caracteres que el usuario ha considerado oportuno eliminar o sustituir en dicho campo. Por ejemplo, se eliminan los caracteres extraños (‘|’, ‘$’,...) y se sustituyen las vocales con tildes por vocales sin tildes.

- Las tablas de búsqueda: sustituyen cada elemento del campo a normalizar por su valor normalizado y, además, le asignan una etiqueta. Por ejemplo, si en el campo a normalizar se encuentra el elemento ‘c/’ se sustituye por ‘calle’ y se le asigna la etiqueta ‘TV’ que significa Tipo de Vía.

- Los Modelos Ocultos de Markov (en inglés *Hidden Markov Models* ó HMM), tratan de reconocer el patrón de los datos que vamos a normalizar a través de una muestra de dichos datos. En este proceso extrapolaremos el conocimiento sobre esas estructuras o patrones que siguen los elementos de la muestra a la totalidad de los registros a normalizar, de tal forma que se consiga segmentar cada uno de los datos en los distintos elementos que los componen.

El proceso de segmentación también se puede realizar mediante técnicas basadas en reglas en lugar de usar Modelos Ocultos de Markov. Sin embargo y a pesar de que para nombres de personas los resultados de utilizar estas técnicas ofrecen resultados similares que al aplicar técnicas basadas en Modelos Ocultos de Markov, para direcciones postales y debido a la gran complejidad que muestran estos datos, sería muy complicado disponer de reglas que abarcasen la mayoría de los casos a segmentar; de esa forma el usuario tendría que crear una nueva regla conforme aparezca un caso nuevo que no haya sido tratado anteriormente. Usando los Modelos Ocultos de Markov para direcciones postales, el proceso se simplifica ya que a través de una muestra del campo a normalizar y estableciendo las estructuras y patrones de esos datos podremos estandarizar y segmentar todo el campo.

La siguiente figura nos muestra el esquema general del proceso de normalización:

Imagen 3: Proceso general de normalización.

Podemos ver como para el subproceso de limpieza necesitamos las listas de corrección y para el subproceso de estandarización y segmentación necesitamos las tablas de búsqueda y el Modelo Oculto de Markov.

ADYN Herramienta de Normalización incluye un conjunto de listas de corrección y tablas de búsqueda para nombres de personas, para direcciones postales y para identificadores de personas físicas y jurídicas. Estos ficheros podrán ser editados en su totalidad de tal forma que el usuario podrá ir enriqueciendo y personalizando la información contenida en ellos.

Para poder utilizar los Modelos Ocultos de Markov en el proceso de estandarización y segmentación, previamente han de ser generados. Para crear estos modelos, partiremos de

una muestra ya sea del fichero de datos original con el que vamos a trabajar o bien de otro fichero que tenga datos con una estructura similar a los que vamos a estandarizar y segmentar. Una vez obtenida esa muestra (operación automática) se realizan los procesos de etiquetado (operación automática), asignación de estados (operación manual) y entrenamiento (operación automática). Finalmente obtendremos un Modelo Oculto de Markov que será usado para normalizar el campo elegido del fichero original de datos.

El esquema de creación de Modelos Ocultos de Markov queda reflejado en la siguiente figura:

Imagen 4: Proceso de entrenamiento o de obtención del Modelo Oculto de Markov.

Y un esquema general de todos los procesos interconectados se muestra en la siguiente imagen:

Imagen 5: Esquema general del proceso de normalización.

3. Consideraciones previas a la normalización.

Antes de llevar a cabo el proceso de normalización con la aplicación informática *ADYN Herramienta de Normalización* es necesario analizar el contenido del fichero de datos con el que vamos a trabajar.

En primer lugar habrá que observar si el campo a normalizar contiene información separada por comas ',' y si es así habrá que eliminarlas. Por ejemplo, supongamos que vamos a normalizar el campo dirección postal o nombre de persona y tenemos información del tipo:

Direcciones	Nombres
Avda. Leonardo Da Vinci, nº 21	Ruíz Torre, Ana Francisca
C/ Doctor Fleming, 7	Rodríguez Miguel, Francisco
…	…

En estos casos habría que eliminar las comas para obtener un proceso de normalización eficiente.

En segundo lugar tendremos que observar si el campo a normalizar está registrado en el fichero de trabajo en una sola columna o por el contrario en varias. Si está en una sola columna no hay ningún problema para utilizar la aplicación pero si el campo a normalizar está segmentado en varias columnas hay que realizar un tratamiento especial.

Por ejemplo, supongamos que queremos normalizar el campo nombre de persona y los datos contenidos en éste aparecen registrados en el fichero de trabajo de la siguiente forma:

Nombre	Apellido1	Apellido2
Ana Francisca	Ruíz	Torre
Francisco	Rodríguez	Miguel
…	…	…

Como la aplicación no permite normalizar varios campos a la vez, para llevar a cabo la normalización del campo nombre de persona podemos proceder de dos formas:

- **Unir la información en un solo campo**, es decir, agregar la información de los campos 'Nombre', 'Apellido1' y 'Apellido2' en uno solo y normalizar dicho campo.

- **Normalizar cada campo por separado.** El usuario puede elegir el orden que desee para empezar a normalizar los campos. Por ejemplo, si decidimos normalizar en primer lugar el campo 'Nombre', a continuación, partiendo de este fichero con el nombre normalizado, normalizaríamos el campo 'Apellido1' y por último, partiendo del fichero con el nombre y el primer apellido normalizado, normalizaríamos el campo 'Apellido2'. Aunque este procedimiento es sencillo hay que tener en cuenta una serie de consideraciones que se explican más detenidamente en el **Anexo I**.

4. ADYN Herramienta de Normalización.

Con esta aplicación conseguiremos realizar de forma sencilla un proceso de normalización de nombres de personas, de direcciones postales y de identificadores de personas físicas y jurídicas de tal forma que a partir del conocimiento de la estructura o patrón que presentan los datos contenidos en una muestra podamos normalizar la totalidad del fichero de datos.

Es preciso aclarar que la aplicación *ADYN* no permite realizar el proceso de normalización de nombres de personas, de direcciones postales y de identificadores de personas físicas y jurídicas a la vez. Es decir, si se quiere normalizar el campo nombre y el campo dirección postal de un fichero de trabajo, primero se tendrá que normalizar uno y posteriormente el otro.

El proceso de normalización de datos a través de *ADYN Herramienta de Normalización* consta de las siguientes fases:

- **4.1 Creación del Modelo Oculto de Markov.**
- **4.2 Normalización de datos.**
- **4.3 Validación del proceso de normalización.**

Veámoslas detenidamente.

4.1 Creación del Modelo Oculto de Markov.

Los Modelos Ocultos de Markov reconocen ciertos patrones de comportamiento que siguen los datos contenidos en nuestro fichero, permitiéndonos estandarizar y segmentar dichos datos.

Por ejemplo si tenemos la dirección postal ‘C/ Jorge Morales 26’ el modelo reconocerá, estandarizará y segmentará el patrón de la siguiente manera:

Valor a normalizar:	**C/ Jorge Morales 26**		
Patrón:	Tipo de Vía	Nombre de Vía	Número
Estandarización y segmentación:	Calle	Jorge Morales	26

Es decir, ‘C/’ lo reconoce como tipo de vía y lo estandariza por ‘Calle’, ‘Jorge Morales’ lo reconoce como nombre de vía y lo estandariza por el mismo valor ya que no contiene ningún error y ‘26’ lo reconoce como número y lo estandariza por el mismo valor por la misma razón anterior.

De igual forma, si trabajamos con nombres de personas y tenemos estructuras de datos del tipo 'Ruíz Torre Ana Francisca' el modelo reconocerá, estandarizará y segmentará estas estructuras de la siguiente forma:

Valor a normalizar:	**Ruíz Torre Ana Francisca**			
Patrón:	Apellido1	Apellido2	Nombre1	Nombre2
Estandarización y segmentación:	Ruiz	Torre	Ana	Francisca

Finalmente, para el caso de identificadores de personas físicas y jurídicas tendremos datos del tipo 'A1245218-1' o similar. El modelo reconocerá, estandarizará y segmentará estos datos de la forma:

Valor a normalizar:	**A1245218-1**		
Patrón:	Letra de inicio	Número de identificación	Carácter de control
Estandarización y segmentación:	a	1245218	1

Para crear el Modelo Oculto de Markov seguiremos los siguientes tres pasos:

- Paso 1: Selección y etiquetado de la muestra.
- Paso 2: Asignación manual de estados.
- Paso 3: Entrenamiento de la muestra.

Hemos de incidir en el hecho de que para normalizar el campo identificador de persona física o jurídica **no es necesario construir** dicho modelo, como sí habría que hacerlo en el caso de nombres de personas y direcciones postales. El motivo se debe a que con la aplicación se suministra un modelo útil que se considera representa bastante bien toda la casuística que sobre esta materia podemos encontrar. Exactamente, el modelo se encuentra ubicado en el directorio *'adyn\ejemplos\idpersonas\modelo_propuesto\modelo HMM_idpersona.hmm'*.

A continuación se explica detenidamente cada paso del proceso de construcción del Modelo Oculto de Markov.

4.1.1 Paso 1: Selección y etiquetado de la muestra.

A partir del fichero de trabajo, la aplicación selecciona una muestra aleatoria con reposición del campo que deseamos normalizar, ya sea nombres de personas o direcciones postales. El usuario será quien especifique el tamaño de la muestra a seleccionar. En ella, se eliminan los símbolos o caracteres extraños mediante las 'listas de corrección' y se etiquetan los elementos que la componen mediante el uso de las 'tablas de búsqueda'.

El proceso de etiquetado consiste en lo siguiente: la aplicación irá buscando cada uno de los elementos en las 'tablas de búsqueda' de tal forma que si lo encuentra, le asignará la etiqueta correspondiente a esa tabla. En el caso de que un elemento no aparezca en las tablas de búsqueda será etiquetado como 'UN' (*unknown*, desconocido en inglés).

Por ejemplo, supongamos un fichero de datos en el que deseamos normalizar el campo nombre de persona. Si el usuario elige tomar una muestra de tres registros, uno de los posibles resultados que la aplicación ofrece de forma automática es el siguiente:

```
# 565 (0): |ruíz torre ana francisca|
#          |ruiz torre ana francisca|
           UN:, UN:, NF:, NF:

# 594 (1): |rodríguez miguel francisco|
#          |rodriguez miguel francisco|
           UN:, NM:, NM:

# 640 (2): |paz gonzález rodríguez|
#          |paz gonzalez rodriguez|
           NF:, UN:, UN:
```

Para cada uno de estos registros tenemos la siguiente información:

- La primera línea nos indica:
  - #: toda la información que va detrás de ella es un simple comentario y sirve de información adicional al usuario pero no es utilizada en ningún proceso.
  - A continuación se muestra la línea del fichero original en la que se encuentra el registro. En el ejemplo: 565, 594 y 640.
  - Al tomar la muestra los registros se enumeran automáticamente comenzando por 0. De esta forma, entre paréntesis aparece el número que tiene asignado el registro en la muestra. En el ejemplo, vemos que el primer registro tiene asignado el número (0), el segundo el (1) y el tercero el (2).
  - Por último la información original contenida en el campo a normalizar. En nuestro caso: |ruíz torre ana francisca|, |rodríguez miguel francisco| y |paz gonzález rodríguez|.

- En la segunda línea se muestra el valor del campo tras las correcciones definidas en las 'listas de corrección' y las 'tablas de búsqueda'.

- La tercera línea muestra las etiquetas asignadas a cada componente del campo a normalizar. Para el primer registro de la muestra las etiquetas asignadas son UN:, UN:, NF:, NF:. Esto quiere decir que:

    - 'ruiz' y 'torre' han sido etiquetados como 'UN' (desconocidos) ya que son elementos que no se han encontrado en ninguna tabla de búsqueda.

    - y los elementos 'ana' y 'francisca' han sido etiquetados como 'NF' porque se han encontrado en la tabla de búsqueda de nombres femeninos (knombres_femeninos.tbl).

Si hubiésemos trabajado con direcciones postales, el procedimiento de selección de la muestra y etiquetado, hubiera sido análogo al de nombres con la restricción de que las etiquetas serían las definidas para direcciones.

En el **Anexo IV** se pueden consultar todas las etiquetas definidas para nombres de personas, direcciones postales, e identificadores de personas físicas y jurídicas.

Para realizar este proceso de selección y etiquetado de direcciones postales o de nombres de personas mediante la herramienta *ADYN* usaremos la interfaz ***'02. Selección de la muestra'***. Accedemos a ella a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) tal y como se muestra en la siguiente imagen:

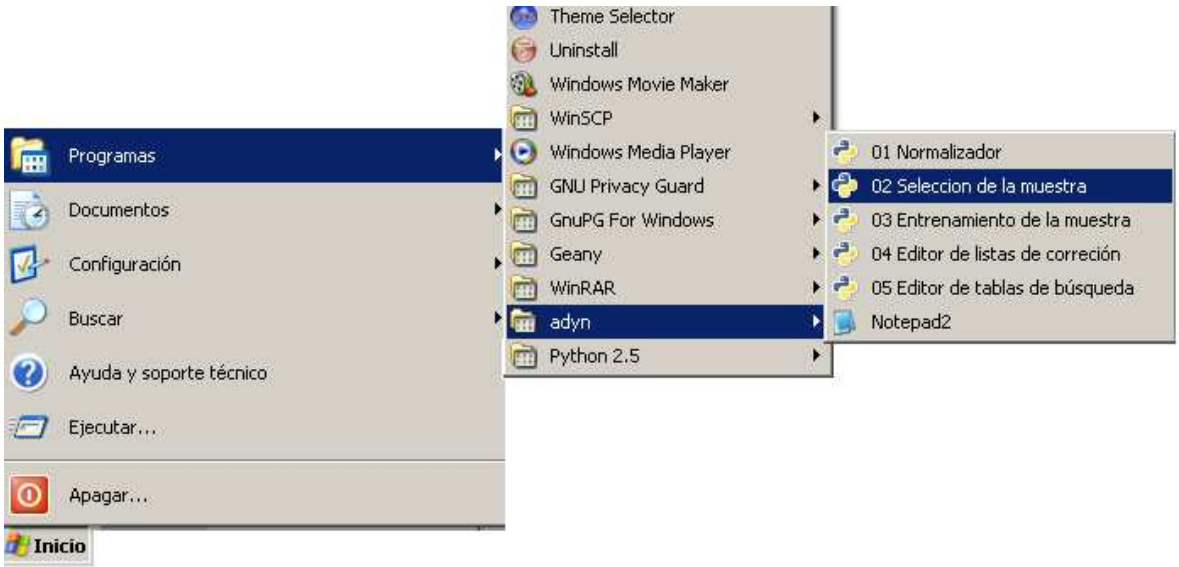


Imagen 6: Acceso a la interfaz '02. Selección de la muestra'.

Y explicaremos cómo funciona utilizando el fichero '*Ejemplo.csv*' de direcciones postales.

Una vez abierta la interfaz, nos recibirá la siguiente pantalla:

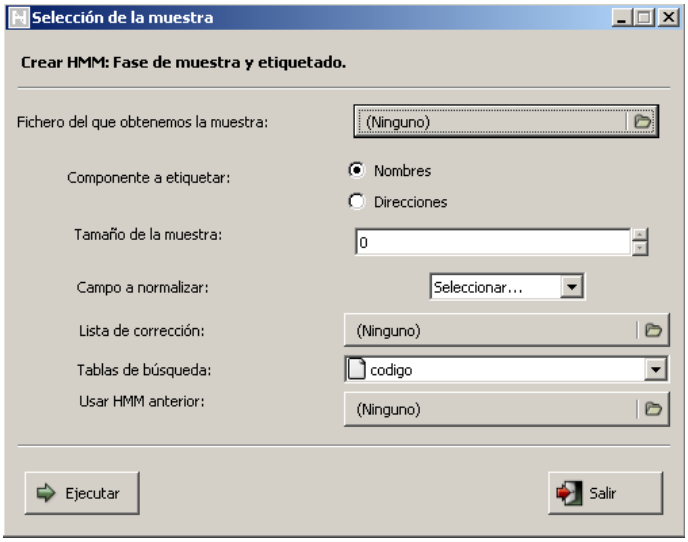


Imagen 7: Interfaz de selección de la muestra y etiquetado de componentes.

Esta pantalla será análoga tanto para el caso de seleccionar una muestra y realizar el etiquetado de nombres de personas como de direcciones postales, ya que lo único que variará será el tipo de '**Componente a etiquetar**' elegida.

Lo primero que debemos seleccionar es el fichero original de trabajo del que vamos a obtener la muestra, '*Ejemplo.csv*'. Para ello haremos 'click' en el botón correspondiente al **'Fichero del que obtenemos la muestra'** y obtendremos un navegador de archivos como el de la siguiente imagen:

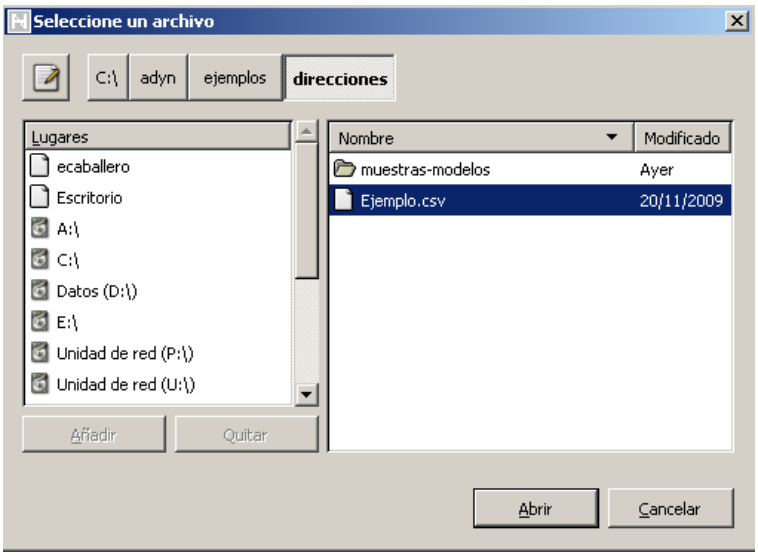


Imagen 8: Selección del fichero del que obtenemos la muestra.

En este navegador seleccionamos el fichero *'Ejemplo.csv'* ubicado en 'C:\adyn\ejemplos\direcciones' y hacemos click en 'Abrir' (o simplemente doble click sobre el nombre del fichero) y quedará seleccionado en nuestra interfaz.

A continuación marcaremos la **'Componente a etiquetar'** que como su nombre indica hace referencia a la componente que vamos a etiquetar para posteriormente normalizar: nombres de personas (Nombres) o direcciones postales (Direcciones); en nuestro caso, 'Direcciones'.

El siguiente paso es seleccionar el **'Tamaño de la muestra'**. El valor por defecto en la aplicación es 1 pero podríamos indicar cualquier otro valor teniendo en cuenta que como máximo el tamaño de la muestra será igual al tamaño del fichero de datos menos uno. El valor óptimo a seleccionar depende de lo heterogéneos que sean nuestros datos, es decir, a mayor heterogeneidad mayor tiene que ser el tamaño de muestra tomado. En nuestro caso tomaremos, por ejemplo, un tamaño de muestra igual a 4, con lo cual se seleccionarán 4 registros.

En el cuadro combinado **'Campo a Estandarizar'** se listan todos los campos que tiene el fichero *'Ejemplo.csv'* y seleccionaremos el campo que contiene la dirección postal que vamos a normalizar. En nuestro caso será el campo '*direcciones*'.

La interfaz está quedando configurada de la siguiente forma:

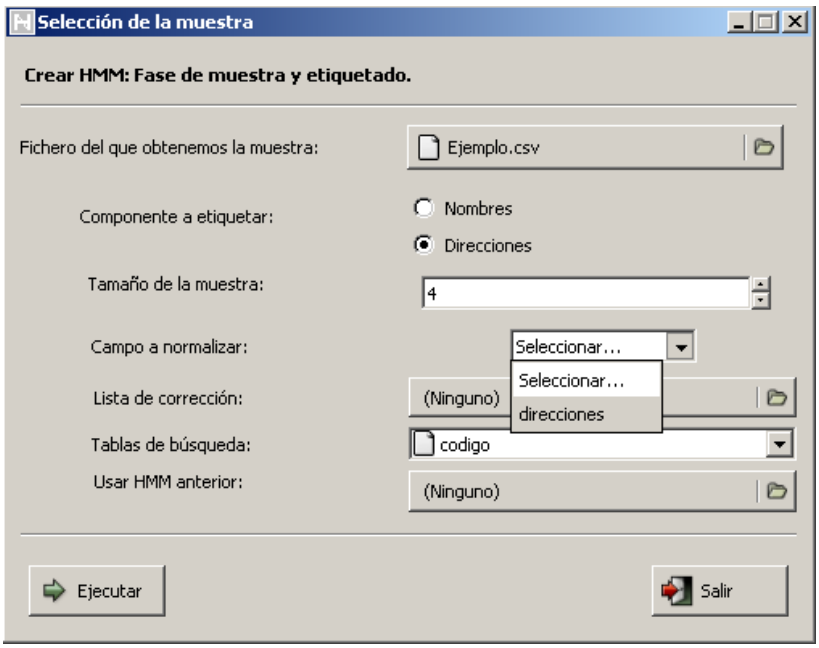


Imagen 9: Selección del campo a normalizar.

A continuación, la interfaz solicita la '**Lista de corrección'**. Para ello abrimos el cuadro de diálogo y seleccionamos '*direcciones_correccion.lst'* que se encuentra dentro de la carpeta

'ListasDeCorreccion' en la ruta 'C:\adyn\codigo\datos\ListasDeCorreccion'.

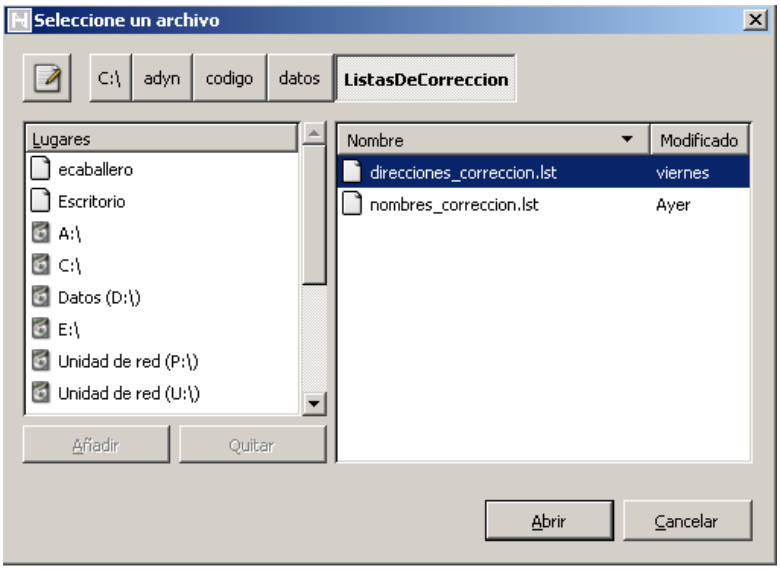


Imagen 10: Selección de lista de corrección de direcciones postales.

Otro de los elementos solicitados son las **'Tablas de búsqueda'**. En el desplegable debemos seleccionar la opción 'Otro' y buscar la carpeta '*direccion-tbl*' donde se encuentran las tablas de búsqueda para direcciones postales. La ubicación de esta carpeta es 'C:\adyn\codigo\datos'. Al seleccionar esta carpeta quedarán seleccionadas automáticamente todas las tablas de búsqueda.

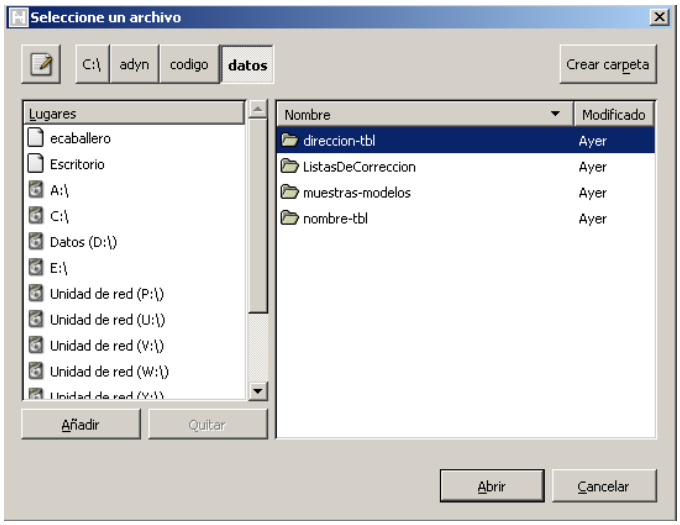


Imagen 11: Selección de las tablas de búsqueda de direcciones postales.

En el **Anexo V** se ofrece información más detallada sobre estas dos herramientas.

Por último, en la interfaz, encontramos la opción **'Usar HMM anterior'** que nos permite utilizar un Modelo Oculto de Markov creado con anterioridad a partir de otro fichero de datos que tiene una estructura similar al nuestro. En el **Anexo VI** se puede consultar más información sobre los Modelos Ocultos de Markov y además se explican más detalladamente las ventajas de utilizar un modelo HMM creado previamente.

Como en nuestro caso no tenemos ningún modelo creado, no introduciremos ningún fichero.

La siguiente imagen muestra como queda definida la interfaz ***'02. Selección de la muestra'*:**

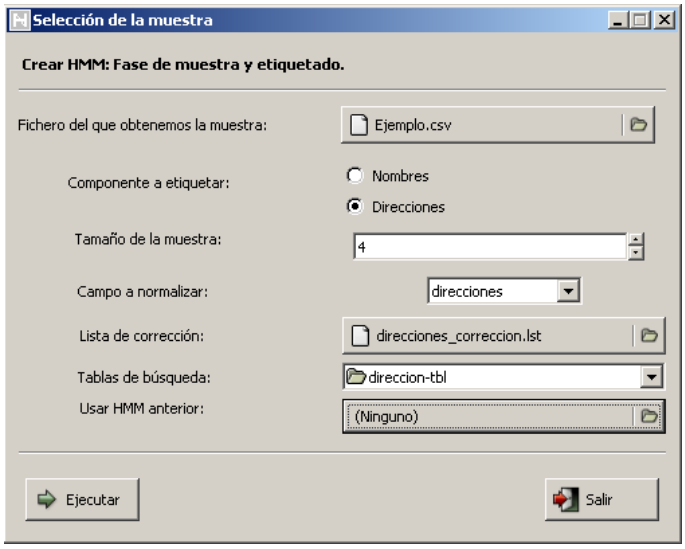


Imagen 12: Interfaz de selección de la muestra.

Hacemos click sobre el botón '**Ejecutar**' y, cuando el proceso termine, nos aparecerá la siguiente pantalla de información:

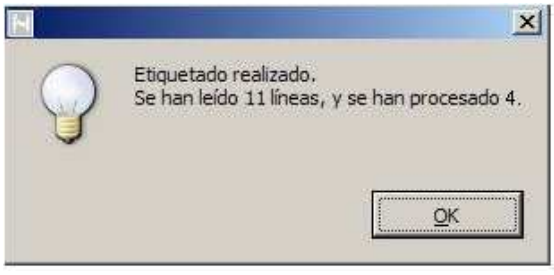


Imagen 13: Finalización del proceso de selección de la muestra.

En ella se indica el número de líneas que se han leído aleatoriamente del campo a normalizar y dentro de ellas las que se han procesado, es decir las que se han elegido para formar parte de la muestra. Así pues, el número de líneas leído va a ser igual al tamaño del fichero de datos o inferior a éste pero nunca inferior al número de líneas procesado, es decir al tamaño de la muestra.

En el ejemplo el número de líneas que se han leído coincide con el tamaño del fichero de datos original (11) y las que se han procesado son cuatro que es el tamaño de la muestra elegida.

En el caso de que hayamos olvidado especificar algún parámetro de la interfaz, al pulsar '**Ejecutar**' aparecerá un mensaje advirtiéndonos del error, consiguiendo así que el proceso de selección de muestra y etiquetado se realice con el éxito esperado.

Como resultado de este proceso se genera un fichero con la muestra etiquetada que se guardará automáticamente en la misma carpeta que el fichero de datos origen *'Ejemplo.csv'.* El nombre del fichero tendrá la forma:

'*muestra_etiquetada_<fecha_creación>-<hora_creación>_<fichero_origen>.csv*'

Por ejemplo, si tenemos el fichero '*muestra_etiquetada_20091030-1241_Ejemplo.csv*', sabremos que la muestra fue creada el día 30 de Octubre de 2009 a las 12:41 a partir del fichero *'Ejemplo.csv'.*

Aunque, por defecto, la aplicación guarda el fichero con este nombre, el usuario puede decidir darle otro nombre, siempre con extensión '*.csv*'.

El contenido del fichero será el siguiente:

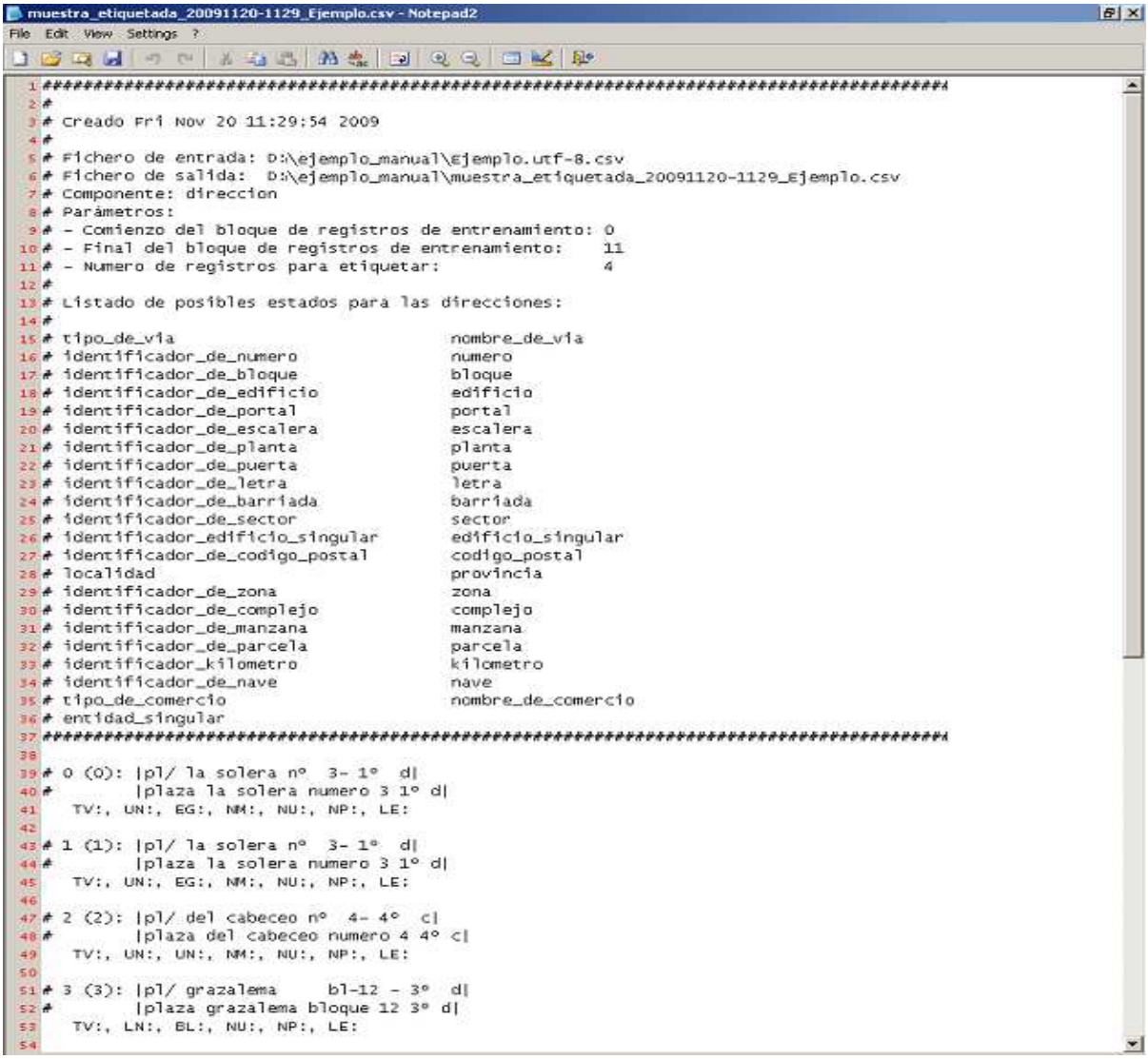


Imagen 14: Fichero resultante de la interfaz 02: componentes de la muestra etiquetadas.

La primera parte del fichero está rodeada de almohadillas '#' por tratarse de comentarios que consideramos necesarios conozca el usuario pero dicha información no será leída por la aplicación.

En concreto, se muestra información de la siguiente naturaleza:

- Fecha de creación del fichero.
- Fichero de origen: se indica su ruta de origen. Notar que se ha realizado automáticamente un cambio de codificación a UTF-8 para subsanar posibles problemas que se pueden presentar con la codificación de caracteres.
- Fichero de salida: se indica la ruta donde se encuentra almacenado este fichero.
- Componente: hace referencia a la componente que hemos etiquetado, en nuestro caso, Direcciones.

- Parámetros de selección: nos indica el número de registros del fichero original de trabajo (11 registros) y el tamaño de muestra seleccionado (4 registros).
- Listado de posibles estados para las Direcciones: en dos columnas tenemos la lista de posibles estados (en estos momentos se han definido 45) que se pueden asignar a cada una de las etiquetas. Como ya se ha comentado en párrafos anteriores, la definición de cada una de estas etiquetas y estados referidos a Direcciones se puede consultar en el **Anexo IV**.

A continuación, se presenta el detalle de los registros que conforman la muestra con sus elementos etiquetados. La información que aparece para cada registro es similar a la que explicamos, al comienzo de este apartado, para nombres de personas.

```
# 0 (0): |pl/ la solera nº  3- 1º  d|
#        |plaza la solera numero 3 1º d|
   TV:, UN:, EG:, NM:, NU:, NP:, LE:

# 1 (1): |pl/ la solera nº  3- 1º  d|
#        |plaza la solera numero 3 1º d|
   TV:, UN:, EG:, NM:, NU:, NP:, LE:

# 2 (2): |pl/ del cabeceo nº  4- 4º  c|
#        |plaza del cabeceo numero 4 4º c|
   TV:, UN:, UN:, NM:, NU:, NP:, LE:

# 3 (3): |pl/ grazalema     bl-12 - 3º  d|
#        |plaza grazalema bloque 12 3º d|
   TV:, LN:, BL:, NU:, NP:, LE:
```

Imagen 15: Detalle de la muestra etiquetada.

Notar que en la muestra aparecen dos registros duplicados (0 y 1) por haber utilizado un muestreo aleatorio simple con reposición.

4.1.2 Paso 2: Asignación manual de estados.

Esta fase será **siempre manual** y requerirá intervención del usuario para asociar a cada etiqueta, del fichero de la muestra etiquetada, su estado correspondiente. Por estado, entendemos el identificador de cada uno de los elementos del campo que vamos a normalizar.

Así pues, para realizar esta asignación de estados editaremos el fichero de salida del paso anterior, '*muestra_etiquetada_20091030-1241_Ejemplo.csv*' con el editor de texto '**Notepad2**' que suministramos junto con la aplicación. Utilizar este editor permite que la codificación de los ficheros con los que trabajamos sea la correcta (UTF-8) y de esta forma se

evita la inserción de caracteres propios de otras codificaciones. Accedemos al editor a través del menú Inicio / adyn / Notepad2 o a través de la ruta 'C:\adyn\notepad2'.

Una vez abierto el fichero '*muestra_etiquetada_20091030-1241_Ejemplo.csv*', nos disponemos a asignar manualmente el estado a cada etiqueta.

Por ejemplo para el primer registro:

```
# 0 (0): |pl/ la solera nº 3-1º d|

#       |plaza la solera numero 3 1º d|

  TV:, UN:, EG:, NM: NU: NP: LE:
```

Procederíamos de la siguiente forma:

- 'pl' se ha etiquetado por la aplicación como TV (tipo de vía) y le asignamos el estado 'tipo_de_via'.
- 'la' se etiqueta por la aplicación como UN (*unknown*, desconocido) y le asignamos el estado 'nombre_de_via' ya que entendemos que forma parte del nombre de la vía.
- 'solera' se etiqueta por la aplicación como EG (entidad singular) y le asignamos el estado 'nombre_de_via' por la misma razón anterior.
- 'nº' se etiqueta por la aplicación como NM (identificador de número) y le asignamos el estado 'identificador_de_numero'.
- '3' se ha etiquetado por la aplicación como NU (número) y le asignamos el estado 'numero'.
- '1º' se ha etiquetado por la aplicación como NP (número de planta) y le asignamos el estado 'planta'.
- 'd' se ha etiquetado por la aplicación como LE (letra) y le asignamos el estado 'puerta'.

Esta asignación de estados se ha realizado bajo el conocimiento que el usuario tiene sobre los campos que pueden componer una dirección postal y según su criterio. Ello supone que otro usuario puede realizar una asignación diferente.

Notar que el nombre de los estados no respeta los signos de puntuación, es decir, los estados se escriben sin tilde.

Tras la asignación, el registro queda de la siguiente forma:

```
# 0 (0): |pl/ la solera nº 3-1º d|

#       |plaza la solera numero 3 1º d|

  TV:tipo_de_via, UN:nombre_de_via, EG:nombre_de_via, NM:identificador_de_numero, NU:numero,   NP:planta, LE:puerta
```

Si repetimos este proceso con todos los registros del fichero, el resultado es el siguiente:

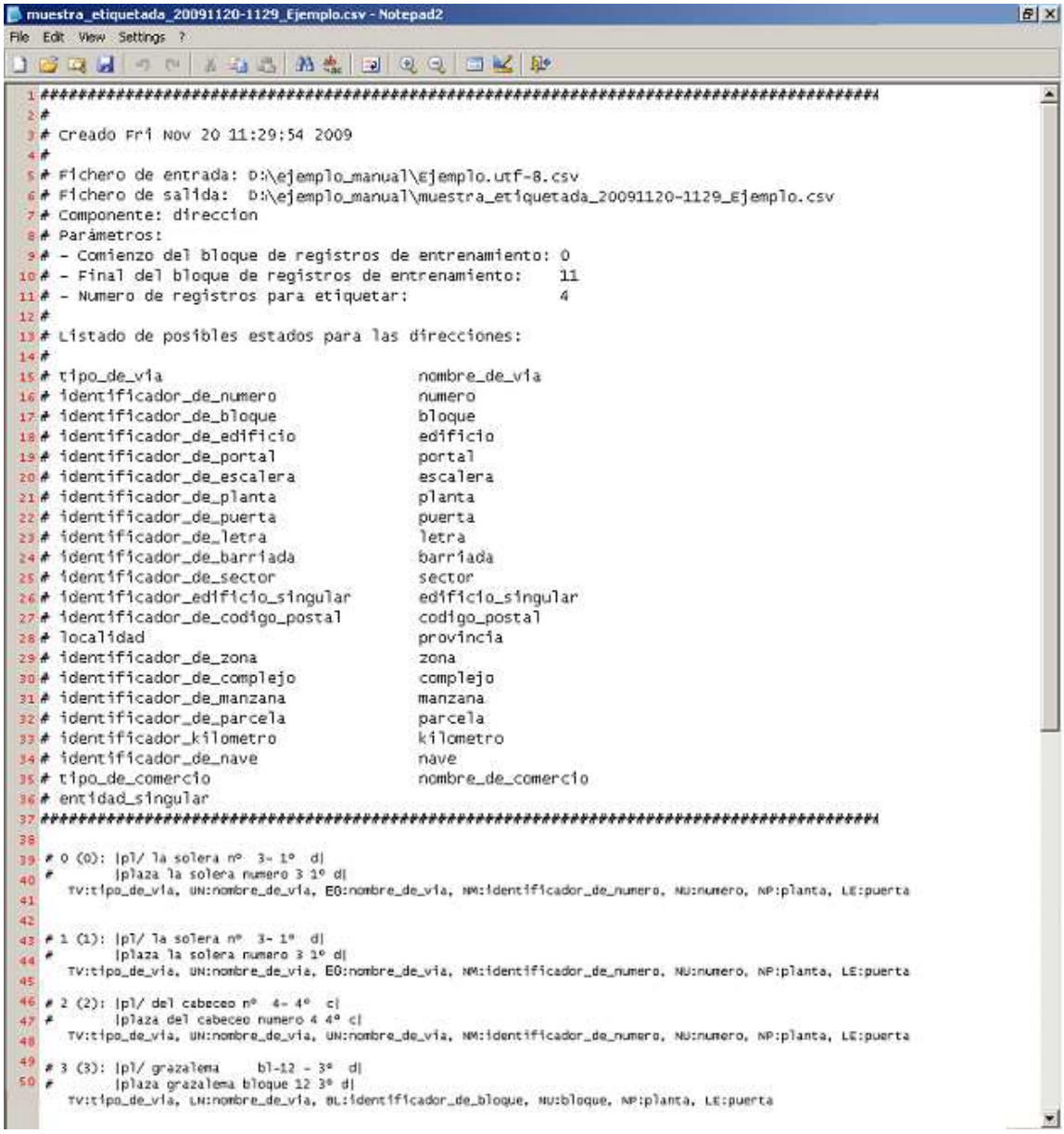


Imagen 16: Fichero con las componentes de la muestra etiquetadas y con estados asignados.

Hay que notar, que el usuario podrá eliminar las estructuras de los registros que considere innecesarias, es decir, si el usuario considera que quiere tener un fichero de muestra donde solamente existan estructuras de datos distintas puede decidir quedarse con una de ellas y eliminar el resto. En este caso, por ejemplo, se podría eliminar el registro identificado por 0 y quedarnos con el identificado por 1, o viceversa.

Antes de cerrar el fichero guardaremos los cambios. Aunque el usuario puede decidir el nombre con el que quiere guardarlo (siempre con extensión '*.csv*'), para este ejemplo se ha decidido hacerlo con el mismo nombre.

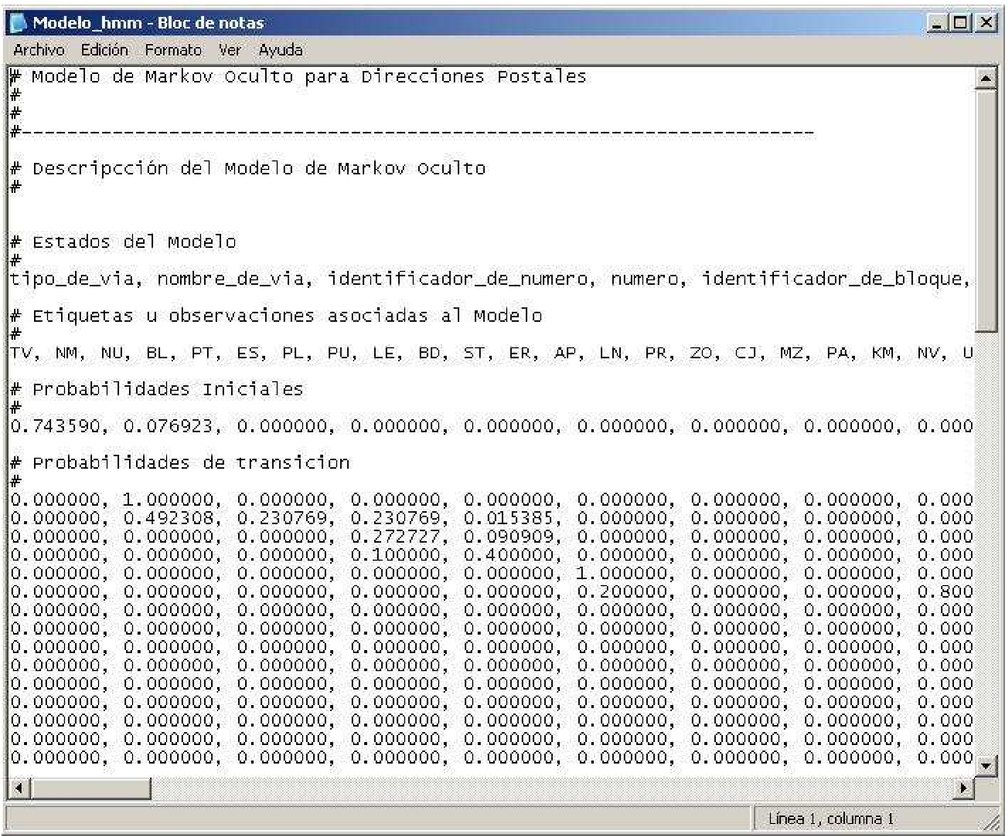
4.1.3 Paso 3: Entrenamiento de la muestra.

El resultado del Entrenamiento de la muestra será la creación del Modelo Oculto de Markov.

En este Paso usamos el fichero resultante del 'Paso 2', que contiene la muestra etiquetada y en la que hemos asignado los estados correspondientes, '*muestra_etiquetada_20091030-1241_Ejemplo.csv*'. A través de esta información, la aplicación genera:

- Un vector de probabilidades iniciales que nos indica la probabilidad de que la dirección postal (siguiendo con nuestro ejemplo) comience por cada uno de los estados.
- Una matriz de probabilidades de transición entre estados. Esta matriz nos indicará la probabilidad de pasar de un estado a otro según la muestra que hemos etiquetado y asociado estados previamente.
- Una matriz de probabilidades de observación (o etiquetas), es decir, muestra la probabilidad de que una etiqueta tenga asociado un estado determinado.

Estas tres matrices conformarán el llamado Modelo Oculto de Markov y quedan recogidas en un fichero de texto con extensión '*.hmm*' como el de la siguiente imagen. Además, aparecen el conjunto de etiquetas y estados ordenados según los elementos que se ha decidido que componen una dirección postal o un nombre de persona, para su correcta interpretación.




```
# Modelo de Markov Oculto para Direcciones Postales
#
#
#----------------------------------------------------------------------

# Descripcción del Modelo de Markov Oculto
#


# Estados del Modelo
#
tipo_de_via, nombre_de_via, identificador_de_numero, numero, identificador_de_bloque,

# Etiquetas u observaciones asociadas al Modelo
#
TV, NM, NU, BL, PT, ES, PL, PU, LE, BD, ST, ER, AP, LN, PR, ZO, CJ, MZ, PA, KM, NV, U

# Probabilidades Iniciales
#
0.743590, 0.076923, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000

# Probabilidades de transicion
#
0.000000, 1.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.492308, 0.230769, 0.230769, 0.015385, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.272727, 0.090909, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.100000, 0.400000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 1.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.200000, 0.000000, 0.000000, 0.800
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
```

Imagen 17: Detalle del fichero que contiene un HMM para direcciones postales.

A continuación vemos detenidamente que información contienen las matrices que conforman el Modelo Oculto de Markov:

El **vector de probabilidades iniciales** aparece en primer lugar. En nuestro ejemplo la probabilidad de que la dirección postal comience por el tipo de vía es de 0,743590 y de que comience por el nombre de la vía es de 0,076923. La probabilidad de que la dirección postal comience por un estado diferente es 0. El detalle del vector de probabilidades iniciales se recoge en la siguiente imagen:

```
# Estados del Modelo
#
tipo_de_via, nombre_de_via, identificador_de_numero, numero, identificador_de_bloque,
# Etiquetas u observaciones asociadas al Modelo
#
TV, NM, NU, BL, PT, ES, PL, PU, LE, BD, ST, ER, AP, LN, PR, ZO, CJ, MZ, PA, KM, NV, U
# Probabilidades Iniciales
#
0.743590, 0.076923, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
```

Imagen 18: Detalle de las probabilidades iniciales para un HMM de direcciones postales.

La **matriz de probabilidades de transición** aparece en segundo lugar. Es una matriz cuadrada, donde el número de filas y columnas es el número de estados. Esta matriz nos indicará la probabilidad de pasar de un estado a otro según la muestra que hemos etiquetado y asociado estados previamente. Estos estados, así como su orden, aparecen en el fichero. Por ejemplo, la primera columna y la primera fila corresponden al estado ‘tipo_de_via’; la segunda columna y la segunda fila corresponden al estado ‘nombre_de_via’. Ello nos permite hacer los cruces para interpretar las probabilidades de transición entre estados.

Interpretando la matriz del ejemplo, tenemos que:

- la probabilidad de pasar del estado ‘tipo_de via’ al estado ‘nombre_de_via’ es uno. Esto es, la probabilidad de que después de ‘calle’ aparezca el ‘nombre de la calle’ o que después de ‘avenida’ aparezca el ‘nombre de la avenida’ es uno.
- la probabilidad de pasar del estado ‘nombre_de_via’ al estado ‘nombre_de_via’ es de 0,492308. Esta situación se presentará en aquellos casos en los que el nombre de la dirección postal sea compuesto, es decir, del tipo ‘Jorge Morales’.

```
0.743590, 0.076923, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000

# Probabilidades de transicion
#
0.000000, 1.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.492308, 0.230769, 0.230769, 0.015385, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.272727, 0.090909, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.100000, 0.400000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 1.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.200000, 0.000000, 0.000000, 0.800
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000

Línea 1, columna 1
```

Imagen 19: Detalle de la matriz de probabilidades de transición para un HMM de direcciones postales.

La **matriz de probabilidades de observación (o etiquetas)** es una matriz que tiene tantas columnas como etiquetas y tantas filas como número de estados.

Cada elemento de la matriz nos indica la probabilidad de que una etiqueta tenga asociado un estado determinado. Así pues, como se puede ver en la siguiente imagen, tenemos que la probabilidad de que la etiqueta 'TV' tenga asociado el estado 'tipo_de_via' es 0,980769 y la de que tenga asociado el estado 'nombre_de_via' es 0,019275.

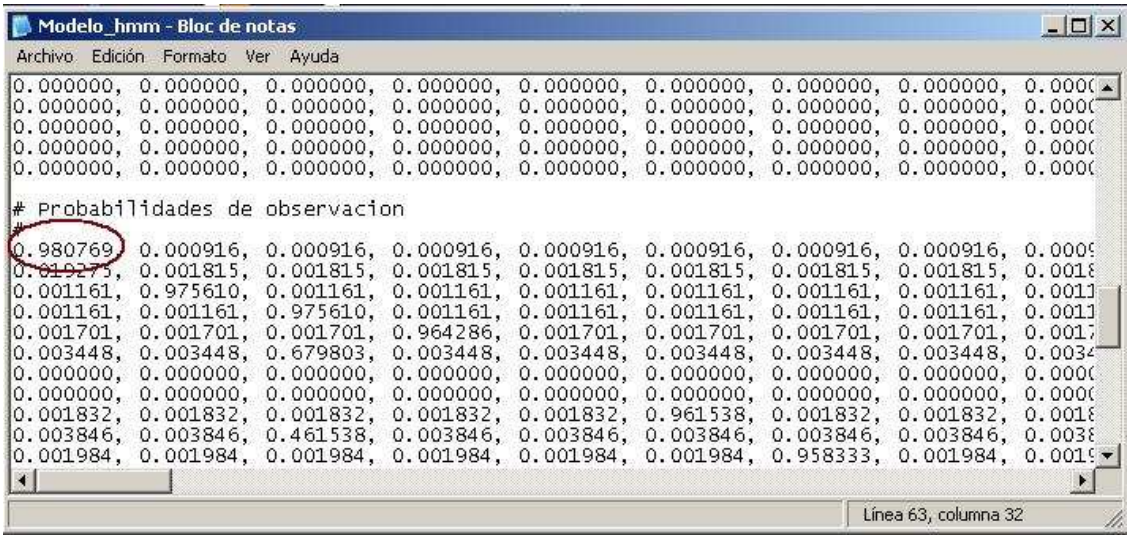


Imagen 20: Detalle de la matriz de probabilidades de observación para un HMM de direcciones postales.

Una vez obtenido el modelo solo nos falta saber cuál es la probabilidad de las secuencias de etiquetas y estados asociados en la muestra. Para encontrar esta probabilidad se utiliza el *algoritmo de Viterbi*.

Vemos a continuación cómo funciona este algoritmo con un ejemplo de dirección postal que tiene asignados dos secuencias.

Cuando la aplicación realiza el proceso de etiquetado es posible que un mismo elemento de la dirección postal tenga asignada dos o más etiquetas. Esto se produce si el elemento se ha encontrado en más de una tabla de búsqueda. Si se da este caso, tendremos dos o más secuencias de etiquetas para la dirección postal.

Por ejemplo, si en el campo dirección postal tenemos el valor 'C/ Luna Sevilla', la aplicación lo podría limpiar, estandarizar y etiquetar como:

#C/	Luna	Sevilla
#Calle	Luna	Sevilla
TV:,	UN:,	LN:
TV:,	UN:,	PR:

Como se puede comprobar el elemento 'Sevilla' se ha etiquetado con 'LN' por haberse encontrado en la tabla de búsqueda de localidades (klocalidad.tbl) y con 'PR' por haberse encontrado en la tabla de búsqueda de provincias (kprovincia.tbl), ya que este elemento podría hacer referencia en una dirección postal tanto a una localidad como a una provincia.

Por el conocimiento que tiene el usuario acerca de los patrones o estructuras que siguen las direcciones postales de su fichero de trabajo, debe decidir si asigna a ambas etiquetas del elemento 'Sevilla', el estado 'localidad' o el estado 'provincia'.

En nuestro ejemplo, entendemos que el elemento 'Sevilla' hace referencia a la Localidad con lo cual en el proceso de asignación manual de estados las etiquetas quedarían de la siguiente forma:

#C/	Luna	Sevilla
#Calle	Luna	Sevilla
TV:tipo_de_via,	UN:nombre_de_via,	LN:localidad
TV:tipo_de_via,	UN:nombre_de_via,	PR:localidad

Ahora la aplicación calcularía la probabilidad de las dos secuencias de la siguiente forma:

- La probabilidad de la **Secuencia TV:tipo_de_via, UN:nombre_de_via, LN:localidad,** es según el algoritmo de Viterbi:

  Probabilidad de que la secuencia comience por el estado 'tipo_de_via' (0,743590) por la probabilidad de que ese estado esté etiquetado por 'TV' (0,980769) por la probabilidad de pasar del estado 'tipo_de_via' al estado 'nombre_de_via' (1) por la probabilidad de que el estado 'nombre_de via' esté etiquetado con 'UN' (0,988812) por la probabilidad de pasar del estado 'nombre_de_via' al estado 'localidad' (0,961410) por la probabilidad de que el estado 'localidad' esté etiquetado con 'LN' (0,925693).

  Así pues, su probabilidad es 0,641785.

- La probabilidad de la **Secuencia TV:tipo_de_via, UN:nombre_de_via, PR:localidad,** es según el algoritmo de Viterbi:

  Probabilidad de que la secuencia comience por el estado 'tipo_de_via' (0,743590) por la probabilidad de que ese estado esté etiquetado por 'TV' (0,980769) por la probabilidad de pasar del estado 'tipo_de_via' al estado 'nombre_de_via' (1) por la probabilidad de que el estado 'nombre_de via' esté etiquetado con 'UN' (0,988812) por la probabilidad de pasar del estado 'nombre_de_via' al estado 'localidad' (0,961410) por la probabilidad de que el estado 'localidad' esté etiquetado con 'PR' (0,379926).

  Así pues, su probabilidad es 0,263403.

Con lo cual la aplicación se quedaría automáticamente con la primera secuencia ya que tiene mayor probabilidad asociada.

Es necesario aclarar, que por seleccionar una muestra aleatoria simple con reposición para conocer la estructura que siguen los datos del campo a normalizar, no van a estar representadas todas las estructuras con lo cual algunas de estas secuencias no van a tener una probabilidad asociada. Esta situación se subsana con la práctica de **Métodos de Suavizado** que asignan una determinada probabilidad a las secuencias no registradas en la muestra. Su teoría puede consultarse en el **Anexo VII**.

A continuación vemos como se llevaría a cabo el entrenamiento de la muestra u obtención del Modelo Oculto de Markov a través de la aplicación *ADYN Herramienta de Normalización*, siguiendo con nuestro ejemplo de direcciones postales.

Para ello accederemos a la interfaz ***'03. Entrenamiento de la muestra'*** a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) tal y como se muestra en la siguiente imagen:

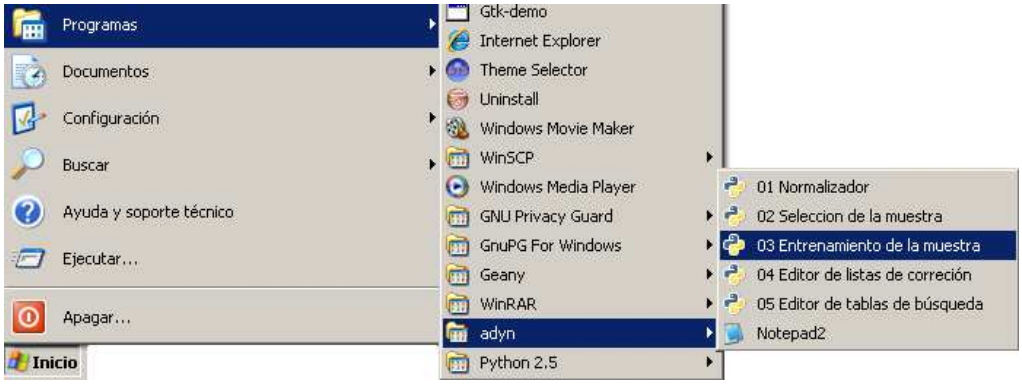


Imagen 21: Acceso a la interfaz '03. Entrenamiento de la muestra'.

Al abrir la interfaz ***'03.Entrenamiento de la muestra'*** nos recibirá la siguiente pantalla:

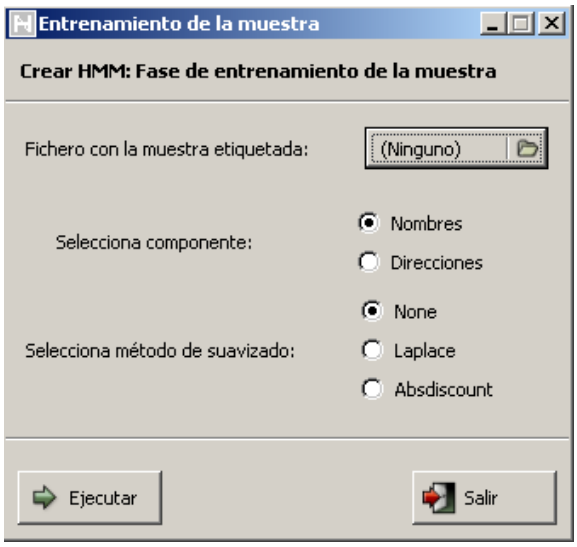


Imagen 22: Interfaz de entrenamiento de la muestra o de creación del HMM.

Primero el botón **'Fichero con la muestra etiquetada'**, nos solicita el fichero obtenido en el Paso 2, '*muestra_etiquetada_20091030-1241_Ejemplo.csv*'.

Seguidamente en **'Selecciona componente'** hemos de marcar si la componente a normalizar es un nombre o una dirección postal, en nuestro caso 'Direcciones'.

Por último en el apartado **'Selecciona método de suavizado'** solicita que indiquemos si queremos utilizar algún método de suavizado. En el ejemplo no usaremos ninguno.

Finalmente haremos click en **'Ejecutar'** y esperaremos a que el programa nos comunique que ha terminado con la siguiente pantalla:

Imagen 23: Ventana de verificación del proceso de entrenamiento.

El resultado de este paso será el Modelo Oculto de Markov que utilizaremos para normalizar el fichero original '*Ejemplo.csv*'. Este modelo será un fichero de extensión *'.hmm'* que encontraremos en la misma carpeta que '*Ejemplo.csv*' y tendrá un nombre con la estructura:

<fichero_de_origen>_<fecha_creación>-<hora_creación>.hmm

Se recomienda renombrarlo con un nombre más intuitivo a libre elección del usuario, como por ejemplo '*modeloX.hmm*'. Nosotros lo hemos denominado *'modelo1.hmm'*.

4.2 Normalización del fichero de datos.

Una vez que tenemos el Modelo Oculto de Markov creado lo utilizaremos para la normalización de los datos. Para ello, usaremos la interfaz ***'01. Normalizador'*** que nos proporciona *ADYN Herramienta de Normalización*. Accedemos a esta interfaz a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) tal y como se muestra en la siguiente imagen:

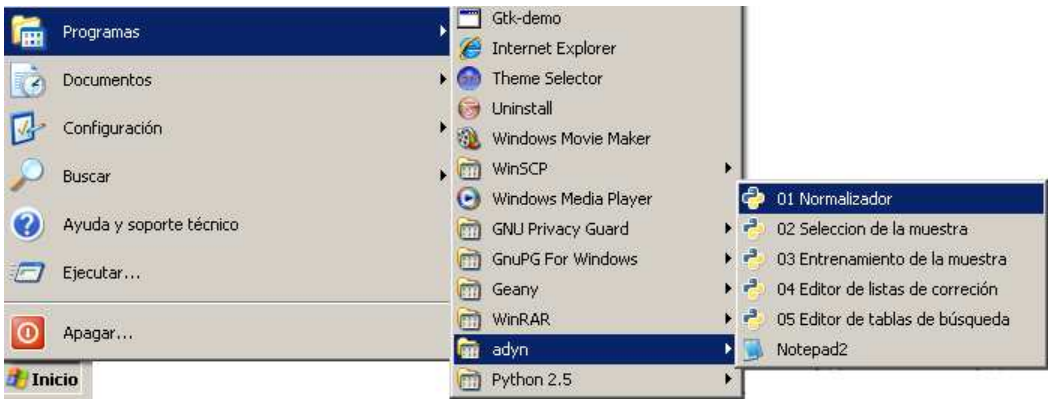


Imagen 24: Acceso a la interfaz '01. Normalizador'.

El resultado es la siguiente pantalla:

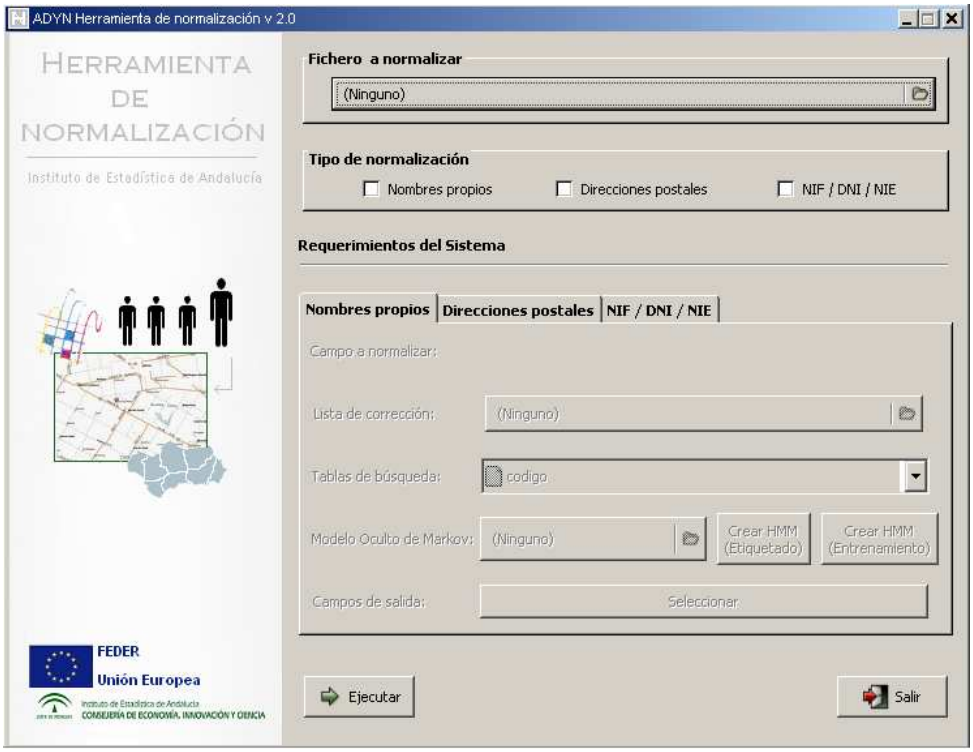


Imagen 25: ADYN Herramienta de Normalización: Interfaz de normalización.

En ella habrá que especificar en el botón **'Fichero a normalizar'** el fichero original de datos que queremos normalizar, es decir, '*Ejemplo.csv*'.

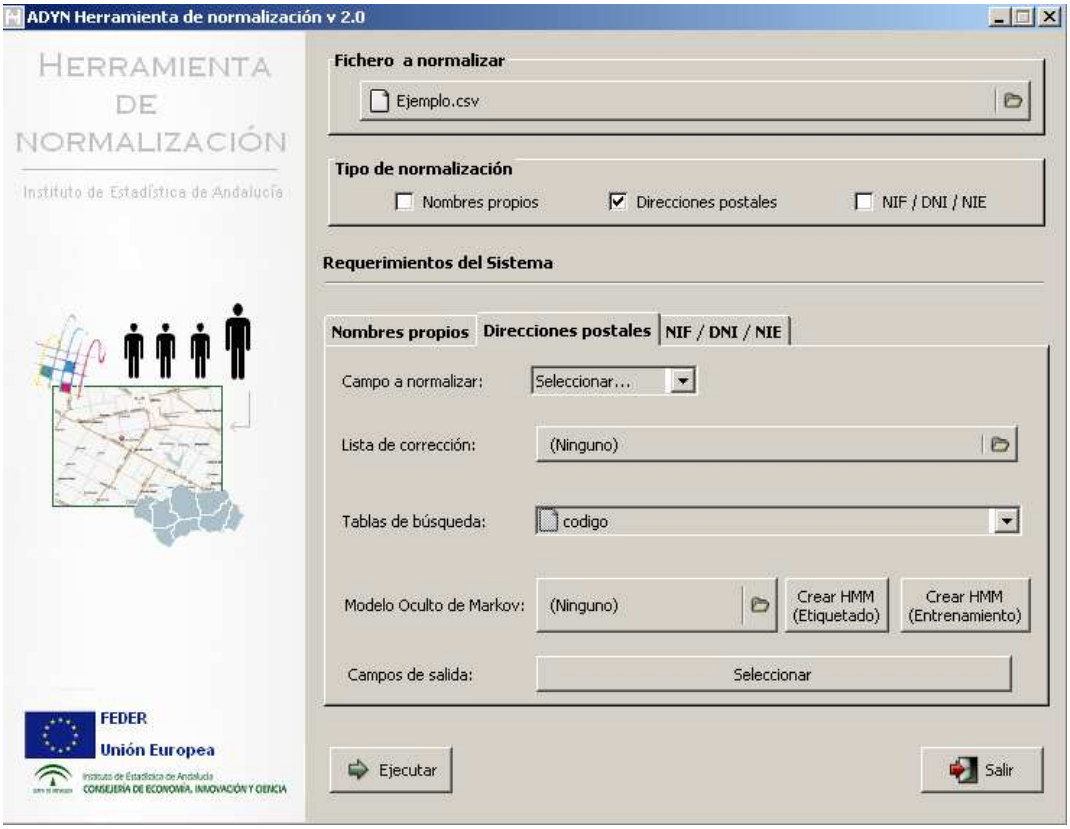


Imagen 26: Interfaz de normalización una vez seleccionado el tipo de normalización.

Seguidamente en **'Tipo de normalización'** hemos de marcar el campo que deseamos normalizar, es decir, Nombres propios, Direcciones postales o NIF/DNI/NIE. En nuestro caso marcaremos 'Direcciones postales'. Al marcar esta casilla se habilitará la pestaña **'Direcciones postales'** para que cumplimentemos los siguientes requerimientos del sistema, que son:

- **'Campo a normalizar'**: en este cuadro se listan todos los campos que contiene el fichero '*Ejemplo.csv*' y seleccionaremos el campo que contiene la dirección postal que es el que queremos normalizar. En nuestro caso será el campo '*direcciones*'.

- **'Lista de corrección'**: el proceso de elección de la lista de corrección es idéntico al realizado en la interfaz de selección y etiquetado de la muestra (Paso1). Es decir, que para especificar la lista de corrección tendremos que navegar por los directorios de la aplicación y encontrar el fichero '*direcciones_correccion.lst'*, que se encuentra dentro de la carpeta 'ListasDeCorreccion' en la ruta 'C:\adyn\codigo\datos\ListasDeCorreccion'.

- **'Tablas de búsqueda'**: al igual que para la lista de corrección el procedimiento de selección de las tablas de búsqueda es análogo al realizado en la interfaz de selección y etiquetado de la muestra (Paso 1). En este caso para especificar las tablas de búsqueda navegaremos por los directorios de la aplicación para buscar la carpeta '*direccion-tbl*' donde se encuentran las tablas de búsqueda para direcciones postales. La ubicación de esta carpeta es 'C:\adyn\codigo\datos' y al marcarla quedarán seleccionadas

automáticamente todas las tablas de búsqueda.

- **'Modelo Oculto de Markov'**: en este cuadro debemos especificar el Modelo Oculto de Markov creado previamente en el Paso 3 (Entrenamiento de la muestra) o cualquier otro que ya tengamos creado. Nosotros siguiendo con nuestro ejemplo utilizaremos el modelo creado en el Paso 3 y al que hemos denominado '*modelo1.hmm'*.

De esta forma nuestra interfaz queda definida de la siguiente forma:

Imagen 27: Interfaz de normalización tras seleccionar la lista de corrección y las tablas de búsqueda.

Hemos de notar que a través de esta interfaz podemos acceder directamente a las interfaces de selección y etiquetado de la muestra ('**Crear HMM (Etiquetado**')), así como a la de entrenamiento de ésta ('**Crear HMM (Entrenamiento)**'). El motivo de que estos botones se hayan insertado aquí se debe a que puede darse el caso de que no hayamos construido previamente el Modelo Oculto de Markov necesario para el proceso de normalización.

- Por último, si pulsamos sobre 'Seleccionar' del apartado **'Campos de salida'**, se abrirá una ventana con todos los posibles campos de salida del fichero normalizado y podremos desmarcar aquellos que no queremos que se muestren en el fichero de salida.

Imagen 28: Selección de los campos de salida de direcciones postales.

Por defecto aparecerán marcados todos los campos de salida y una vez elegidos todos o los seleccionados por el usuario pulsaremos 'OK'.

No obstante, si el usuario no pulsara sobre el apartado 'Seleccionar' la aplicación actuaría de igual forma que si se hubiesen marcado todos los campos.

Una vez especificados todos los parámetros necesarios para llevar a cabo el proceso de normalización pulsaremos **'Ejecutar'** y esperaremos unos segundos (o unos minutos si el fichero es grande) hasta que la interfaz nos avise de que se han normalizado todos los registros. El aviso nos llegará a través de la siguiente pantalla:

Imagen 29: Ventana de finalización del proceso de normalización.

El proceso de normalización generará dos ficheros de salida que se guardarán en la carpeta donde se encuentra el fichero original de datos, *'Ejemplo.csv'*. Estos serán:

- Fichero ***'est_<fecha_creación>-<hora_creación>_<fichero_datos>.csv'***: contendrá todos los campos del fichero original, junto con el campo direcciones estandarizado y segmentado en los campos que se han seleccionado previamente. Se recomienda abrirlo con '**Scalc**' del paquete ofimático Open Office 2.4.

- Fichero de proyecto ***'proy_<fecha_creación>-<hora_creación>_<fichero_datos>.py'***: contendrá el conjunto de parámetros con los que hemos realizado el anterior proceso de normalización, permitiendo reproducir o modificar este proceso posteriormente. Para poder ejecutar correctamente este fichero deberá guardarse en la carpeta ***'codigo'*** de la aplicación *ADYN* que se encuentra en la ruta 'C:\adyn\codigo' y haremos doble clic sobre él.

4.3 Validación del proceso de normalización.

Al abrir el fichero de datos normalizado, podrá apreciarse una columna extra llamada 'validacion' que contiene los valores 0 ó 1. Esta columna nos servirá para determinar si el proceso de normalización ha sido bueno o no según el Modelo Oculto de Markov utilizado. Para abrir el fichero hemos utilizado el programa '**Scalc**' del paquete ofimático Open Office 2.4, resultando:

est_20091120-0834_Ejemplo

	A	B	C	D	E	F	G	H	I	J	K
1	tipo_de_via	nombre_de_via	id_de_numero	numero	id_de_bloque	bloque	id_de_planta	planta	id_de_puerta	puerta	validacion
2	plaza	la solera	numero	3				1		d	0
3	plaza	del cabeceo	numero	4				4		c	0
4	plaza	grazalema			bloque	12		3		d	0
5	plaza	grazalema			bloque	12		2		c	0
6	calle	atalaya	numero	16				1		c	0
7		espiritu santo		2				bajo		c	0
8	parque	atlantico	numero	4				8		b	0
9	calle	la_cartuja			bloque	12	piso	1	puerta	a	0
10	camino	de vista alegre	numero	2							0
11	paseo	de las delicias	numero	5				11		r	0
12	plaza	andromeda			bloque	9		2		b	0

Imagen 30: Fichero normalizado donde se muestran varios campos de salida, entre ellos el campo de validación.

Si para un registro, la columna 'validacion' tiene un valor igual a 1 significa que la dirección postal contenida en ese registro está incorrectamente normalizada, es decir, los valores que aparecen en los campos de salida en los que se ha recogido la normalización de la dirección postal no se corresponden con los valores reales que deberían aparecer.

Si, por el contrario, un registro presenta valor 0 en esta columna, significa que el algoritmo de validación no ha encontrado nada que pueda indicar que la dirección postal de este registro está incorrectamente normalizada.

Por lo tanto, la importancia del proceso de validación es primordial ya que permite:

- Reconocer aquellas estructuras de datos que han sido mal normalizadas debido a que hay registros cuyas estructuras **NO** se han introducido en la muestra con la que se generó el Modelo Oculto de Markov que hemos utilizado para normalizar el fichero original de datos.

- Darnos cuenta de la existencia de valores que no están incluidos en las tablas de búsqueda y por lo tanto no pueden ser reconocidos por el Modelo Oculto de Markov a la hora de normalizar el fichero original de datos.

Con el fin de ir corrigiendo estos errores y construir un proceso de validación lo más eficiente posible tendremos que:

- Enriquecer el Modelo Oculto de Markov con las nuevas estructuras de datos no presentes en la muestra seleccionada aleatoriamente por la aplicación. Esto es, introducir las secuencias en el fichero '*muestra_etiquetada_20091030-1241_Ejemplo.csv*'.

- Insertar en las tablas de búsqueda esos nuevos elementos que han aparecido y que no estaban recogidos previamente en ellas.

Para llevar a cabo el enriquecimiento del Modelo Oculto de Markov tenemos dos opciones:

a) Para explicar la **primera opción** haremos uso del siguiente ejemplo:

   Supongamos que al realizar el proceso de normalización de direcciones postales el registro '*espiritu santo 2 bajo*' no está bien normalizado, es decir, su campo de validación tiene asignado el valor 1.

   Por el conocimiento que tenemos sobre los registros del fichero de trabajo y sobre las etiquetas y sus posibles estados, esta normalización errónea nos va a indicar que la estructura de dirección postal que comienza por un nombre de vía compuesto y sigue por el número de la vía y la planta no está presente en la muestra. Es decir, que la secuencia de etiquetas y estados asignados:

   UN:nombre_de_via, UN:nombre_de_via, UN:numero, PL:planta

   no está incluida en la muestra (podemos comprobarlo observando la Imagen 14).

   Entonces para corregir situaciones de este tipo e ir construyendo un Modelo Oculto de Markov lo más eficiente posible, o lo que es lo mismo, que contenga más estructuras de datos, podemos introducir en el fichero de la muestra etiquetada que hemos generado en el Paso 1 (Selección y etiquetado de la muestra) las estructuras de los registros mal normalizados. Esto significa introducir manualmente las etiquetas y estados correspondientes a esos registros. En nuestro ejemplo la forma de proceder será la siguiente:

   1. Abrimos el fichero '*muestra_etiquetada_20091030-1241_Ejemplo.csv*' con el editor de texto '**Notepad2**', fichero que contiene la muestra etiquetada junto con sus estados asignados:

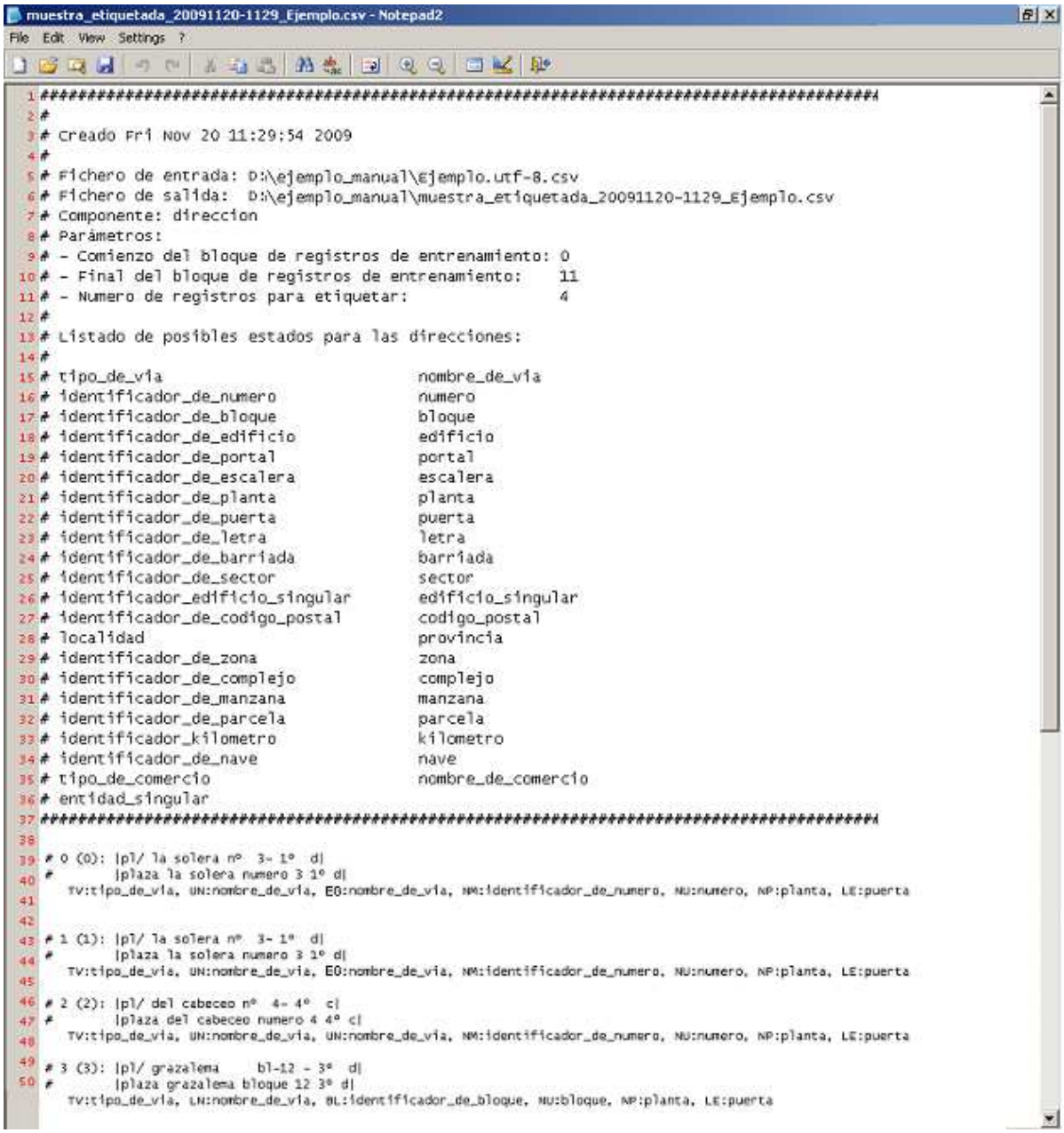


Imagen 31: Fichero de la muestra etiquetada con estados asignados.

2. Introduciremos en el fichero muestra etiquetada, la estructura o patrón correspondiente al registro: *'espiritu santo 2 bajo*'. Para ello, escribiremos debajo del cuarto registro de la muestra lo siguiente:

   #espiritu santo 2 bajo
   UN:nombre_de_via , UN:nombre_de_via , NU:numero , PL:planta

   De esta manera ya tenemos identificada la estructura o patrón de este registro y el fichero de la muestra etiquetada aparecerá de la forma:

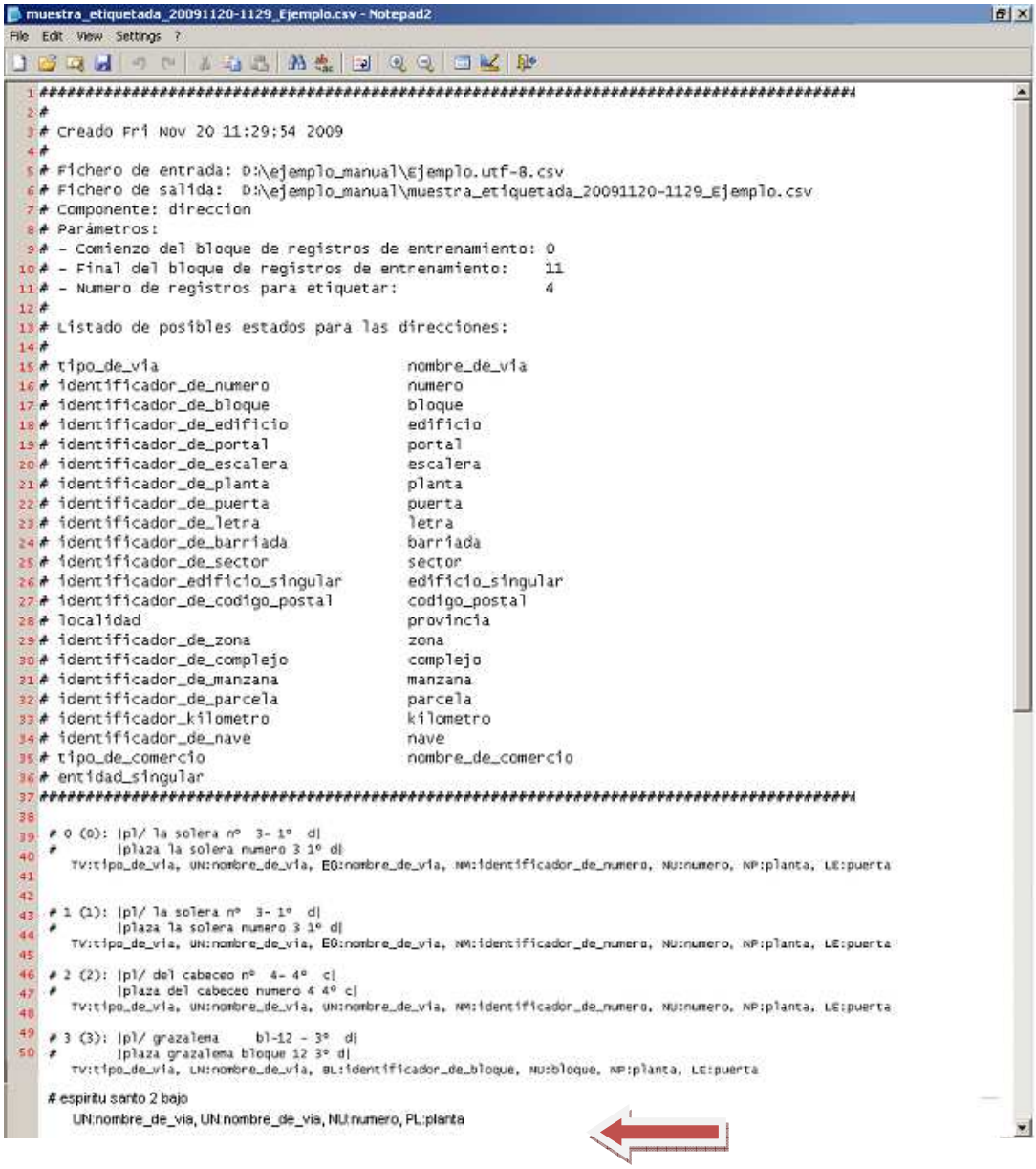


Imagen 32: Fichero de la muestra etiquetada con estados asignados y un nuevo patrón.

Como se puede observar lo importante es que se introduzca la secuencia de etiquetas y estados ya que es la parte que influirá en la construcción del Modelo Oculto de Markov. Si deseamos incluir información adicional sobre el registro al que hace referencia la secuencia, podemos hacerlo añadiendo una almohadilla al comienzo de la línea como hemos hecho:

#espiritu santo 2 bajo

A continuación utilizando esta nueva muestra etiquetada se volverá a construir el Modelo Oculto de Markov a través de la interfaz ***'03. Entrenamiento de la muestra'***. Seguidamente utilizando la interfaz ***'01. Normalizador'*** volveremos a realizar el

proceso de normalización del fichero original de datos con este nuevo Modelo Oculto de Markov creado.

b) La **segunda opción** que nos permite enriquecer el modelo HMM consiste en lo siguiente: una vez normalizado el campo del fichero de datos que se desee (nombres de personas o direcciones postales), supongamos por ejemplo el campo '*direcciones*' del fichero '*Ejemplo.csv*', se va a crear otro fichero de datos a partir de éste que sólo contenga los registros que estén mal normalizados. A continuación a través de la interfaz ***'02. Selección de la muestra'*** la aplicación seleccionará una muestra de esos registros. En este caso la componente a normalizar (Direcciones) quedará etiquetada y manualmente se introducirán los estados correspondientes.

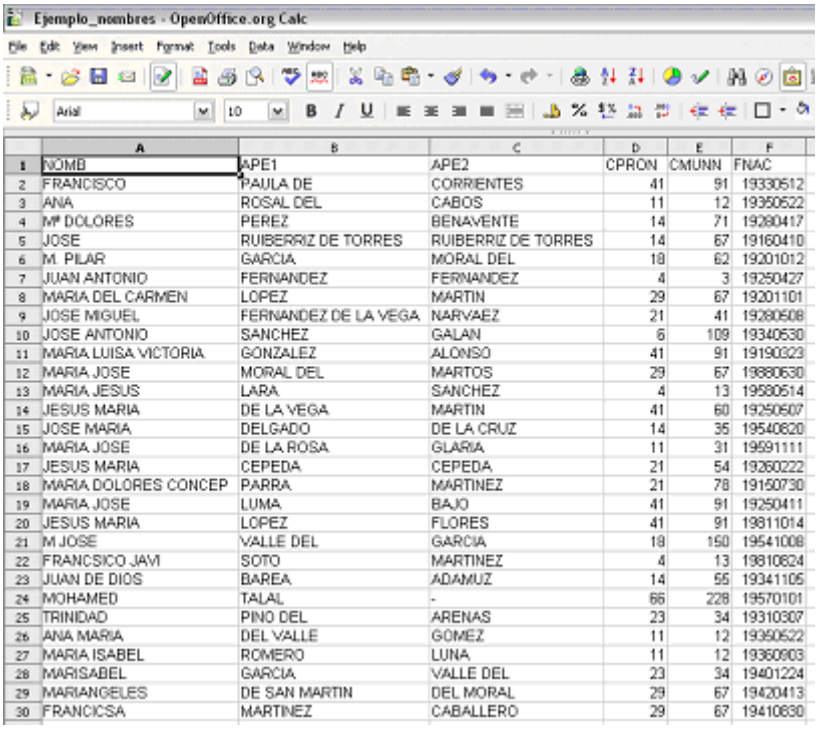
   Posteriormente, incluiremos en esta nueva muestra etiquetada, la muestra que nos ha servido para entrenar el primer modelo HMM y que nos ha permitido normalizar el fichero original de datos en primera instancia (en nuestro ejemplo '*muestra_etiquetada_20091030-1241_Ejemplo.csv*'). De esta forma tendremos un solo fichero que contendrá la unión de las dos muestras etiquetadas del fichero original de datos. Por último utilizaremos la interfaz ***'03. Entrenamiento de la muestra'*** para generar con esta nueva muestra etiquetada el nuevo Modelo Oculto de Markov.

La experiencia determina que la opción más eficiente es la descrita en a), sin embargo dejamos al usuario la elección de la misma según estime conveniente.

Anexo I: Normalización del campo Nombre de Persona.

Para llevar a cabo este proceso vamos a utilizar el fichero '*Ejemplo_nombres.csv*' ubicado en 'C:\adyn\ejemplos\nombres'. El fichero consta de los seis campos siguientes: nombre propio del individuo (NOMB), primer apellido (APE1), segundo apellido (APE2), código de la provincia de nacimiento (CPRON), código del municipio de nacimiento (CMUNN) y fecha de nacimiento (FNAC).

El fichero tiene un tamaño de 29 registros y su contenido se muestra en la siguiente imagen:


Ejemplo_nombres - OpenOffice.org Calc

	NOMB	APE1	APE2	CPRON	CMUNN	FNAC
2	FRANCISCO	PAULA DE	CORRIENTES	41	91	19330512
3	ANA	ROSAL DEL	CABOS	11	12	19350522
4	Mª DOLORES	PEREZ	BENAVENTE	14	71	19280417
5	JOSE	RUIBERRIZ DE TORRES	RUIBERRIZ DE TORRES	14	67	19160410
6	M. PILAR	GARCIA	MORAL DEL	18	62	19201012
7	JUAN ANTONIO	FERNANDEZ	FERNANDEZ	4	3	19250427
8	MARIA DEL CARMEN	LOPEZ	MARTIN	29	67	19201101
9	JOSE MIGUEL	FERNANDEZ DE LA VEGA	NARVAEZ	21	41	19280508
10	JOSE ANTONIO	SANCHEZ	GALAN	6	109	19340530
11	MARIA LUISA VICTORIA	GONZALEZ	ALONSO	41	91	19190323
12	MARIA JOSE	MORAL DEL	MARTOS	29	67	19880630
13	MARIA JESUS	LARA	SANCHEZ	4	13	19580514
14	JESUS MARIA	DE LA VEGA	MARTIN	41	60	19250507
15	JOSE MARIA	DELGADO	DE LA CRUZ	14	35	19540820
16	MARIA JOSE	DE LA ROSA	GLARIA	11	31	19591111
17	JESUS MARIA	CEPEDA	CEPEDA	21	54	19260222
18	MARIA DOLORES CONCEP	PARRA	MARTINEZ	21	78	19150730
19	MARIA JOSE	LUMA	BAJO	41	91	19250411
20	JESUS MARIA	LOPEZ	FLORES	41	91	19811014
21	M JOSE	VALLE DEL	GARCIA	18	150	19541008
22	FRANCSICO JAVI	SOTO	MARTINEZ	4	13	19810824
23	JUAN DE DIOS	BAREA	ADAMUZ	14	55	19341105
24	MOHAMED	TALAL	-	66	228	19570101
25	TRINIDAD	PINO DEL	ARENAS	23	34	19310307
26	ANA MARIA	DEL VALLE	GOMEZ	11	12	19350522
27	MARIA ISABEL	ROMERO	LUNA	11	12	19360903
28	MARISABEL	GARCIA	VALLE DEL	23	34	19401224
29	MARIANGELES	DE SAN MARTIN	DEL MORAL	29	67	19420413
30	FRANCICSA	MARTINEZ	CABALLERO	29	67	19410830

Como se puede observar el campo a normalizar 'nombre de persona' no se encuentra en una sola columna sino que está segmentado en tres: NOMB, APE1 y APE2, con lo cual para llevar a cabo la normalización se ha tomado la decisión de normalizar en primer lugar el campo 'NOMB' que contiene el nombre propio del individuo, a continuación el campo 'APE1' que contiene el primer apellido de éste y por último el campo 'APE2' que contiene el segundo apellido. Hay que decir que se podría haber elegido cualquier otro orden de normalización.

Antes de llevar a cabo la normalización observamos si los campos a normalizar contienen comas ',' y si es así habría que eliminarlas. Observando el fichero se comprueba que este no es el caso, con lo cual empezamos con la normalización de los tres campos utilizando *ADYN Herramienta de normalización*.

Para ello procederemos del mismo modo que cuando normalizamos el fichero '*Ejemplo.csv*' de direcciones postales, es decir, tendremos que seguir las tres etapas siguientes:

1. Creación del Modelo Oculto de Markov.

1.1. Selección y etiquetado de la muestra.

1.2. Asignación manual de estados.

1.3. Entrenamiento de la muestra.

2. Normalización del campo del fichero de datos.

3. Validación del proceso de normalización.

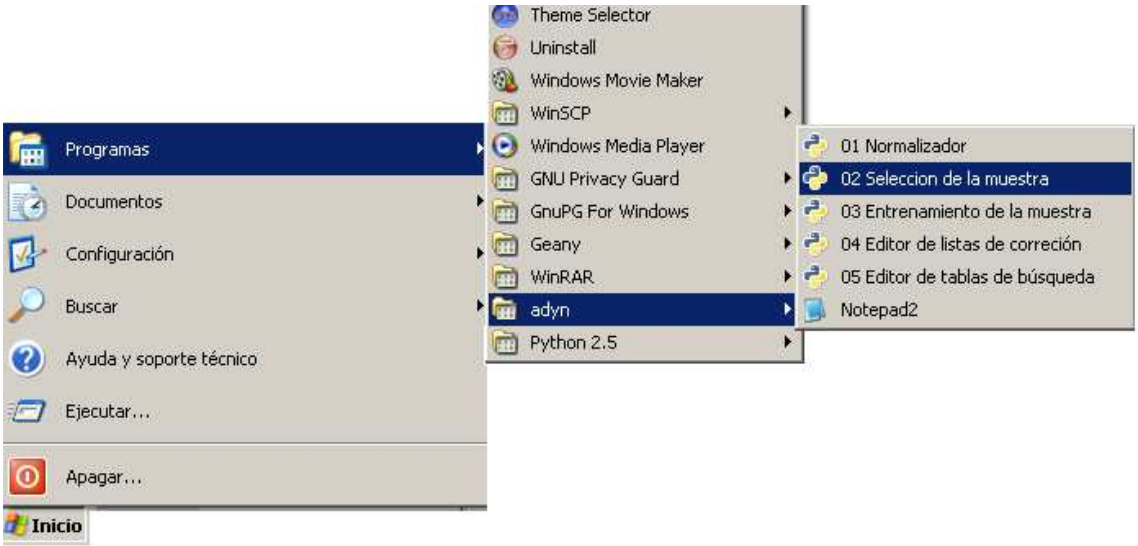
Así pues, empezamos con la normalización del primer campo.

PROCESO DE NORMALIZACIÓN DEL CAMPO '*nomb*'

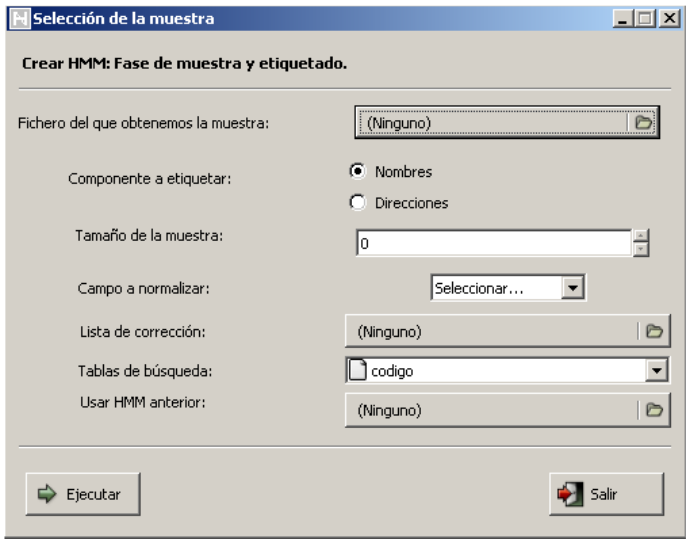
1. Creación del Modelo Oculto de Markov

1.1. Selección de la muestra

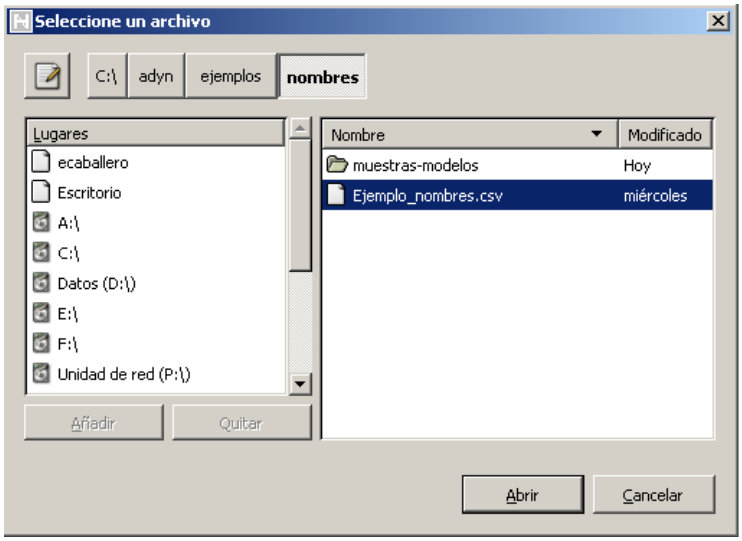
Para seleccionar la muestra utilizamos la interfaz ***'02. Selección de la muestra'*** de la aplicación *ADYN Herramienta de Normalización*. Accedemos a esta interfaz (si trabajamos con Windows) a través de:



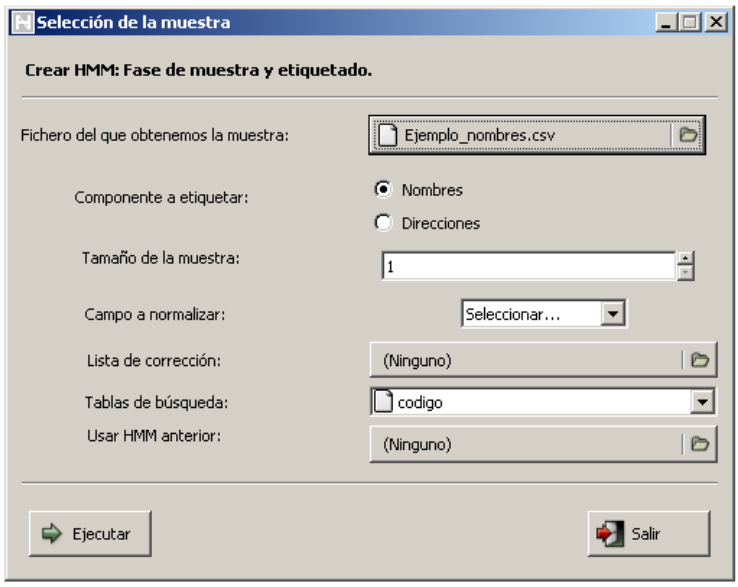
Y nos aparecerá la siguiente ventana:



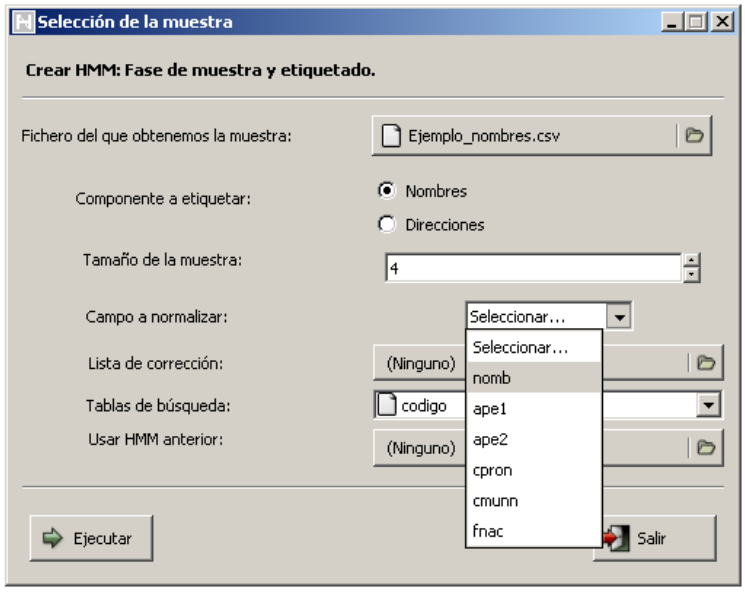
Seleccionamos el **'Fichero del que obtenemos la muestra'**, en nuestro caso '*Ejemplo_nombres.csv*', que se encuentra en la ruta 'C:\adyn\ejemplos\nombres':



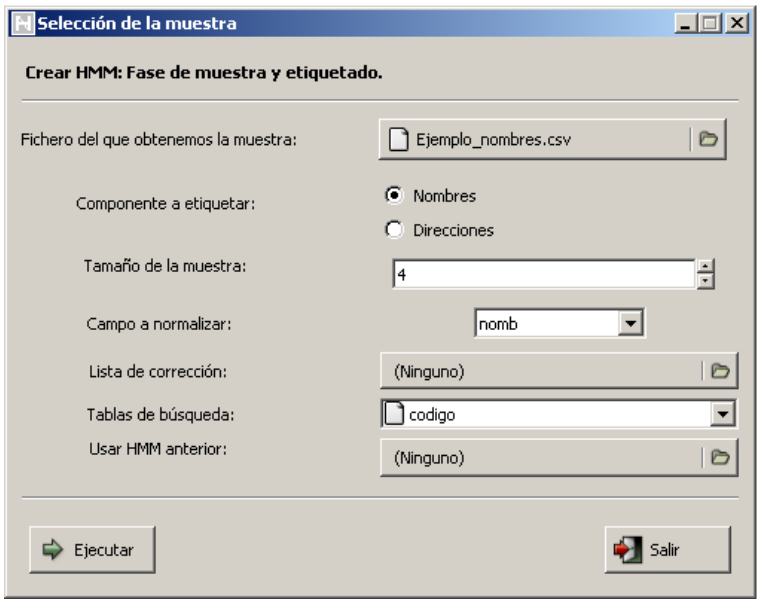
A continuación seleccionamos la **'Componente a etiquetar'**, que en nuestro caso es 'Nombres' y es la que está seleccionada por defecto en la aplicación:



El siguiente paso será especificar el **'Tamaño de la muestra'** a seleccionar y el **'Campo a normalizar'**. En cuanto al tamaño hemos elegido el valor cuatro pero el usuario podría haber indicado otro. Notar que el hecho de que el tamaño de muestra sea igual al del ejemplo de direcciones postales es una simple coincidencia. Y en cuanto al campo a normalizar se ha seleccionado el campo **'***nomb***'** de entre todos los campos que componen el fichero de trabajo '*Ejemplo_nombres.csv*':

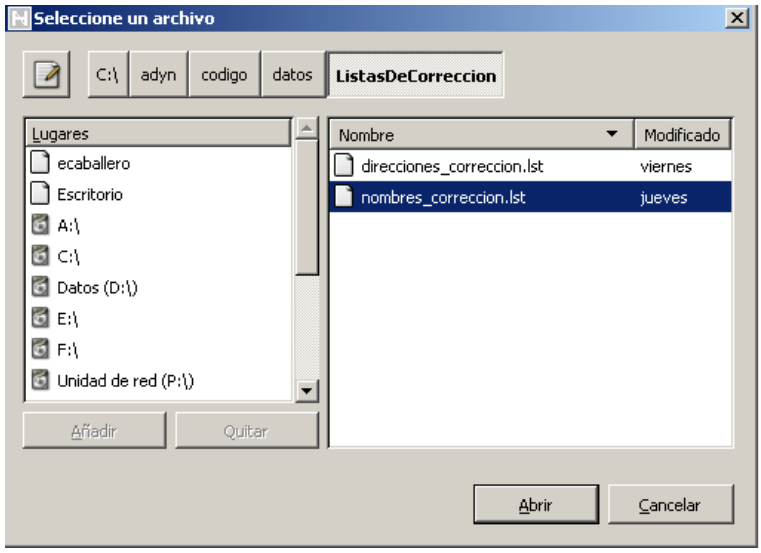


Tras la especificación de estos parámetros, la interfaz ha quedado configurada de la siguiente forma:

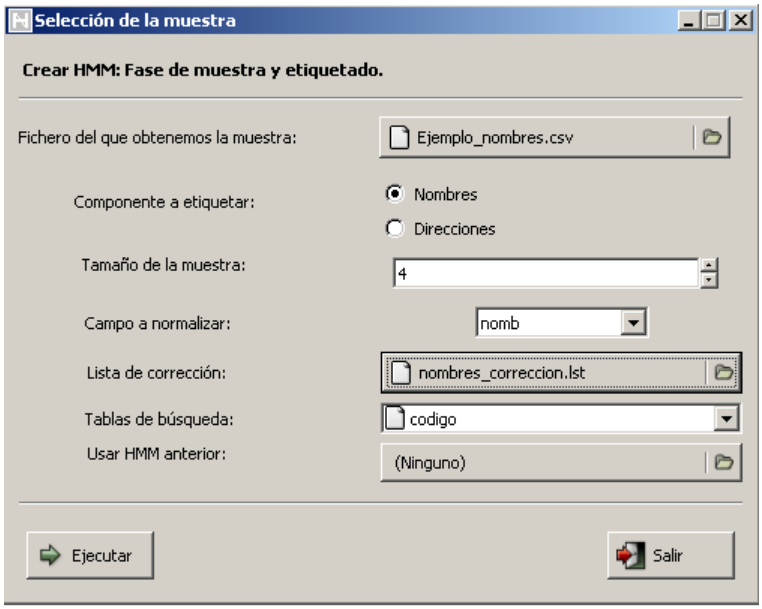


Los siguientes pasos consisten en seleccionar la **'Lista de corrección'** y las **'Tablas de búsqueda'** relativas a nombres de personas que nos van a servir para limpiar, estandarizar y etiquetar la muestra extraída del campo '*nomb*' del fichero de trabajo.

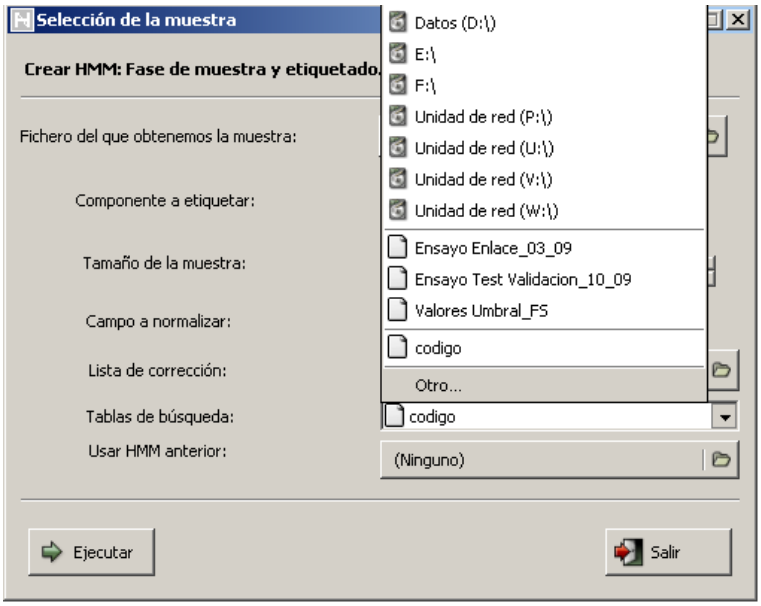
La **'Lista de corrección'** para nombres, '*nombres_correccion.lst*' se encuentra en la ruta 'C:\adyn\codigo\datos\ListasDeCorreccion':



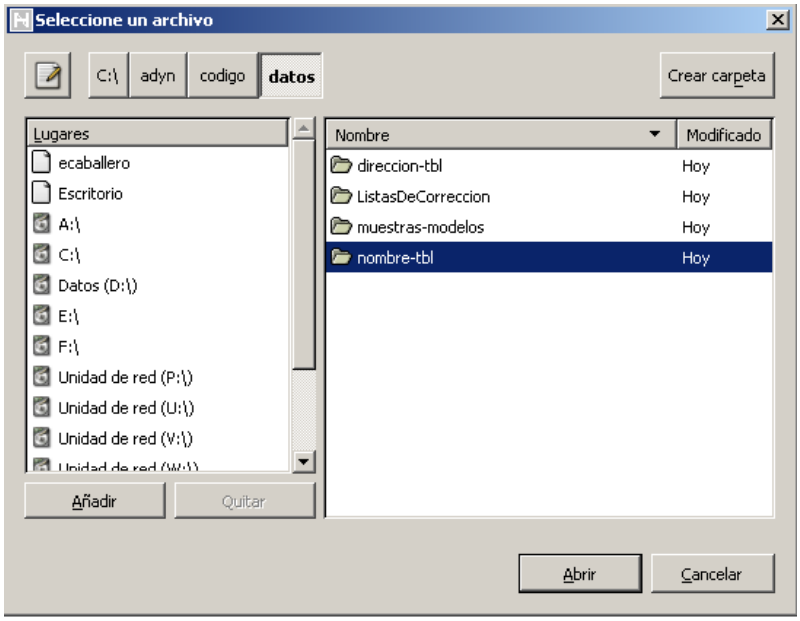
Una vez seleccionada, la interfaz queda configurada como se muestra en la siguiente imagen:



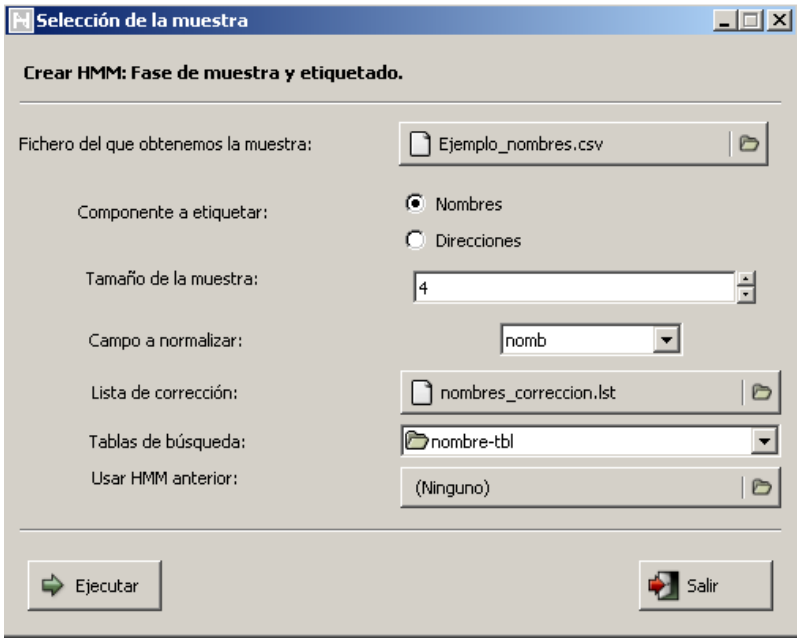
Por otro lado, las **'Tablas de búsqueda'** para nombres se encuentran en la ruta 'C:\adyn\codigo\datos. Para acceder con mayor facilidad a esta ruta se recomienda pulsar 'Otro' en el cuadro de búsqueda y directamente entramos en la carpeta '*codigo*':



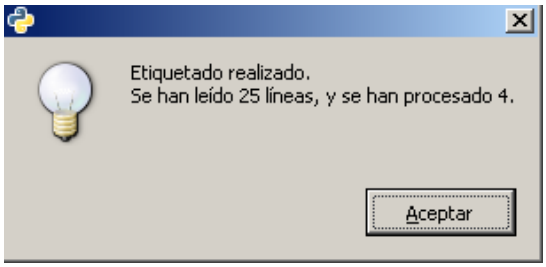
A continuación accedemos a '*datos*' y seleccionando la carpeta '*nombre-tbl*' tendremos seleccionadas automáticamente todas las tablas de búsqueda relativas a nombres de personas:



En caso de tener un Modelo Oculto de Markov ya creado podríamos utilizarlo para agilizar el proceso de asignación de estados a la muestra seleccionada. Como este no es el caso, el apartado **'Usar HMM anterior'** quedará vacío. Con lo cual, la interfaz quedará configurada como se muestra en la siguiente pantalla:



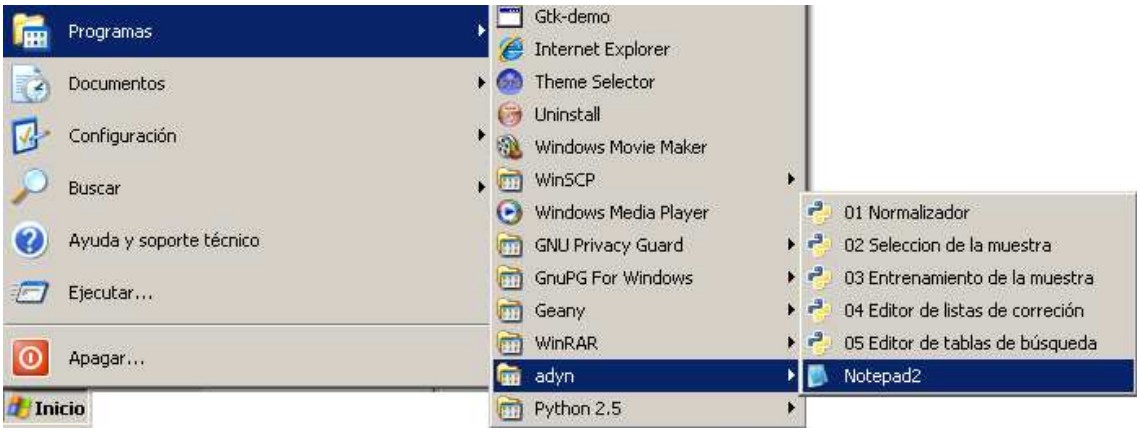
Ahora sólo tenemos que pulsar **'Ejecutar'** y esperaremos a que la aplicación nos indique que el proceso ha finalizado mediante el siguiente mensaje:



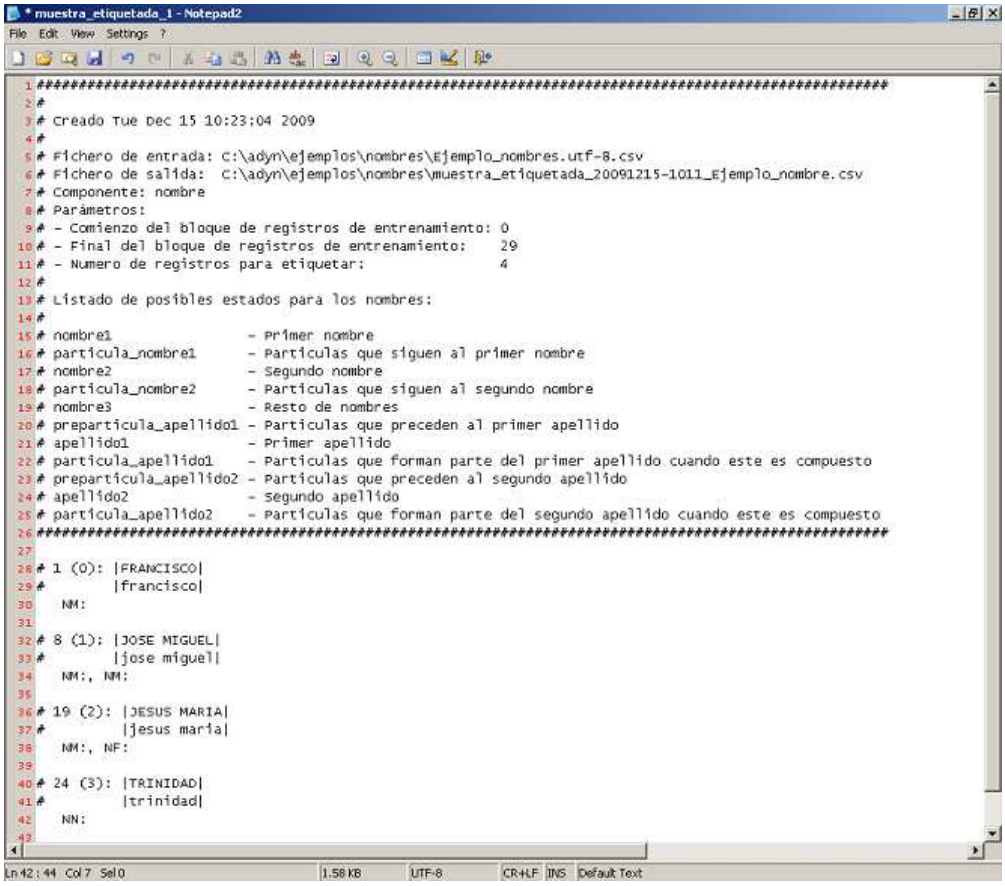
Esta pantalla nos indica que el número de líneas que se han leído aleatoriamente son 25, es decir, no se han leído todas las líneas del campo a normalizar '*nomb*' (29) y de ellas se han procesado 4, es decir, se ha extraído una muestra de cuatro registros.

Como resultado de este proceso de etiquetado obtenemos un fichero de salida del tipo '*muestra_etiquetada_<fecha_creación>-<hora_creación>_<fichero_origen>.csv*', que se encuentra en el mismo lugar que el fichero de datos original '*Ejemplo_nombres.csv*'. Este fichero ha sido renombrado por comodidad y su nueva denominación es '*muestra_etiquetada_1.csv*'.

Para ver su contenido utilizaremos el editor de texto **'Notepad2'** que se incluye en la aplicación *ADYN* y al que se puede acceder (si se trabaja con Windows) desde:



Así pues, el contenido del fichero de la muestra etiquetada para el nombre propio es:




```
################################################################################################################
#
# Creado Tue Dec 15 10:23:04 2009
#
# Fichero de entrada: C:\adyn\ejemplos\nombres\Ejemplo_nombres.utf-8.csv
# Fichero de salida:  C:\adyn\ejemplos\nombres\muestra_etiquetada_20091215-1011_Ejemplo_nombre.csv
# Componente: nombre
# Parámetros:
# - Comienzo del bloque de registros de entrenamiento: 0
# - Final del bloque de registros de entrenamiento:    29
# - Numero de registros para etiquetar:                4
#
# Listado de posibles estados para los nombres:
#
# nombre1                - Primer nombre
# particula_nombre1      - Particulas que siguen al primer nombre
# nombre2                - Segundo nombre
# particula_nombre2      - Particulas que siguen al segundo nombre
# nombre3                - Resto de nombres
# preparticula_apellido1 - Particulas que preceden al primer apellido
# apellido1              - Primer apellido
# particula_apellido1    - Particulas que forman parte del primer apellido cuando este es compuesto
# preparticula_apellido2 - Particulas que preceden al segundo apellido
# apellido2              - Segundo apellido
# particula_apellido2    - Particulas que forman parte del segundo apellido cuando este es compuesto
################################################################################################################

# 1 (0): |FRANCISCO|
#        |francisco|
   NM:

# 8 (1): |JOSE MIGUEL|
#        |jose miguel|
   NM:, NM:

# 19 (2): |JESUS MARIA|
#         |jesus maria|
   NM:, NF:

# 24 (3): |TRINIDAD|
#         |trinidad|
   NN:
```

Y el detalle de la muestra de registros seleccionada por la aplicación se muestra a continuación:

```
# 1 (0): |FRANCISCO|
#        |francisco|
   NM:

# 8 (1): |JOSE MIGUEL|
#        |jose miguel|
   NM:, NM:

# 19 (2): |JESUS MARIA|
#         |jesus maria|
   NM:, NF:

# 24 (3): |TRINIDAD|
#         |trinidad|
   NN:
```

El listado de todas las etiquetas asociadas a nombres de personas se muestra en el **Anexo IV** de este documento.

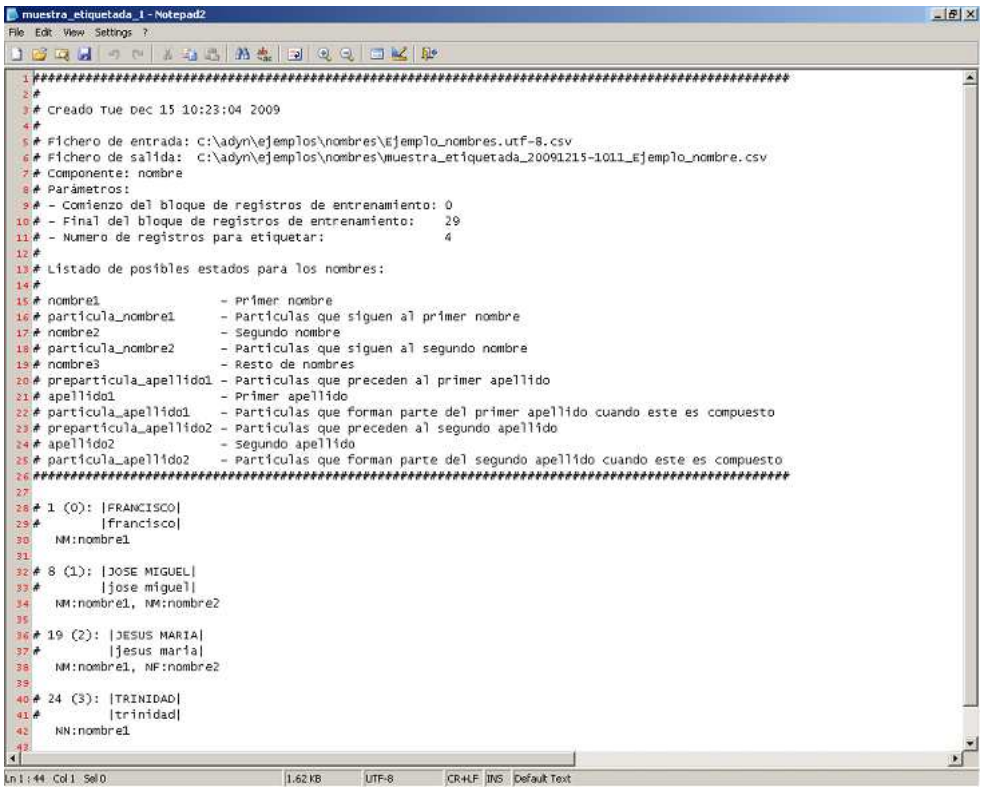
1.2. Asignación manual de estados

El siguiente paso será asignar manualmente estados a estas etiquetas. Si no tenemos abierto el fichero anterior: '*muestra_etiquetada_1.csv*' volveremos a abrirlo con el editor de texto '**Notepad2**' igual que antes.
A continuación iremos asignando a cada etiqueta el estado que consideramos adecuado. Esta asignación se realiza según el conocimiento que se tiene sobre el fichero de trabajo y nuestra experiencia, de tal forma que a continuación se muestra un detalle de cómo quedaría la muestra etiquetada junto con sus estados asociados.

```
# 1 (0): |FRANCISCO|
#        |francisco|
   NM:nombre1

# 8 (1): |JOSE MIGUEL|
#        |jose miguel|
   NM:nombre1, NM:nombre2

# 19 (2): |JESUS MARIA|
#         |jesus maria|
   NM:nombre1, NF:nombre2

# 24 (3): |TRINIDAD|
#         |trinidad|
   NN:nombre1
```

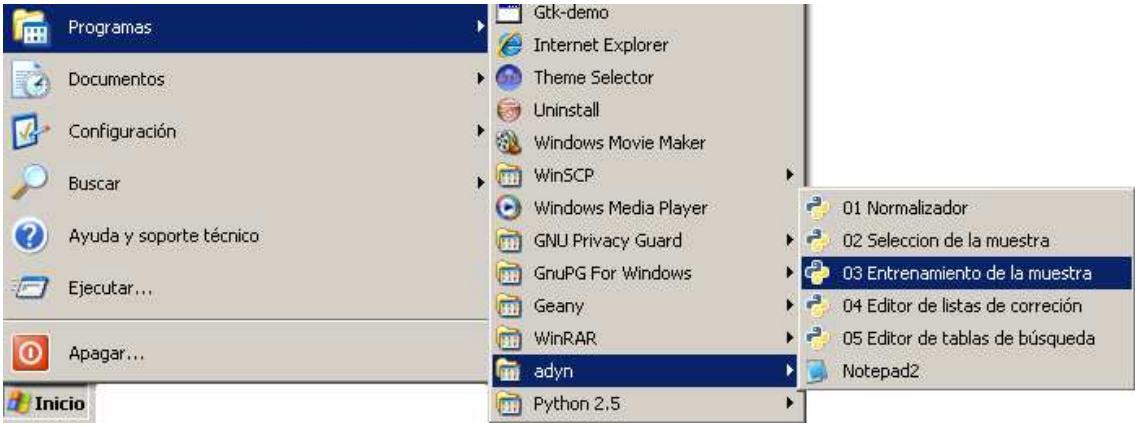
El listado de los estados asociados a nombres de personas se muestra en el **Anexo IV** de este documento.

Guardamos estos cambios y el fichero '*muestra_etiquetada_1.csv*' quedaría de la forma:



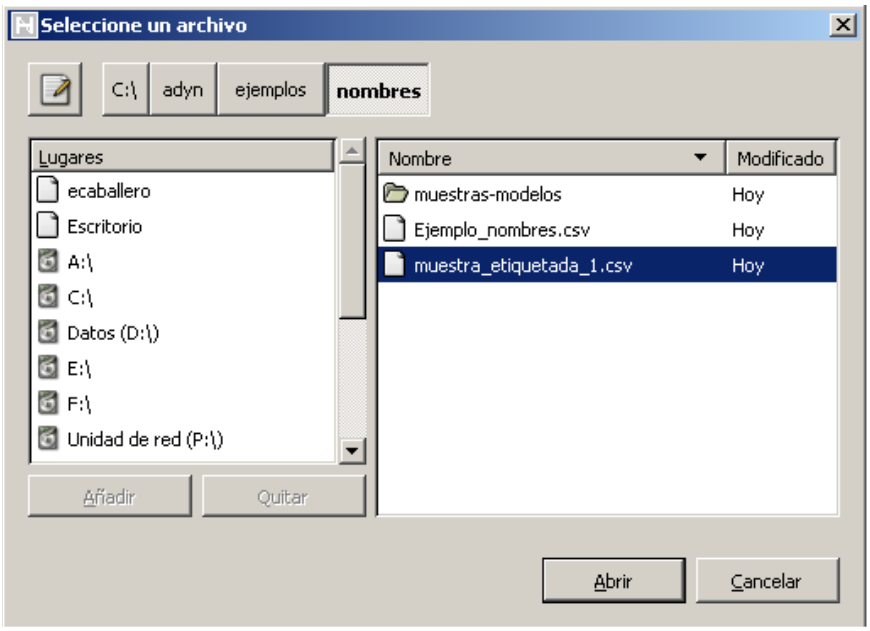
1.3. Entrenamiento de la muestra

A continuación utilizaremos el fichero anterior para llevar a cabo el entrenamiento de la muestra. Para acceder a la interfaz de entrenamiento procedemos como muestra la siguiente imagen (si trabajamos con Windows):

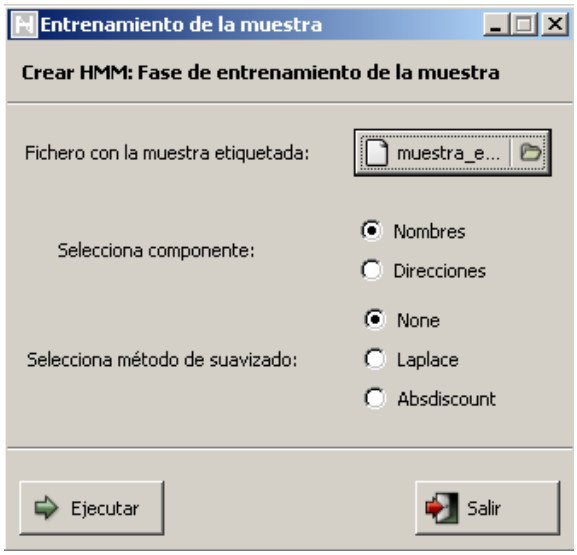


La pantalla que nos recibe es la siguiente:

En el campo **'Fichero con la muestra etiquetada'** seleccionamos el fichero '*muestra_etiquetada_1.csv*':

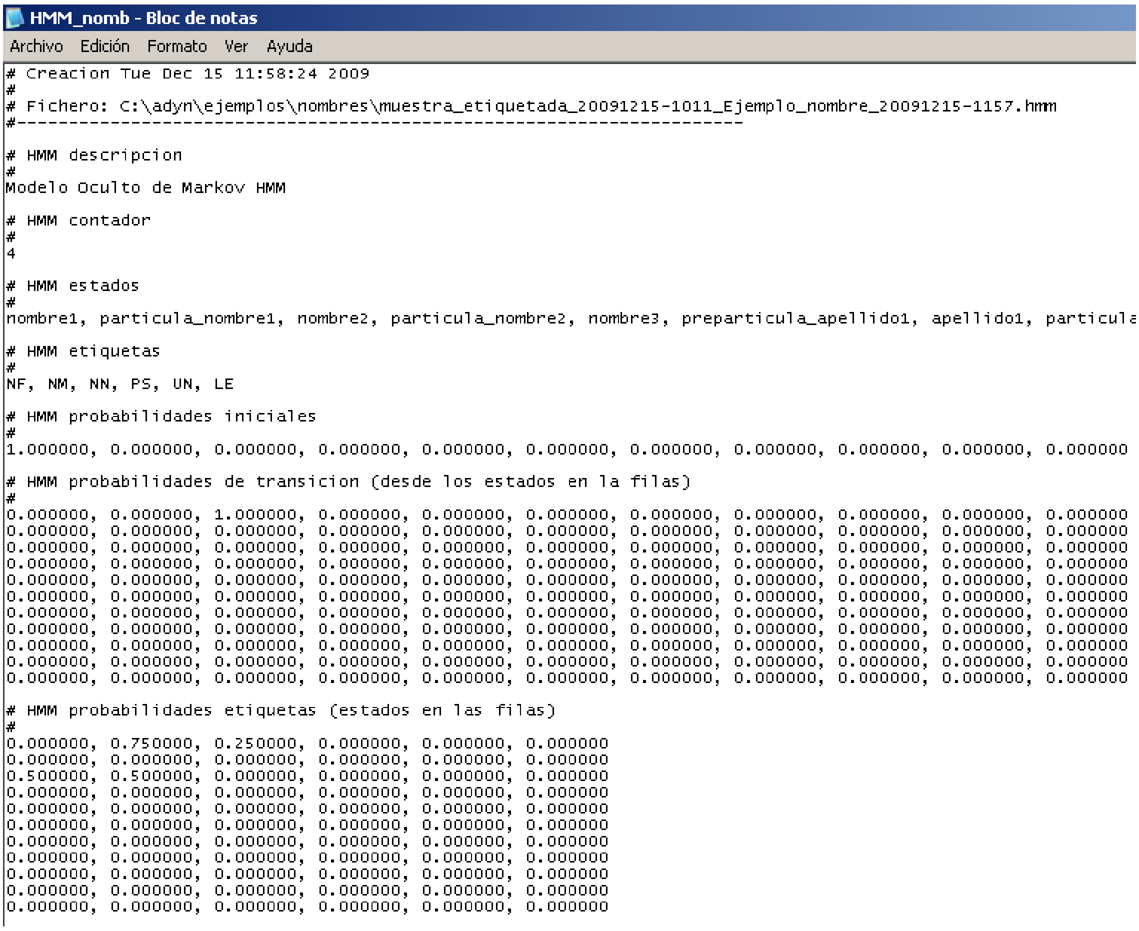


Posteriormente en **'Selecciona componente'** marcamos la componente a normalizar. En nuestro caso 'Nombres'. Y **'Selecciona método de suavizado'** lo dejamos como aparece marcado por defecto, lo que nos indica que no elegiremos ninguno. Por tanto la interfaz ha quedado definida como se muestra en la siguiente imagen:



A continuación pulsamos **'Ejecutar'** y esperamos a que finalice el proceso de entrenamiento, que sabremos que ha terminado cuando se muestre el siguiente mensaje en pantalla:

Como resultado de este proceso obtenemos un fichero de extensión '*.hmm*' que por comodidad lo renombraremos como, '*HMM_nomb.hmm*'. Dicho fichero se encuentra en el mismo directorio que el fichero de trabajo '*Ejemplo_nombres.csv*' y su contenido se muestra en la siguiente imagen:

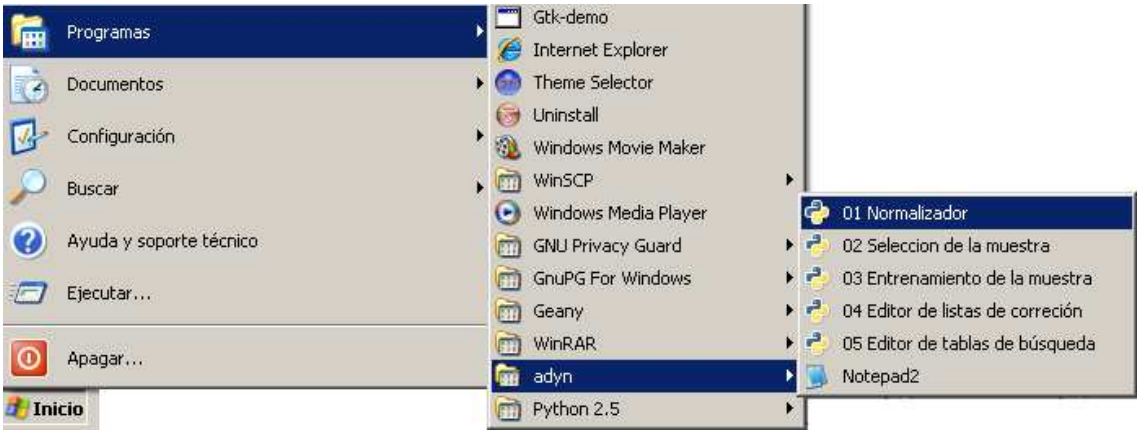


```
HMM_nomb - Bloc de notas
Archivo  Edición  Formato  Ver  Ayuda
# Creacion Tue Dec 15 11:58:24 2009
#
# Fichero: C:\adyn\ejemplos\nombres\muestra_etiquetada_20091215-1011_Ejemplo_nombre_20091215-1157.hmm
#---------------------------------------------------------------------

# HMM descripcion
#
Modelo Oculto de Markov HMM

# HMM contador
#
4

# HMM estados
#
nombre1, particula_nombre1, nombre2, particula_nombre2, nombre3, preparticula_apellido1, apellido1, particula

# HMM etiquetas
#
NF, NM, NN, PS, UN, LE

# HMM probabilidades iniciales
#
1.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000

# HMM probabilidades de transicion (desde los estados en la filas)
#
0.000000, 0.000000, 1.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000

# HMM probabilidades etiquetas (estados en las filas)
#
0.000000, 0.750000, 0.250000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.500000, 0.500000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
```

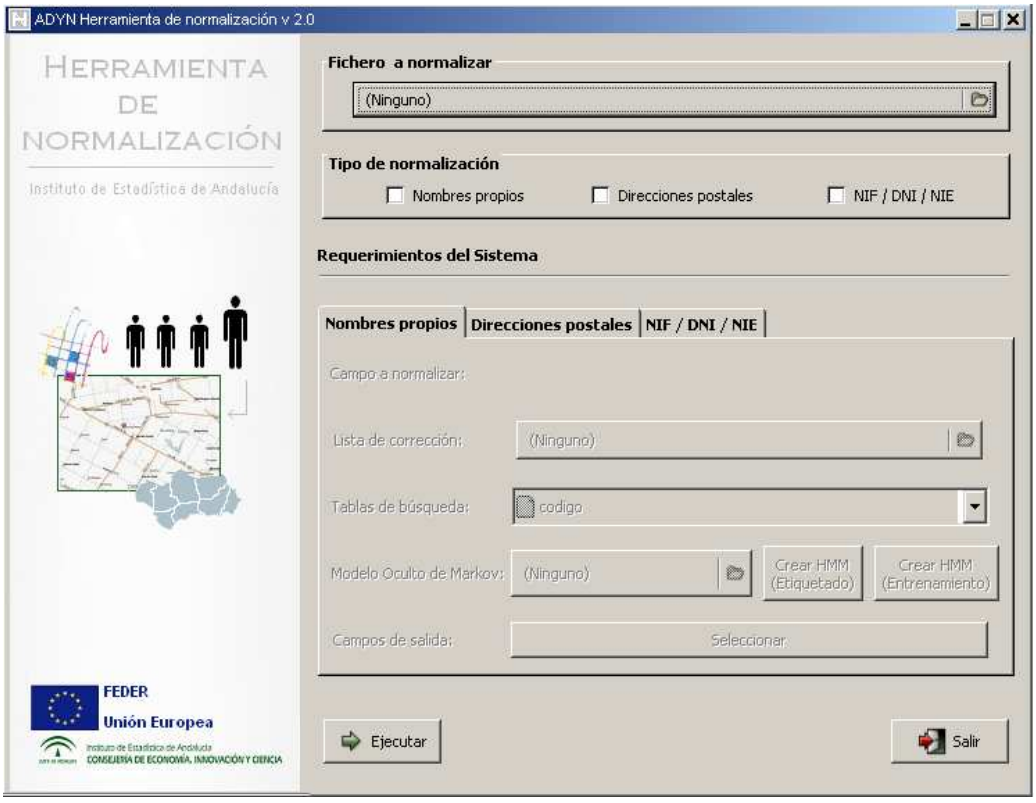
Este fichero será el Modelo Oculto de Markov que utilizaremos para estandarizar y segmentar el campo **'***nomb***'** del fichero de trabajo '*Ejemplo_nombres.csv*'.

2. Normalización del campo '*nomb*'

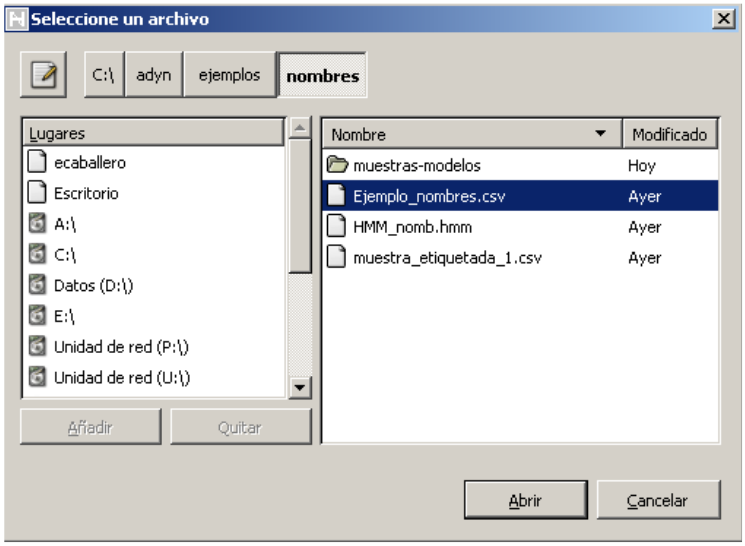
Una vez que tenemos creado el Modelo Oculto de Markov procederemos a llevar a cabo la normalización de dicho campo. Para ello accederemos a la interfaz ***'01. Normalizador'*** de la siguiente forma (si trabajamos con Windows):



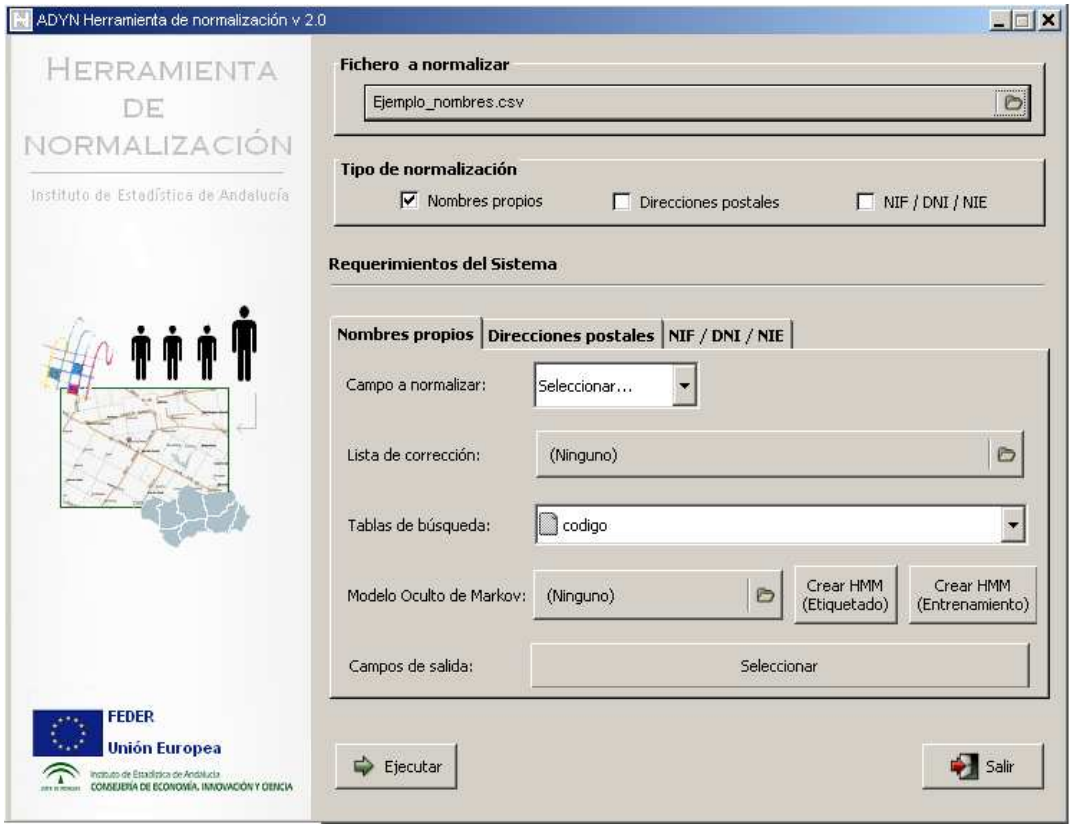
Aparecerá la siguiente pantalla:



En primer lugar seleccionamos el fichero que queremos normalizar:

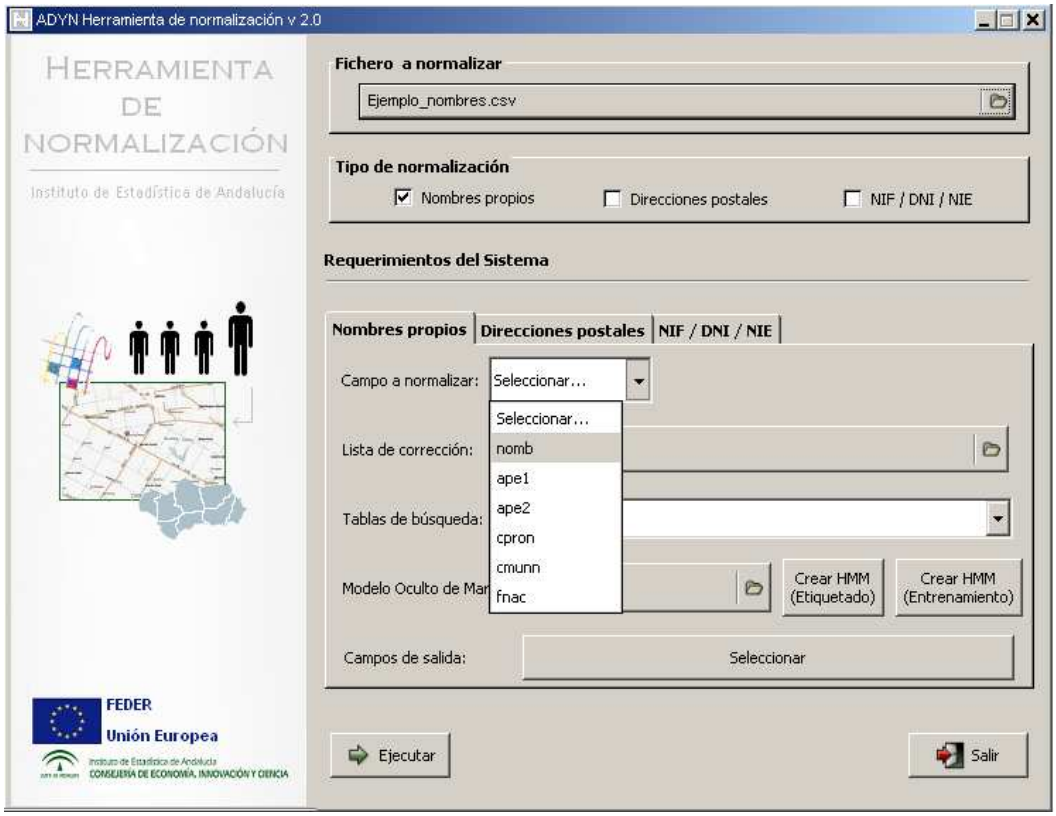


A continuación indicamos el **'Tipo de normalización'** que vamos a llevar a cabo, en este caso marcamos 'Nombres' ya que es lo que queremos normalizar, con lo que la interfaz va quedando configurada como:

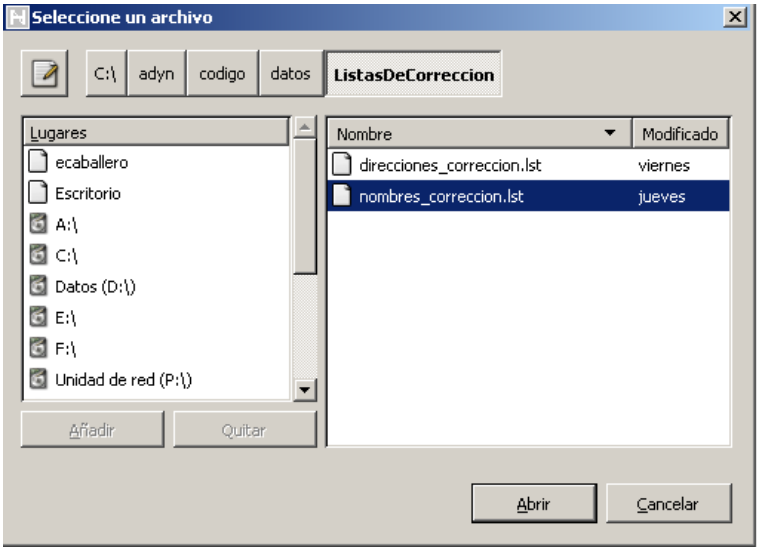


El siguiente paso es determinar los requerimientos del sistema.

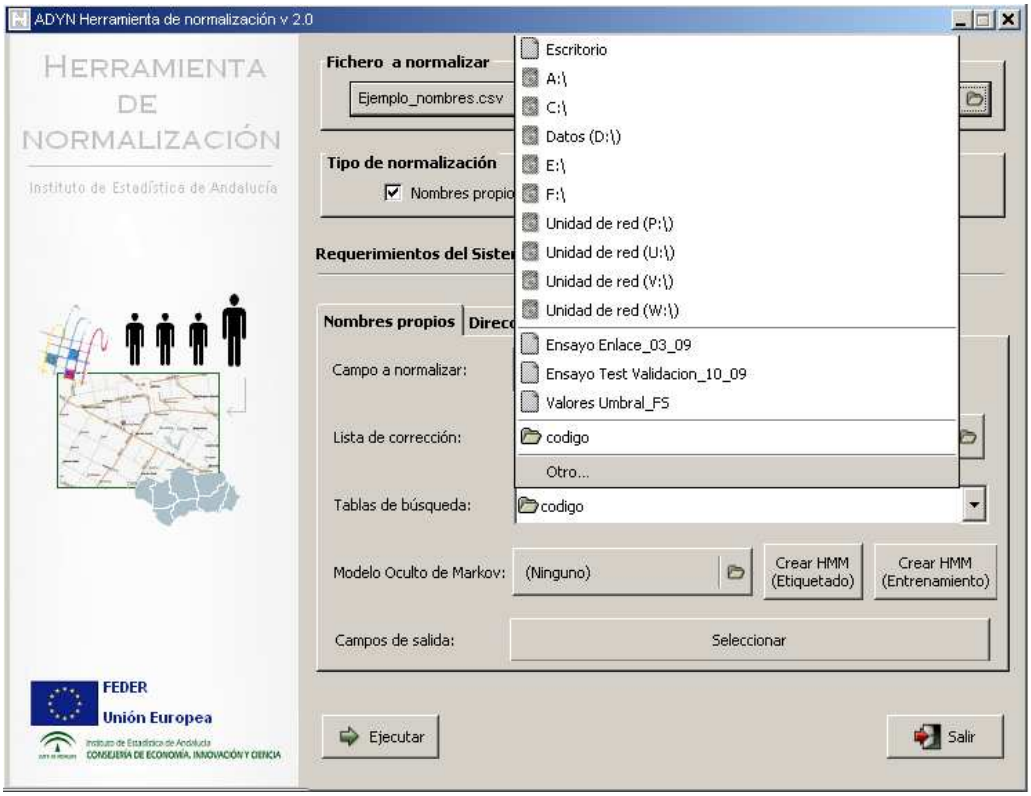
Así pues en **'Campo a normalizar'** elegimos el campo '*nomb*' de entre los existentes:



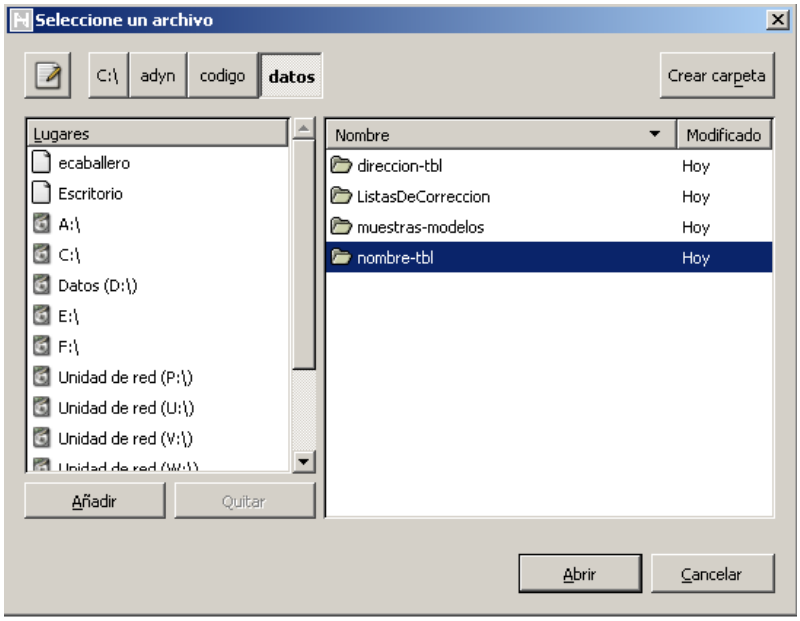
En **'Lista de corrección'** seleccionamos el fichero indicado en la imagen:



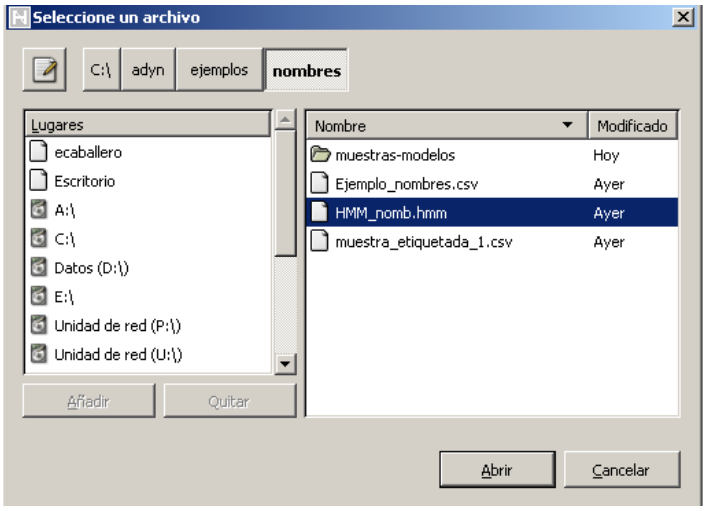
Para las **'Tablas de búsqueda'** pulsamos 'Otro' y accedemos directamente a la carpeta '*codigo*'. Dentro de ella, accederemos a '*datos*' y marcaremos la carpeta '*nombre-tbl*'.



Seleccionamos el directorio correspondiente a las tablas de búsqueda de nombres de persona.

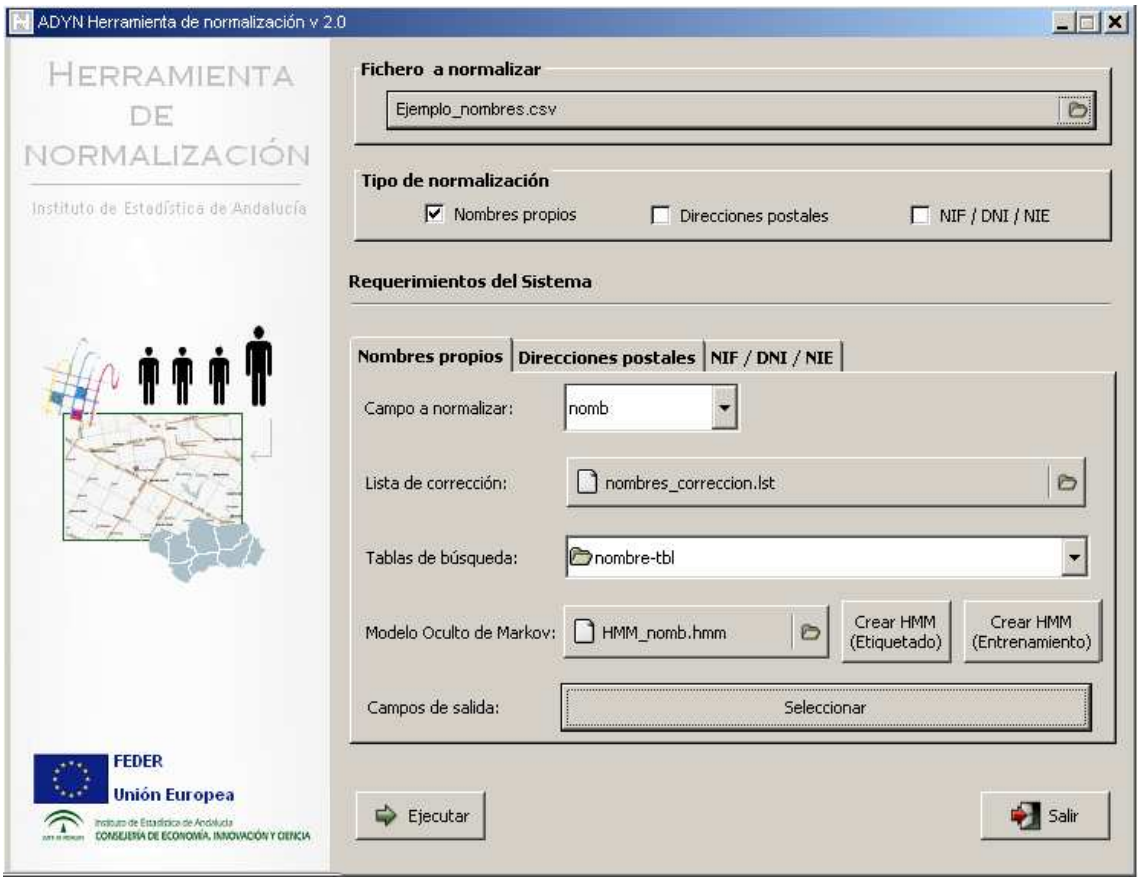


En cuanto al **'Modelo Oculto de Markov'** lo seleccionaremos en esta ruta:

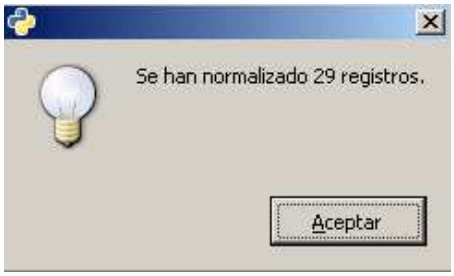


A continuación seleccionaremos los **'Campos de salida'**. Como se observa sólo se han marcado los relativos al nombre propio del individuo:

Tras estas especificaciones la interfaz ha quedado configurada como:

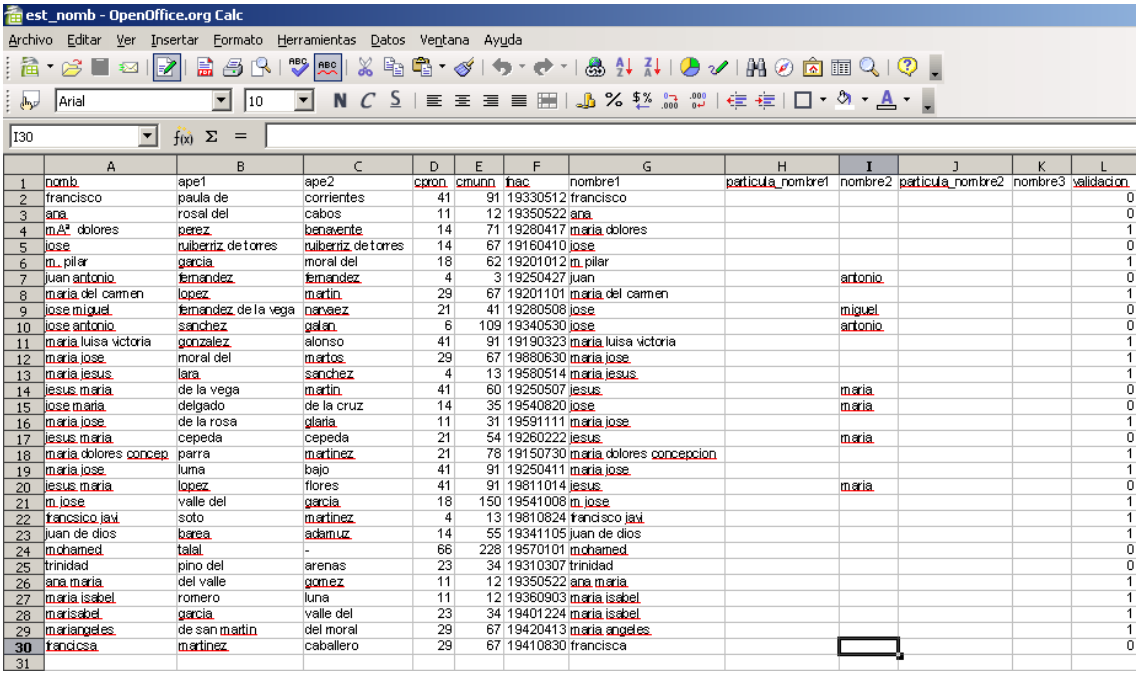


A continuación pulsamos **'Ejecutar'** y esperamos hasta que aparezca el siguiente mensaje que nos indica que el proceso de normalización del campo '*nomb*' ha terminado:



El fichero resultante de la normalización se encuentra en la misma ruta que el fichero de trabajo '*Ejemplo_nombres.csv*' y es el que aparece en la siguiente imagen. Hay que decir que el nombre de este fichero no es el nombre original con el que la aplicación lo

genera, sino que una vez que se ha generado se ha renombrado por comodidad con el nombre '*est_nomb.csv*'.



est_nomb - OpenOffice.org Calc

Archivo Editar Ver Insertar Formato Herramientas Datos Ventana Ayuda

Arial 10

I30

	A	B	C	D	E	F	G	H	I	J	K	L
1	nomb	ape1	ape2	cpron	cmunn	fnac	nombre1	particula_nombre1	nombre2	particula_nombre2	nombre3	validacion
2	francisco	paula de	corrientes	41	91	19330512	francisco					0
3	ana	rosal del	cabos	11	12	19350522	ana					0
4	m.Âª dolores	perez	benavente	14	71	19280417	maria dolores					1
5	jose	ruiberriz de torres	ruiberriz de torres	14	67	19160410	jose					0
6	m. pilar	garcia	moral del	18	62	19201012	m pilar					1
7	juan antonio	fernandez	fernandez	4	3	19250427	juan		antonio			0
8	maria del carmen	lopez	martin	29	67	19201101	maria del carmen					1
9	jose miguel	fernandez de la vega	narvaez	21	41	19280508	jose		miguel			0
10	jose antonio	sanchez	galan	6	109	19340530	jose		antonio			0
11	maria luisa victoria	gonzalez	alonso	41	91	19190323	maria luisa victoria					1
12	maria jose	moral del	martos	29	67	19880630	maria jose					1
13	maria jesus	lara	sanchez	4	13	19580514	maria jesus					1
14	jesus maria	de la vega	martin	41	60	19250507	jesus		maria			0
15	jose maria	delgado	de la cruz	14	35	19540820	jose		maria			0
16	maria jose	de la rosa	glaria	11	31	19591111	maria jose					1
17	jesus maria	cepeda	cepeda	21	54	19260222	jesus		maria			0
18	maria dolores concep	parra	martinez	21	78	19150730	maria dolores concepcion					1
19	maria jose	luna	bajo	41	91	19250411	maria jose					1
20	jesus maria	lopez	flores	41	91	19811014	jesus		maria			0
21	m jose	valle del	garcia	18	150	19541008	m jose					1
22	francsico javi	soto	martinez	4	13	19810824	francisco javi					1
23	juan de dios	barea	adamuz	14	55	19341105	juan de dios					1
24	mohamed	talal	-	66	228	19570101	mohamed					0
25	trinidad	pino del	arenas	23	34	19310307	trinidad					0
26	ana maria	del valle	gomez	11	12	19350522	ana maria					1
27	maria isabel	romero	luna	11	12	19360903	maria isabel					1
28	marisabel	garcia	valle del	23	34	19401224	maria isabel					1
29	mariangeles	de san martin	del moral	29	67	19420413	maria angeles					1
30	francicsa	martinez	caballero	29	67	19410830	francisca					0
31												

3. Validación

Como se puede observar los registros que tienen el valor 1 en la columna 'validacion' están mal normalizados (en total hay 16 registros mal normalizados). Si incluimos las estructuras de estos datos en la muestra que se ha seleccionado previamente a través de la aplicación, el proceso de normalización será más eficiente.

Por ejemplo, para comprobar cómo mejora el proceso de normalización vamos a incluir algunas de las estructuras no representadas en la muestra, en concreto elegimos:

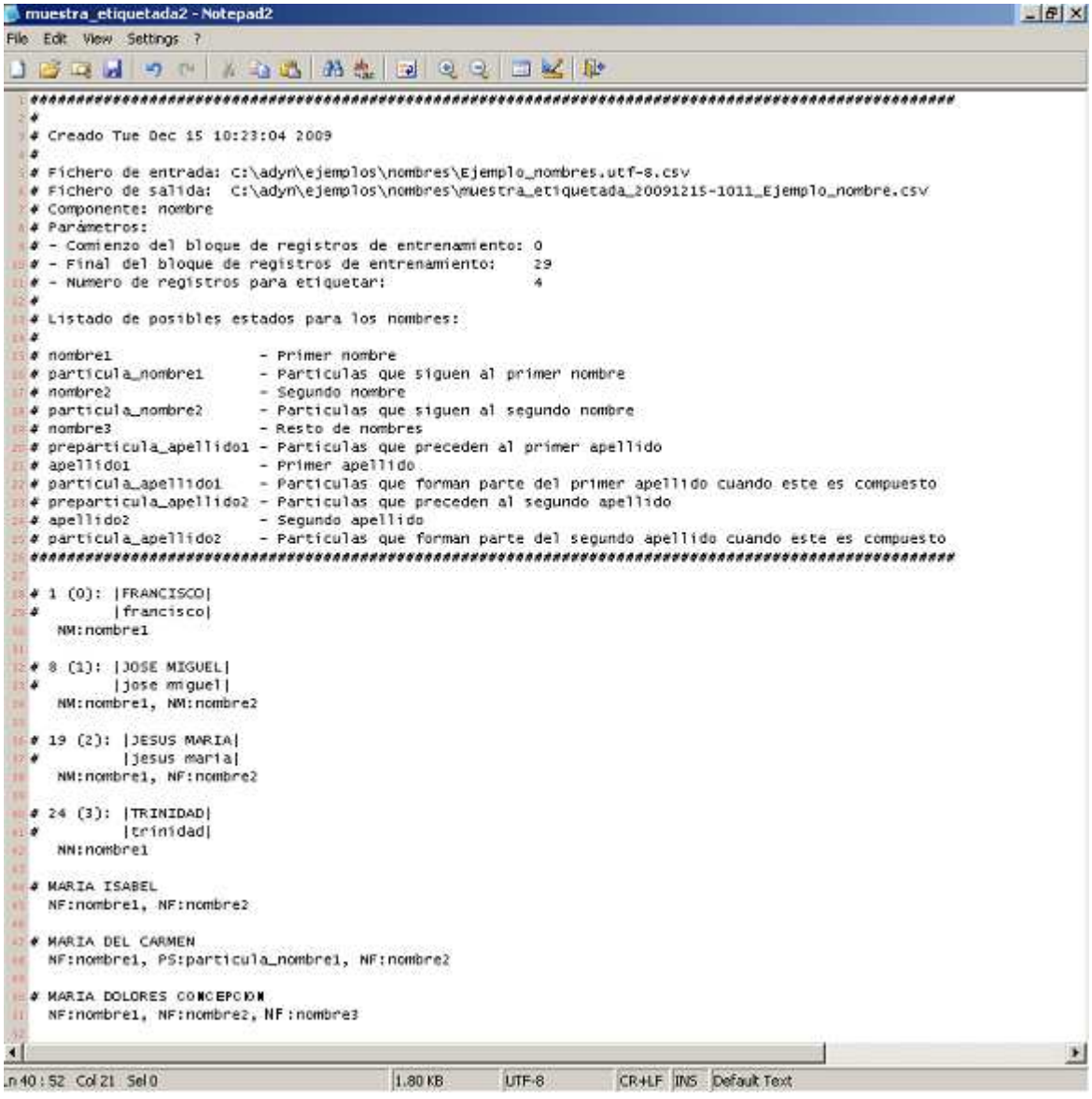
- maria isabel
- maria del carmen
- maria dolores concepcion

Así pues, abriremos el fichero '*muestra_etiquetada_1.csv*' con el editor de texto '**Notepad2**', incluiremos las siguientes estructuras:

```
# MARIA ISABEL
  NF:nombre1, NF:nombre2

# MARIA DEL CARMEN
  NF:nombre1, PS:particula_nombre1, NF:nombre2

# MARIA DOLORES CONCEPCION
  NF:nombre1, NF:nombre2, NF:nombre3
```

y guardaremos los cambios. Al guardar los cambios se ha decidido renombrar el fichero de la muestra etiquetada con el nombre '*muestra_etiquetada2.csv*' pero se podría haber guardado con el mismo nombre.

En la siguiente imagen se muestra cómo ha quedado el nuevo fichero de la muestra etiquetada:



A continuación volveremos a entrenar la nueva muestra etiquetada, es decir, volveremos a crear un nuevo Modelo Oculto de Markov. Este modelo lo usaremos posteriormente para volver a normalizar el campo '*nomb*' del fichero de trabajo '*Ejemplo_nombres.csv*' y ver si el proceso de normalización ha mejorado.

Para ello repetiremos los pasos seguidos en la interfaz ***'03. Entrenamiento de la muestra'*** con el fichero '*muestra_etiquetada2.csv'*. Como resultado obtendremos un nuevo modelo HMM, que renombraremos con el nombre '*HMM_nomb2.hmm'*.

La siguiente fase será volver a normalizar el fichero '*Ejemplo_nombres.csv*' con este nuevo modelo HMM creado, '*HMM_nomb2.hmm'*. El resultado de dicha normalización será el siguiente fichero, al que hemos denominado '*est_nomb2.csv*':

est_nomb2 - OpenOffice.org Calc

Archivo Editar Ver Insertar Formato Herramientas Datos Ventana Ayuda

Arial 10

I33

	A	B	C	D	E	F	G	H	I	J	K	L
1	nomb	ape1	ape2	cpron	cmunn	fnac	nombre1	particula_nombre1	nombre2	particula_nombre2	nombre3	validacion
2	francisco	paula de	corrientes	41	91	19330512	francisco					0
3	ana	rosal del	cabos	11	12	19350522	ana					0
4	mÃ³ dolores	perez	benavente	14	71	19280417	maria		dolores			0
5	jose	ruiberriz de torres	ruiberriz de torres	14	67	19160410	jose					0
6	m. pilar	garcia	moral del	18	62	19201012	m pilar					1
7	juan antonio	fernandez	fernandez	4	3	19250427	juan		antonio			0
8	maria del carmen	lopez	martin	29	67	19201101	maria	del	carmen			0
9	jose miguel	fernandez de la vega	narvaez	21	41	19280508	jose		miguel			0
10	jose antonio	sanchez	galan	6	109	19340530	jose		antonio			0
11	maria luisa victoria	gonzalez	alonso	41	91	19190323	maria		luisa		victoria	0
12	maria jose	moral del	martos	29	67	19880630	maria		jose			0
13	maria jesus	lara	sanchez	4	13	19580514	maria		jesus			0
14	jesus maria	de la vega	martin	41	60	19250507	jesus		maria			0
15	jose maria	delgado	de la cruz	14	35	19540820	jose		maria			0
16	maria jose	de la rosa	glaria	11	31	19591111	maria		jose			0
17	jesus maria	cepeda	cepeda	21	54	19260222	jesus		maria			0
18	maria dolores concep	parra	martinez	21	78	19150730	maria		dolores		concepcion	0
19	maria jose	luna	bajo	41	91	19250411	maria		jose			0
20	jesus maria	lopez	flores	41	91	19811014	jesus		maria			0
21	m jose	valle del	garcia	18	150	19541008	m jose					1
22	francsico javi	soto	martinez	4	13	19810824	francisco javi					1
23	juan de dios	barea	adamuz	14	55	19341105	juan de dios					1
24	mohamed	talal	-	66	228	19570101	mohamed					0
25	trinidad	pino del	arenas	23	34	19310307	trinidad					0
26	ana maria	del valle	gomez	11	12	19350522	ana		maria			0
27	maria isabel	romero	luna	11	12	19360903	maria		isabel			0
28	marisabel	garcia	valle del	23	34	19401224	maria		isabel			0
29	mariangeles	de san martin	del moral	29	67	19420413	maria		angeles			0
30	francicsa	martinez	caballero	29	67	19410830	francisca					0
31												

Como se puede observar el número de registros que están mal normalizados con el nuevo modelo HMM que se ha creado ha disminuido, pasando de 15 registros mal normalizados a 4. Además todos aquellos registros que presentan estructuras similares a las que hemos incluido en la muestra aparecen correctamente normalizados (por ejemplo, ver en la imagen anterior el registro número 26 que tiene una estructura similar al 27).

Se puede comprobar que aún quedan registros mal normalizados con lo cual habría que incluir estas estructuras en el fichero '*muestra_etiquetada2.csv*' y realizaríamos el mismo proceso de enriquecimiento que antes y así seguiríamos sucesivamente hasta tener el campo '*nomb*' perfectamente normalizado.

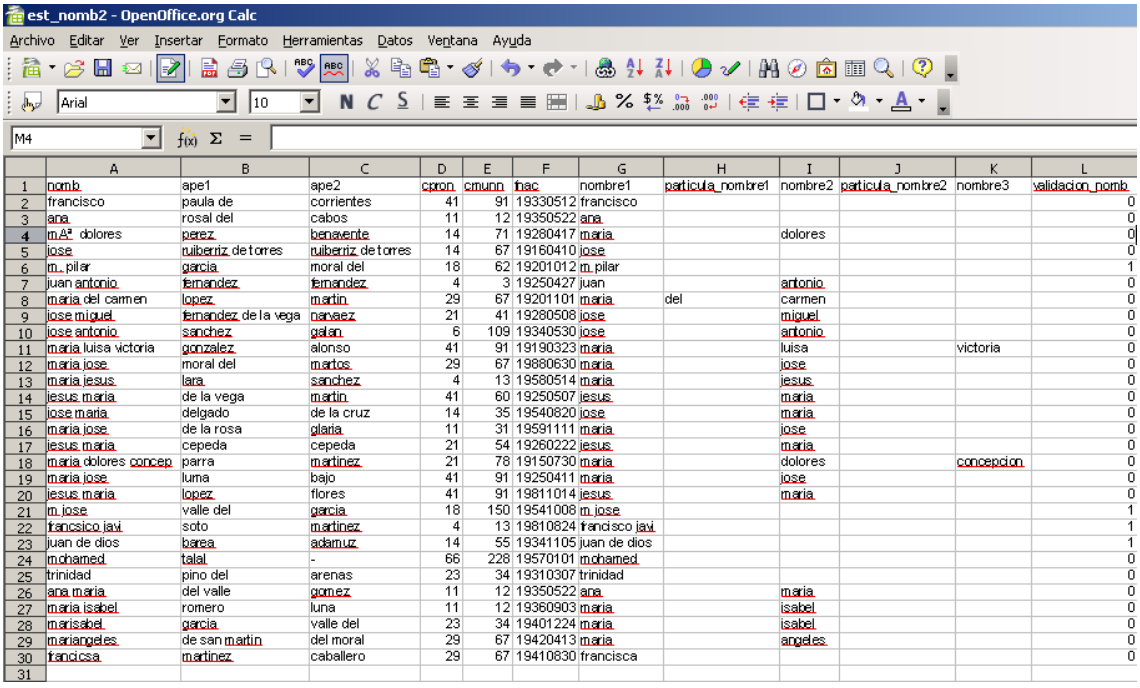
A continuación pasaríamos a normalizar el campo '*ape1*' pero antes de iniciar este proceso es necesario realizar dos observaciones a tener en cuenta cuando normalizamos campos que están segmentados en varias columnas:

1. **Campos de salida**. En la interfaz ***'01. Normalización***' sólo se deberían seleccionar los campos de salida relativos al campo que se normaliza pero si el usuario no hubiese tenido en cuenta esta precaución y hubiese dejado marcados todos los campos (como aparecen por defecto), sería necesario eliminar del fichero normalizado los que no están referidos al campo a normalizar o bien cambiar la

denominación de todos los campos de salida que se muestran, ya que el fichero normalizado será el de partida para realizar la normalización del siguiente campo. Si dejáramos la denominación de los campos de salida tal cual aparecen o no se eliminaran del fichero normalizado habría problemas de compatibilidad entre los nombres de los campos de salida y no se podría llevar a cabo el proceso de normalización.

2. **Campo 'validacion'**. En este caso el problema es el mismo, la incompatibilidad de campos con la misma denominación. Por ejemplo, si como en este caso en primer lugar se normaliza el campo nombre propio, el usuario debe cambiar la denominación de la columna 'validacion' del fichero normalizado '*est_nomb2.csv*' o eliminarla si considera que ésta ya no le es necesaria, para así evitar problemas de compatibilidad con la denominación del campo 'validacion' que también aparecerá al normalizar el siguiente campo, que en este caso sería '*ape1*'.

Como se ha podido comprobar en este caso práctico se ha decidido seleccionar sólo los campos de salida referidos a nombres propios y renombrar la columna 'validacion' con el nombre 'validacion_nomb' con lo que el fichero con el campo '*nomb*' normalizado ha quedado de la siguiente forma:



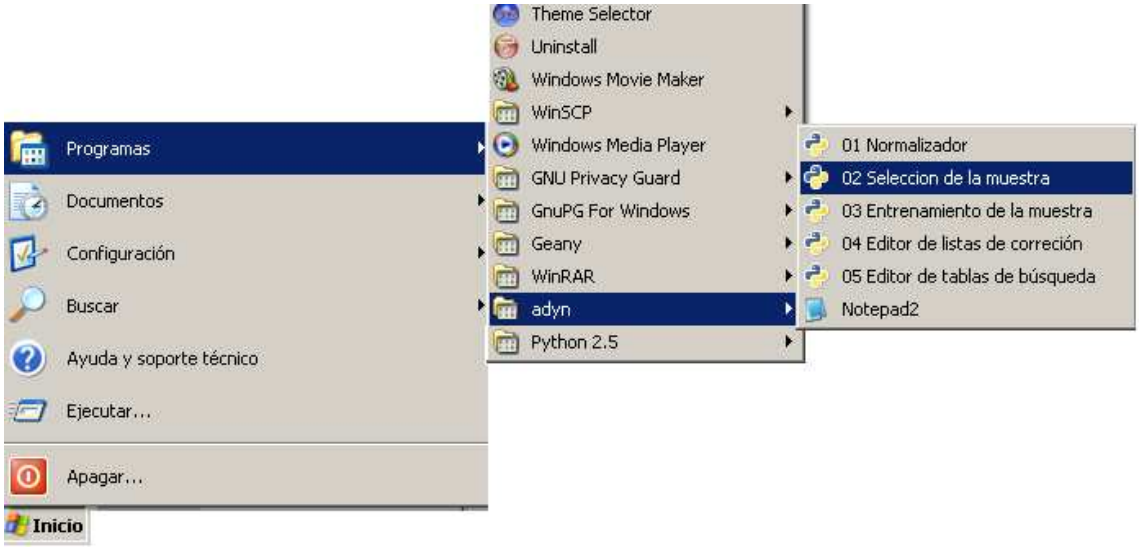
	A	B	C	D	E	F	G	H	I	J	K	L
1	nomb	ape1	ape2	cpron	cmunn	fnac	nombre1	particula_nombre1	nombre2	particula_nombre2	nombre3	validacion_nomb
2	francisco	paula de	corrientes	41	91	19330512	francisco					0
3	ana	rosal del	cabos	11	12	19350522	ana					0
4	mÃª dolores	perez	benavente	14	71	19280417	maria		dolores			0
5	jose	ruiberriz de torres	ruiberriz de torres	14	67	19160410	jose					0
6	m. pilar	garcia	moral del	18	62	19201012	m.pilar					1
7	juan antonio	fernandez	fernandez	4	3	19250427	juan		antonio			0
8	maria del carmen	lopez	martin	29	67	19201101	maria	del	carmen			0
9	jose miguel	fernandez de la vega	narvaez	21	41	19280508	jose		miguel			0
10	jose antonio	sanchez	galan	6	109	19340530	jose		antonio			0
11	maria luisa victoria	gonzalez	alonso	41	91	19190323	maria		luisa		victoria	0
12	maria jose	moral del	martos	29	67	19880630	maria		jose			0
13	maria jesus	lara	sanchez	4	13	19580514	maria		jesus			0
14	jesus maria	de la vega	martin	41	60	19250507	jesus		maria			0
15	jose maria	delgado	de la cruz	14	35	19540820	jose		maria			0
16	maria jose	de la rosa	glaria	11	31	19591111	maria		jose			0
17	jesus maria	cepeda	cepeda	21	54	19260222	jesus		maria			0
18	maria dolores concep	parra	martinez	21	78	19150730	maria		dolores		concepcion	0
19	maria jose	luna	bajo	41	91	19250411	maria		jose			0
20	jesus maria	lopez	flores	41	91	19811014	jesus		maria			0
21	m jose	valle del	garcia	18	150	19541008	m jose					1
22	francsico javi	soto	martinez	4	13	19810824	francsico javi					1
23	juan de dios	barea	adamuz	14	55	19341105	juan de dios					1
24	mohamed	talal	-	66	228	19570101	mohamed					0
25	trinidad	pino del	arenas	23	34	19310307	trinidad					0
26	ana maria	del valle	gomez	11	12	19350522	ana		maria			0
27	maria isabel	romero	luna	11	12	19360903	maria		isabel			0
28	marisabel	garcia	valle del	23	34	19401224	maria		isabel			0
29	mariangeles	de san martin	del moral	29	67	19420413	maria		angeles			0
30	francicsa	martinez	caballero	29	67	19410830	francisca					0
31												

PROCESO DE NORMALIZACIÓN DEL CAMPO '*ape1*'

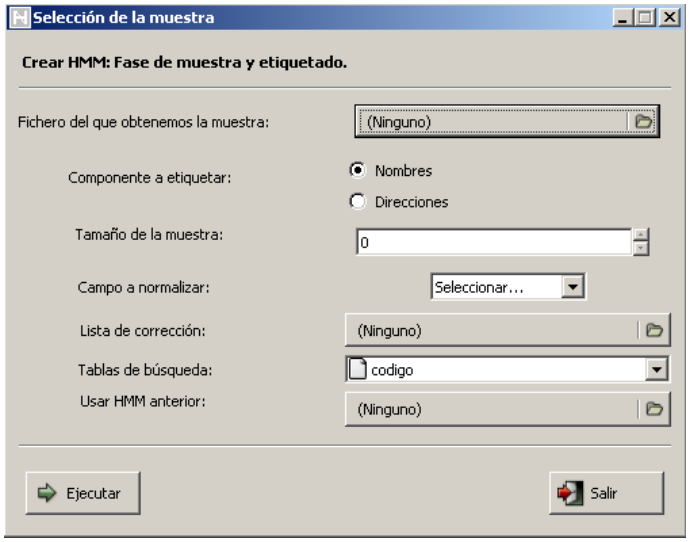
1. Creación del Modelo Oculto de Markov

1.1. Selección de la muestra

Para seleccionar la muestra accedemos a la interfaz a través de (si trabajamos con Windows):

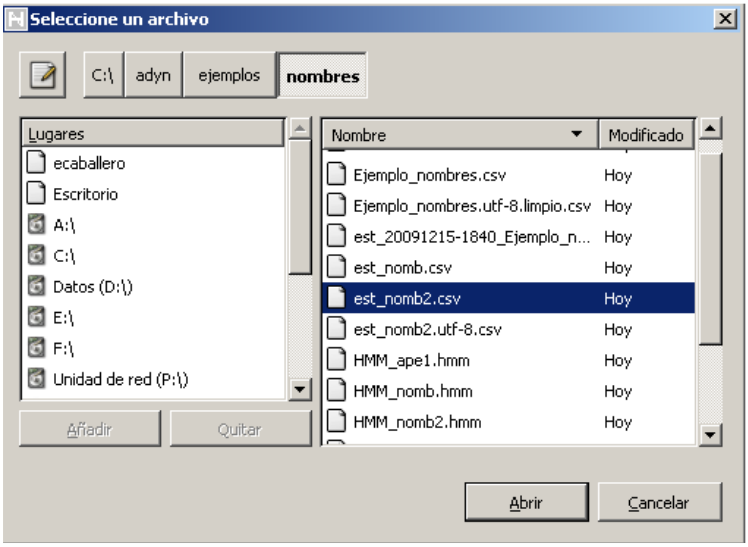


Y nos aparecerá la siguiente ventana:

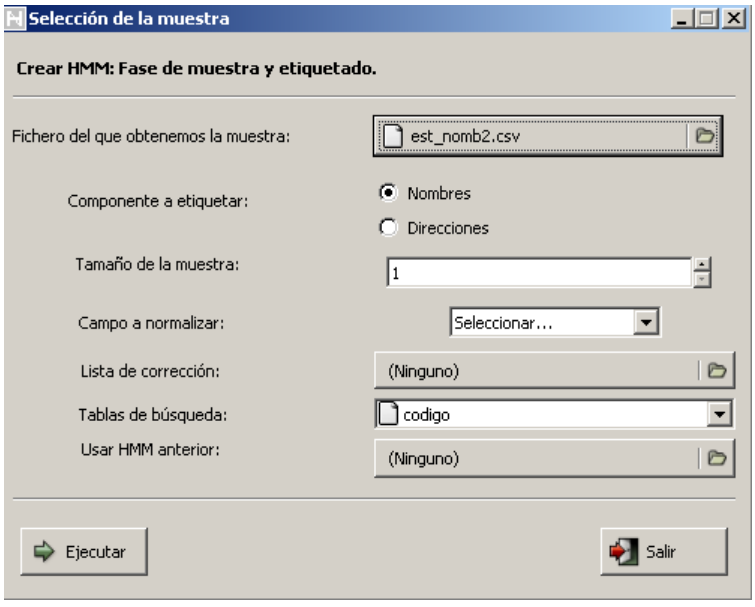


Seleccionamos el **'Fichero del que obtenemos la muestra'**, en nuestro caso: '*est_nomb2.csv*', que es el fichero con el campo '*nomb*' normalizado. La ruta de este fichero es la misma que la del fichero original de datos, es decir,

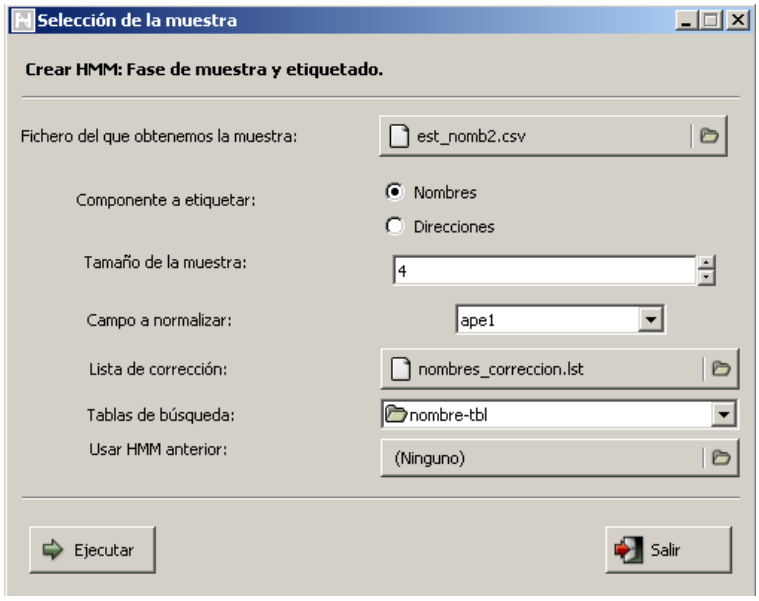
'C:\adyn\ejemplos\nombres'. La siguiente ventana muestra cómo se llega al fichero que vamos a utilizar:



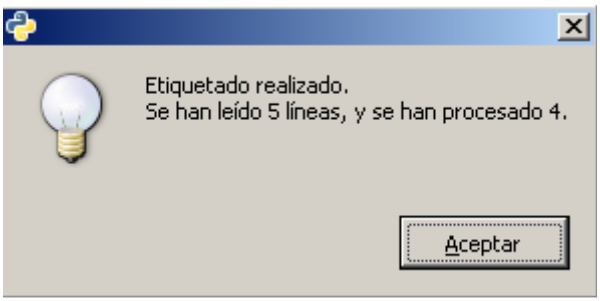
A continuación seleccionamos la **'Componente a etiquetar'**, que en nuestro caso es 'Nombres' y es la que está seleccionada por defecto en la aplicación:



Los siguientes pasos serán especificar el **'Tamaño de la muestra'** a seleccionar, el **'Campo a normalizar'**, la **'Lista de corrección'**, las **'Tablas de búsqueda'** y la opción **'Usar HMM anterior'.** Para ello procederemos como en el caso del campo '*nomb*' solo que ahora en el cuadro **'Campo a normalizar'** elegiremos la opción '*ape1*' . De esta forma la interfaz queda configurada como:

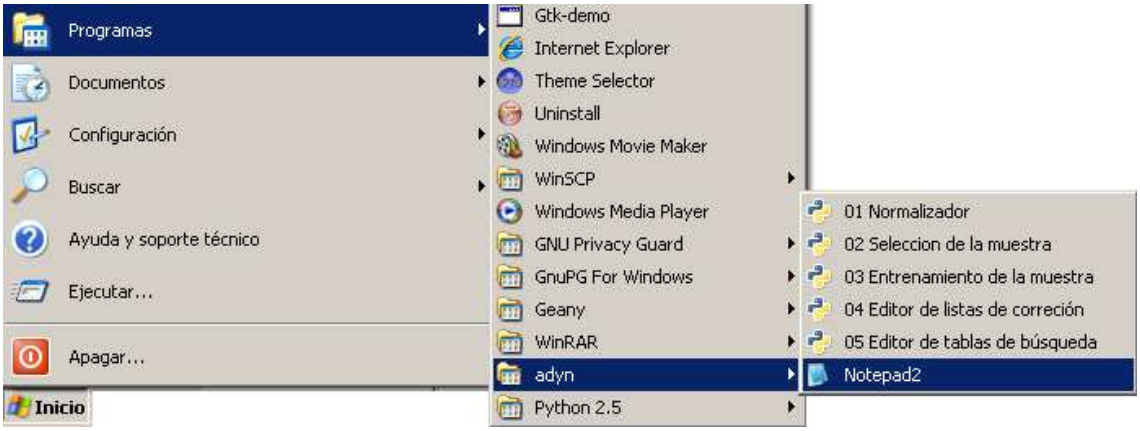


Ahora sólo tenemos que pulsar **'Ejecutar'** y esperaremos a que la aplicación nos indique que el proceso ha finalizado mediante el siguiente mensaje:

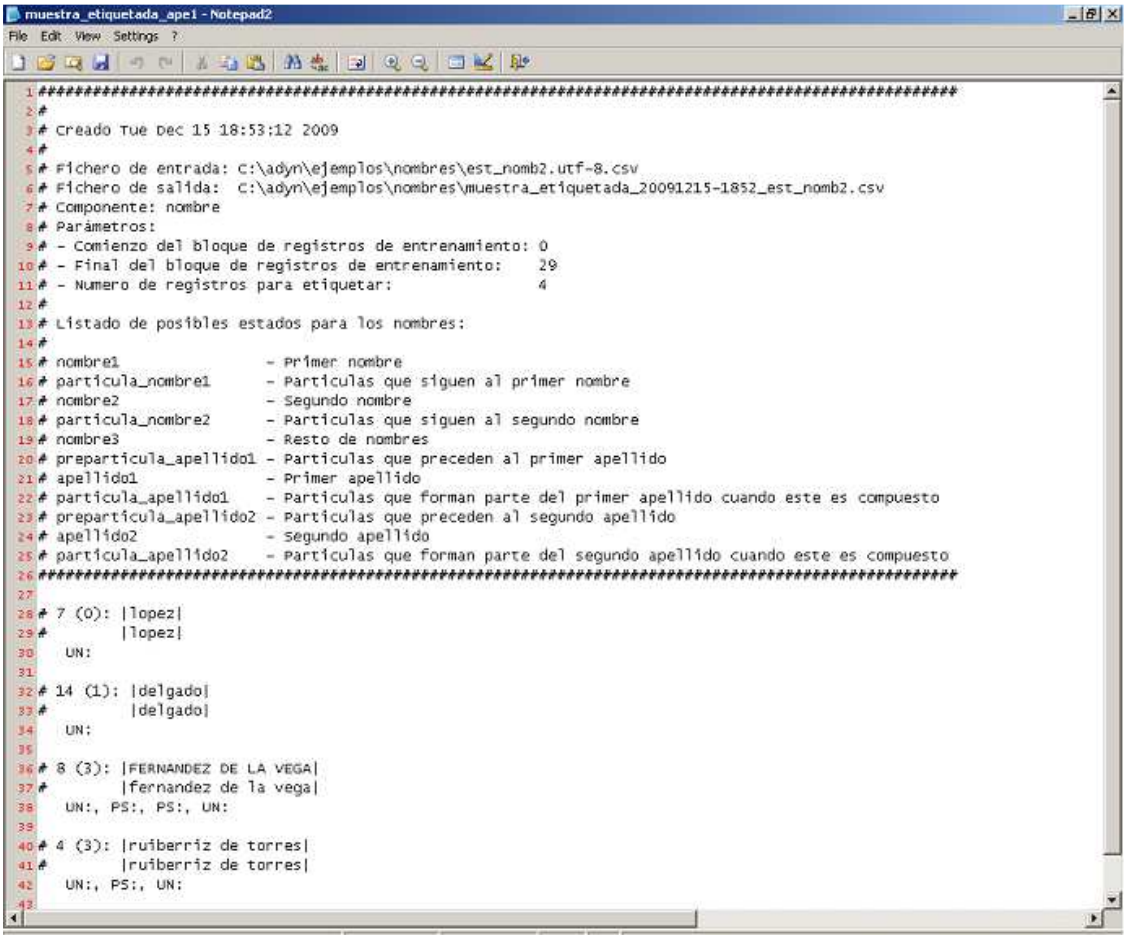


Se puede observar que en este caso el número de líneas que se han leído aleatoriamente son 5, es decir, de todas las líneas del campo a normalizar '*ape1*' (29) sólo se han leído aleatoriamente 5 y de ellas se han procesado 4, es decir, se ha extraído una muestra de cuatro registros.

El fichero de la muestra etiquetada que hemos obtenido para este campo se ha renombrado como '*muestra_etiquetada_ape1.csv*' y se encuentra en el mismo lugar que el fichero de datos original '*Ejemplo_nombres.csv*'. Para ver su contenido utilizaremos el editor de texto **'Notepad2'** que se incluye en la aplicación *ADYN* y al que se puede acceder (si se trabaja con Windows) desde:



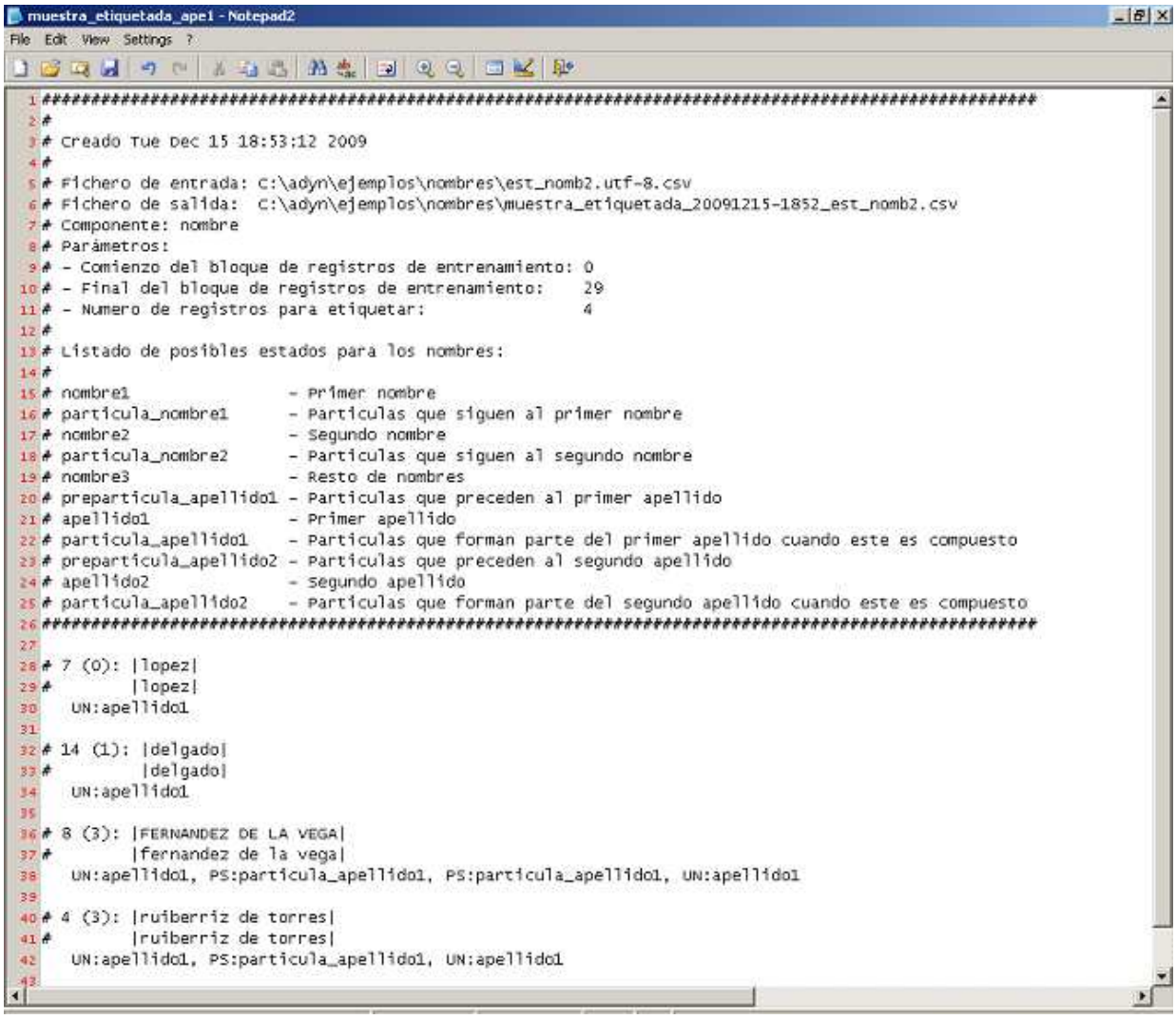
Así pues, el contenido del fichero de la muestra etiquetada para el primer apellido es:



1.2. Asignación manual de estados

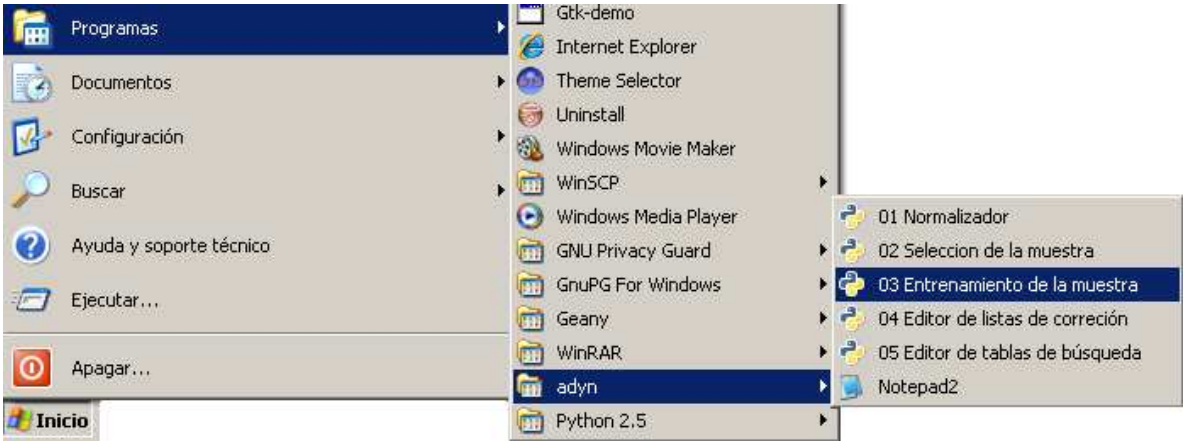
El siguiente paso será asignar manualmente estados a estas etiquetas. Para ello editamos el fichero anterior con el editor de texto '**Notepad2**'.

A continuación iremos asignando a cada etiqueta el estado que consideramos adecuado. Esta asignación se ha realizado según el conocimiento que se tiene sobre el fichero de trabajo y nuestra experiencia, de tal forma que el fichero '*muestra_etiquetada_ape1.csv*' quedaría de la forma:



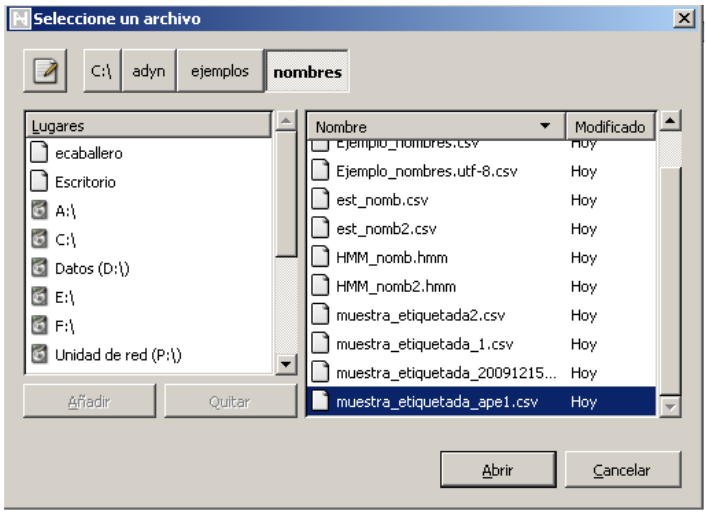
1.3. Entrenamiento de la muestra

A continuación utilizaremos el fichero anterior para llevar a cabo el entrenamiento de la muestra. Para acceder a la interfaz de entrenamiento procedemos como se muestra a continuación (si trabajamos con Windows):

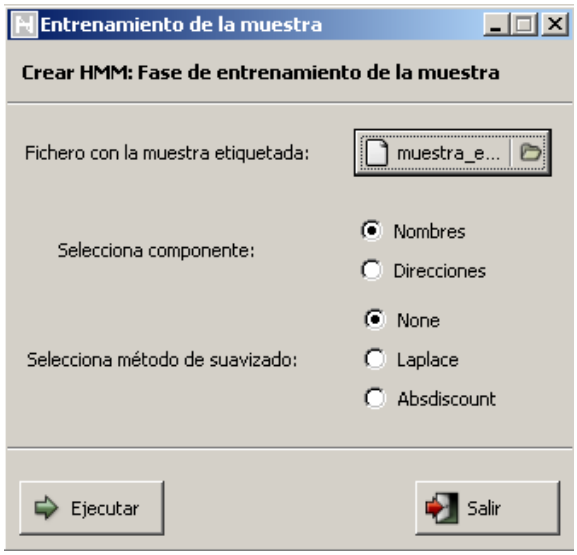


La pantalla que nos recibe es la siguiente:

En el campo **'Fichero con la muestra etiquetada'** seleccionamos el fichero '*muestra_etiquetada_ape1.csv*':

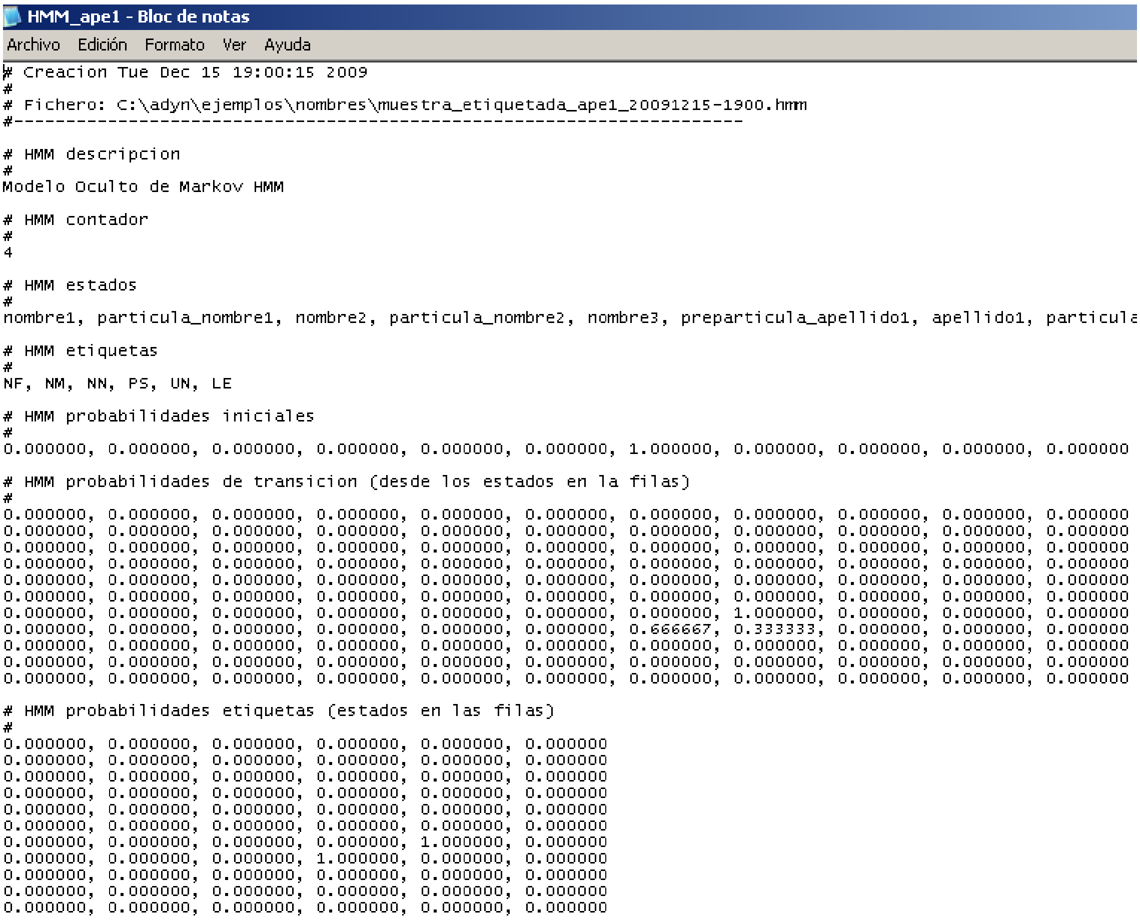


Posteriormente en **'Selecciona componente'** marcamos la componente a normalizar, en nuestro caso 'Nombres' y no elegiremos ningún método de suavizado.



A continuación pulsamos **'Ejecutar'** y esperamos a que finalice el proceso de entrenamiento, que sabremos que ha terminado cuando se muestre el siguiente mensaje en pantalla:

Como resultado de este proceso obtenemos un fichero de extensión '*.hmm*' al que por comodidad lo denominaremos '*HMM_ape1.hmm*'. Dicho fichero se encuentra en el mismo directorio que el fichero de trabajo '*Ejemplo_nombres.csv*' y su contenido se muestra en la siguiente imagen:

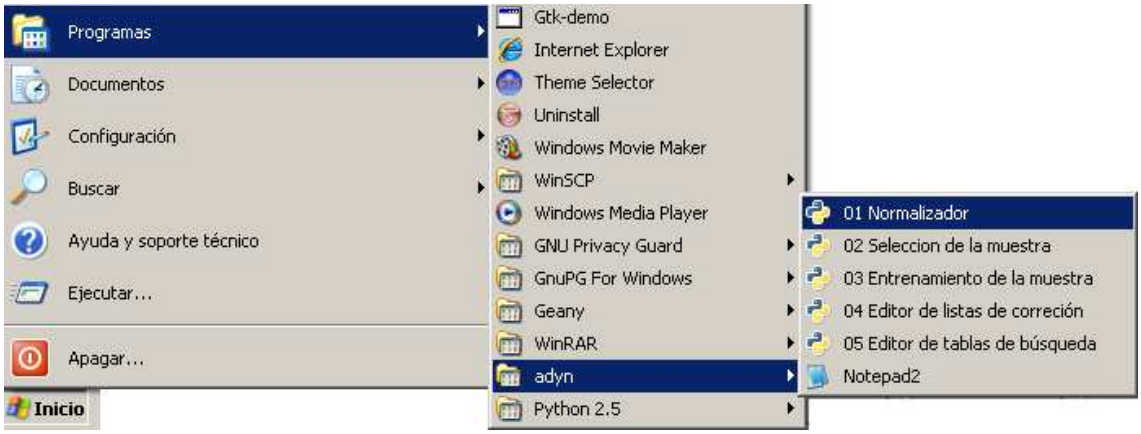


```
HMM_ape1 - Bloc de notas
Archivo  Edición  Formato  Ver  Ayuda
# Creacion Tue Dec 15 19:00:15 2009
#
# Fichero: C:\adyn\ejemplos\nombres\muestra_etiquetada_ape1_20091215-1900.hmm
#-----------------------------------------------------------------------

# HMM descripcion
#
Modelo Oculto de Markov HMM

# HMM contador
#
4

# HMM estados
#
nombre1, particula_nombre1, nombre2, particula_nombre2, nombre3, preparticula_apellido1, apellido1, particula

# HMM etiquetas
#
NF, NM, NN, PS, UN, LE

# HMM probabilidades iniciales
#
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 1.000000, 0.000000, 0.000000, 0.000000, 0.000000

# HMM probabilidades de transicion (desde los estados en la filas)
#
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 1.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.666667, 0.333333, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000

# HMM probabilidades etiquetas (estados en las filas)
#
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 1.000000, 0.000000
0.000000, 0.000000, 0.000000, 1.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000
```

Este fichero será el Modelo Oculto de Markov que utilizaremos para estandarizar y segmentar el campo **'***ape1***'** del fichero de trabajo '*est_nomb2.csv*'.

2. Normalización del campo '*ape1*'

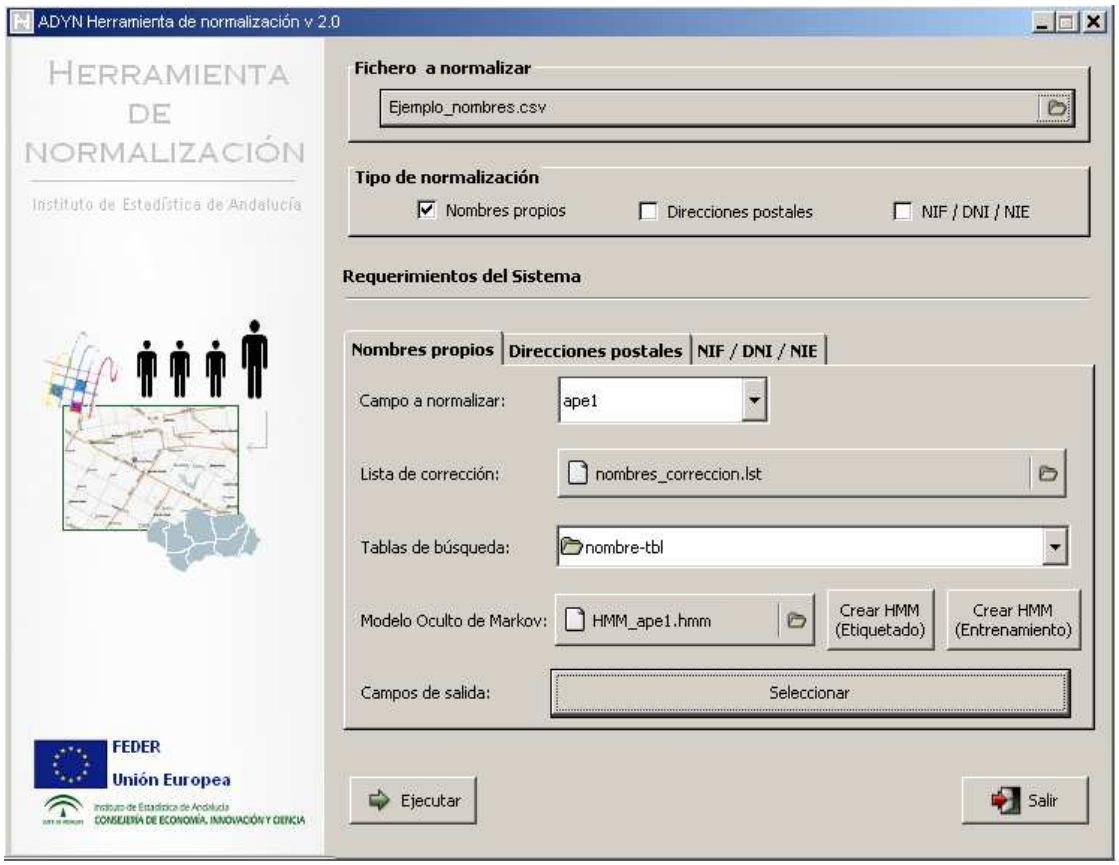
Una vez que tenemos creado el Modelo Oculto de Markov, para llevar a cabo la normalización del campo primer apellido ('*ape1*') accederemos a la interfaz ***'01. Normalizador'*** de la siguiente forma (si trabajamos con Windows):



En la pantalla que nos recibe especificamos los siguientes parámetros que se muestran en la imagen:

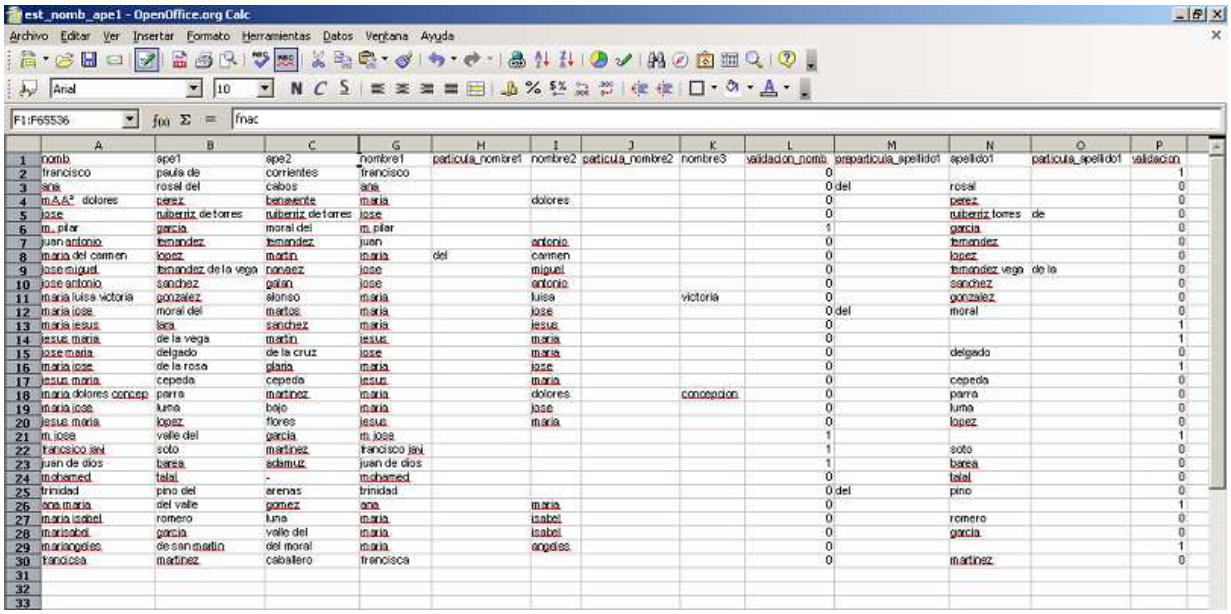
Como se puede ver sólo se han seleccionado los campos de salida referidos al primer apellido para evitar incompatibilidades con las denominaciones de los campos del siguiente campo a normalizar '*ape2*'.

Así pues, la interfaz ha quedado configurada de la siguiente forma:



A continuación pulsamos **'Ejecutar'** y esperamos hasta que aparezca el siguiente mensaje que nos indica que el proceso de normalización del campo '*ape1*' ha terminado:

El fichero con el campo '*ape1*' normalizado lo hemos denominado '*est_nomb_ape1.csv*' y se muestra en la siguiente imagen:



	A	B	C	G	H	I	J	K	L	M	N	O	P
1	nomb	ape1	ape2	nombre1	particula_nombre1	nombre2	particula_nombre2	nombre3	validacion_nomb	preparticula_apellido1	apellido1	particula_apellido1	validacion
2	francisco	paula de	corrientes	francisco					0				1
3	ana	rosal del	cabos	ana					0	del	rosal		0
4	mÃ¡ dolores	perez	benavente	maria		dolores			0		perez		0
5	jose	ruiberriz de torres	ruiberriz de torres	jose					0		ruiberriz torres	de	0
6	m. pilar	garcia	moral del	m. pilar					1		garcia		0
7	juan antonio	fernandez	fernandez	juan		antonio			0		fernandez		0
8	maria del carmen	lopez	martin	maria	del	carmen			0		lopez		0
9	jose miguel	fernandez de la vega	narvaez	jose		miguel			0		fernandez vega	de la	0
10	jose antonio	sanchez	galan	jose		antonio			0		sanchez		0
11	maria luisa victoria	gonzalez	alonso	maria		luisa		victoria	0		gonzalez		0
12	maria jose	moral del	martos	maria		jose			0	del	moral		0
13	maria jesus	lara	sanchez	maria		jesus			0				1
14	jesus maria	de la vega	martin	jesus		maria			0				1
15	jose maria	delgado	de la cruz	jose		maria			0		delgado		0
16	maria jose	de la rosa	plana	maria		jose			0				1
17	jesus maria	cepeda	cepeda	jesus		maria			0		cepeda		0
18	maria dolores concep	parra	martinez	maria		dolores		concepcion	0		parra		0
19	maria jose	luna	bajo	maria		jose			0		luna		0
20	jesus maria	lopez	flores	jesus		maria			0		lopez		0
21	m jose	valle del	garcia	m jose					1				1
22	francisco javi	soto	martinez	francisco javi					1		soto		0
23	juan de dios	barea	adamuz	juan de dios					1		barea		0
24	mohamed	talal	-	mohamed					0		talal		0
25	trinidad	pino del	arenas	trinidad					0	del	pino		0
26	ana maria	del valle	gomez	ana		maria			0				1
27	maria isabel	romero	luna	maria		isabel			0		romero		0
28	marisabel	garcia	valle del	maria		isabel			0		garcia		0
29	mariangeles	de san martin	del moral	maria		angeles			0				1
30	francisca	martinez	caballero	francisca					0		martinez		0
31													
32													
33													

Notar que se han ocultado las columnas del fichero original de datos: 'cmunn', 'cpron' y 'fnac' para poder mostrar los resultados en la imagen anterior.

3. Validación

Como se puede observar en la imagen anterior los apellidos cuya estructura o patrón de datos no están representados en la muestra seleccionada ('*muestra_etiquetada_ape1.csv*') no están bien normalizados, de hecho no se muestran ni en el campo de salida 'apellido1' (ver registros número 2 y 13 por ejemplo).

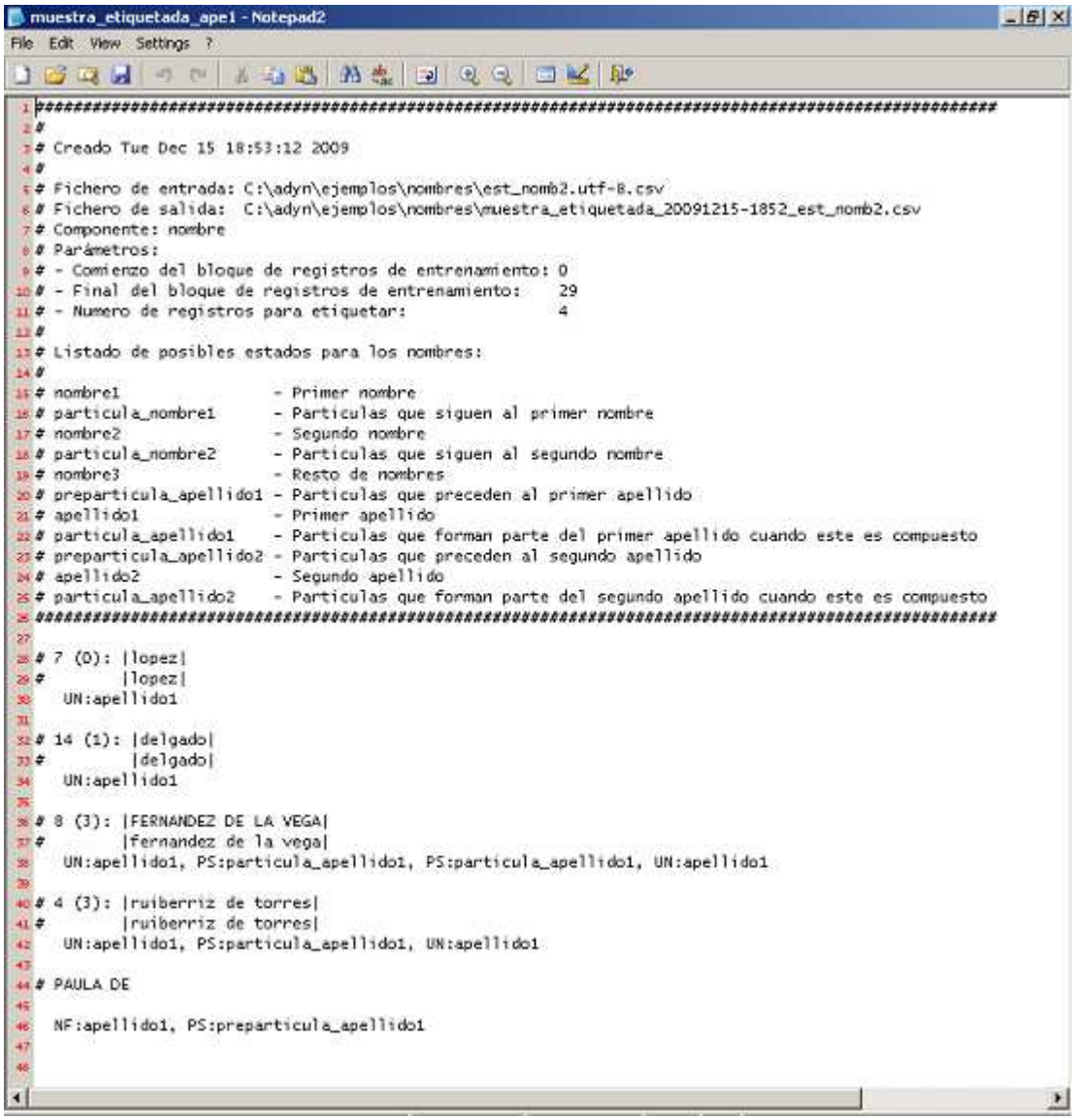
Si hubiésemos cambiado la denominación de los campos de salida del fichero con el campo '*nomb*' normalizado y hubiésemos marcado todos los campos de salida al normalizar el campo '*ape1*' entonces se podría observar que la información referida al primer apellido se encuentra recogida en el campo de salida 'nombre1'.

Para corregir los errores de normalización actuaremos de igual forma que en el caso de la normalización del campo '*nomb*'. Es decir, tendríamos que abrir el fichero '*muestra_etiquetada_ape1.csv*' con '**Notepad2**' e incluir las estructuras de los registros mal normalizados ya que no están representadas en ella.

Por ejemplo supongamos que en el fichero '*muestra_etiquetada_ape1.csv*' incluimos la estructura correspondiente a un apellido del tipo '*paula de*', es decir:

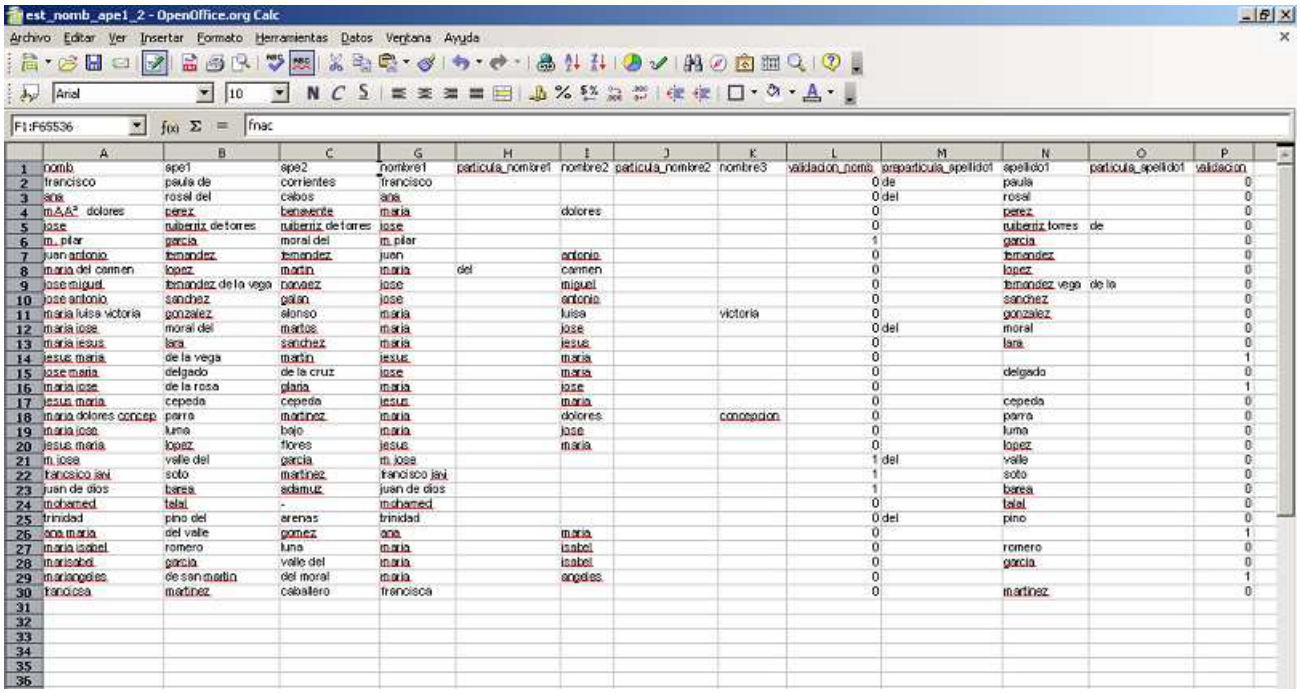
NF:apellido1, PS:preparticula_apellido1

Entonces el dicho fichero quedaría de la forma:



Con este fichero entrenamos de nuevo la muestra a través de la interfaz ***'03. Entrenamiento de la muestra'*** y el modelo HMM obtenido en este proceso lo utilizaremos para normalizar de nuevo el campo '*ape1*' del fichero '*est_nomb2.csv*', fichero que ya tiene normalizado el campo nombre propio.

El resultado de la normalización del campo '*ape1*' con este nuevo modelo HMM creado es el siguiente fichero al que hemos denominado '*est_nomb_ape1_2.csv*':



	A	B	C	G	H	I	J	K	L	M	N	O	P
1	nomb	ape1	ape2	nombre1	particula_nombre1	nombre2	particula_nombre2	nombre3	validacion_nomb	preparticula_apellido1	apellido1	particula_apellido1	validacion
2	francisco	paula de	corrientes	francisco					0	de	paula		0
3	ana	rosal del	cabos	ana					0	del	rosal		0
4	mÃ¡Â² dolores	perez	benavente	maria		dolores			0		perez		0
5	jose	ruiberriz de torres	ruiberriz de torres	jose					0		ruiberriz torres	de	0
6	m. pilar	garcia	moral del	m pilar					1		garcia		0
7	juan antonio	fernandez	fernandez	juan		antonio			0		fernandez		0
8	maria del carmen	lopez	martin	maria	del	carmen			0		lopez		0
9	jose miguel	fernandez de la vega	narvaez	jose		miguel			0		fernandez vega	de la	0
10	jose antonio	sanchez	galan	jose		antonio			0		sanchez		0
11	maria luisa victoria	gonzalez	alonso	maria		luisa		victoria	0		gonzalez		0
12	maria jose	moral del	martos	maria		jose			0	del	moral		0
13	maria jesus	lara	sanchez	maria		jesus			0		lara		0
14	jesus maria	de la vega	martin	jesus		maria			0				1
15	jose maria	delgado	de la cruz	jose		maria			0		delgado		0
16	maria jose	de la rosa	glaria	maria		jose			0				1
17	jesus maria	cepeda	cepeda	jesus		maria			0		cepeda		0
18	maria dolores concep	parra	martinez	maria		dolores		concepcion	0		parra		0
19	maria jose	luna	bajo	maria		jose			0		luna		0
20	jesus maria	lopez	flores	jesus		maria			0		lopez		0
21	m jose	valle del	garcia	m jose					1	del	valle		0
22	francisco javi	soto	martinez	francisco javi					1		soto		0
23	juan de dios	barea	adamuz	juan de dios					1		barea		0
24	mohamed	talal	-	mohamed					0		talal		0
25	trinidad	pino del	arenas	trinidad					0	del	pino		0
26	ana maria	del valle	gomez	ana		maria			0				1
27	maria isabel	romero	luna	maria		isabel			0		romero		0
28	marisabel	garcia	valle del	maria		isabel			0		garcia		0
29	mariangeles	de san martin	del moral	maria		angeles			0				1
30	francisca	martinez	caballero	francisca					0		martinez		0
31													
32													
33													
34													
35													
36													

En la imagen anterior se puede observar que los registros que tienen estructuras similares a la que hemos incluido en la muestra '*muestra_etiquetada_ape1.csv*' ya están bien normalizados (ver registro número 1 y 21).

PROCESO DE NORMALIZACIÓN DEL CAMPO '*ape2*'

Este proceso es análogo al seguido para el caso de la normalización del campo '*ape1*', solo que en este caso el fichero del que partimos para llevar a cabo la normalización del campo '*ape2*' es '*est_nomb_ape1_2.csv*', es decir, el fichero que tiene los campos nombre propio y primer apellido normalizados. Por tanto, renombraremos el campo '*validacion*' del fichero '*est_nomb_ape1_2.csv*', por ejemplo con el nombre '*validacion_ape1*', y realizaremos el proceso de normalización del campo '*ape2*' con la misma metodología descrita para los campos '*nomb*' y '*ape1*'.

Anexo II: Normalización del campo identificador de persona física o jurídica.

Cuando disponemos de un fichero de datos con información relativa a personas es muy probable que éste contenga variables que las identifiquen físicamente como por ejemplo, el documento nacional de identidad (DNI), número de pasaporte o en el caso de que el individuo carezca de nacionalidad española, el número de identificación de extranjero (NIE). En teoría, estos identificadores deberían determinar unívocamente a cada individuo pero en realidad puede que esto no suceda porque estas variables contengan errores.

Por otro lado, también es posible trabajar con ficheros de datos que contengan información tributaria, que contengan datos personales junto con información tributaria, etc.; la casuística puede ser muy variada. En esta situación lo normal sería encontrarnos con datos que identifiquen inequívocamente tanto a personas físicas como jurídicas en sus relaciones de naturaleza o con trascendencia tributaria, como por ejemplo, el número de identificación fiscal (NIF). La importancia de esta información resulta vital a la hora de tener identificadas a todas aquellas personas físicas o jurídicas susceptibles de tributar y es por este motivo por lo que se considera necesario disponer de información depurada y corregida para conseguir una mejor calidad de ésta.

Así pues, dada la relevancia de los identificadores de personas físicas y jurídicas parece fundamental llevar a cabo un proceso de normalización antes de iniciar algún tipo de análisis o estudio con ellos.

Es por este motivo por el que se ha decidido implementar en *ADYN Herramienta de Normalización* un nuevo módulo que realice este proceso. Con el vamos a poder normalizar datos que identifican personalmente a un individuo como puede ser el DNI o el NIE, así como datos que identifican tributariamente tanto a personas físicas como jurídicas, es decir, el NIF.

Normalización del DNI ó NIE

A pesar de que los formatos del documento nacional de identidad (DNI) y del número de identificación de extranjero (NIE) son distintos se puede observar que ambos identificadores constan de un carácter de control que comprueba si los caracteres de identificación que componen estos documentos son correctos. Sus formatos son los siguientes:

- **DNI**: está compuesto por ocho caracteres numéricos más un carácter de control.
- **NIE**: está integrado por nueve caracteres con la siguiente composición:
  - una letra inicial, que será la X.
  - siete dígitos o caracteres numéricos.
  - un código o carácter de verificación alfabético.

  Y una vez agotada la serie numérica correspondiente a la letra X, se continuará siguiendo el orden alfabético, es decir, con la Y y así sucesivamente.

En este sentido la normalización de este tipo de identificadores se ha realizado verificando la bondad del carácter de control para lo cual se han utilizado una serie de reglas de verificación de dicho carácter.

Normalización del NIF

En cumplimiento de la disposición adicional sexta de la Ley 58/2003, de 17 de diciembre, General Tributaria, el Real Decreto 1065/2007, de 27 de julio, establece que las personas físicas y jurídicas, así como los obligados tributarios a que se refiere el artículo 35.4 de esta Ley, tendrán obligatoriamente un número de identificación fiscal para sus relaciones de naturaleza o con trascendencia tributaria.

En relación al concepto del NIF es necesario decir que cuando se habla de él no se hace distinción entre personas físicas y personas jurídicas, sino que se habla del NIF en general. El motivo se debe a la entrada en vigor el 1 de enero de 2008 del Real Decreto 1065/2007, de 27 de julio, ya que hasta entonces el número de identificación fiscal asignado a personas jurídicas y entidades sin personalidad se correspondía con el Código de Identificación Fiscal (CIF). Así pues, cuando se hable de normalizar el campo identificador fiscal se hablará en general de NIF independientemente de que se refiera a personas físicas o jurídicas.

El proceso de normalización del NIF es análogo al realizado para el caso del DNI o NIE. En esencia, consiste en verificar que el carácter de control que acompaña a los caracteres que conforman este identificador es correcto. No obstante, habrá que hacer algunas distinciones en este proceso en función de una serie de casos que se muestran a continuación.

1. **Normalización del NIF de persona física**

   a) Si la persona es de nacionalidad española y tiene DNI o si por el contrario es extranjera y tiene NIE, el NIF coincide con éstos respectivamente, con lo cual se normalizan como DNI y NIE.

   b) Si la persona tiene nacionalidad española y realiza o participa en operaciones de naturaleza o con transcendencia tributaria pero no está obligada a obtener el DNI, bien por residir en el extranjero o bien por ser menor de 14 años, deberá obtener un número de identificación fiscal propio. Para ello puede optar por solicitar voluntariamente el DNI, en cuyo caso el NIF coincidirá con éste y la normalización coincidirá con la del DNI, o solicitar a la Administración tributaria la asignación de un número de identificación fiscal propio. Este último estará integrado por nueve caracteres con la siguiente composición:

      - Una letra inicial destinada a indicar la naturaleza de este número (**L** para españoles residentes en el extranjero y **K** para españoles que, residiendo en España, sean menores de 14 años).
      - Siete caracteres alfanuméricos.
      - Un carácter de verificación alfabético.

Para llevar a cabo la normalización de este NIF propio se utilizarán las mismas operaciones que para el DNI sin considerar la letra inicial.

c) Si la persona carece de la nacionalidad española y no dispone del número de identidad de extranjero, deberá solicitar a la Administración tributaria la asignación de un número de identificación fiscal cuando vaya a realizar operaciones de naturaleza o con trascendencia tributaria. Dicho número estará integrado por nueve caracteres con la siguiente composición:

   - Una letra inicial, que será la **M**, destinada a indicar la naturaleza de este número.
   - Siete caracteres alfanuméricos.
   - Un carácter de verificación alfabético.

   En este caso el proceso de normalización que se realiza es análogo al realizado en el apartado b) anterior.

2. Normalización del NIF de personas jurídicas y entidades sin personalidad jurídica

A partir del 1 de julio de 2008, el NIF para personas jurídicas y entidades sin personalidad jurídica (antes conocido como CIF) es un número que las identifica y que es invariable cualesquiera que sean las modificaciones que experimenten aquellas, salvo que cambie su forma jurídica o nacionalidad. Así pues, los cambios realizados en la normativa existente hasta ese momento, han hecho que el formato del NIF haya variado, siendo su composición actual la que se muestra a continuación.

El número de identificación fiscal de las personas jurídicas y entidades sin personalidad jurídica está compuesto por nueves caracteres distribuidos de la siguiente forma:

- Una letra, que muestra información sobre la forma jurídica, si se trata de una entidad española, o en su caso, el carácter de entidad extranjera o de establecimiento permanente de una entidad no residente en España.
- Un número aleatorio de siete dígitos.
- Un carácter de control, que puede ser un número o una letra.

En la siguiente tabla se muestran las claves sobre la forma jurídica de entidades españolas, la clave de entidad extranjera y la de establecimiento permanente de una entidad no residente en España.

Tabla de claves		
Clave	**Forma jurídica de entidades españolas**	**Tipo de clave**
A	Sociedades anónimas	Número
B	Sociedades de responsabilidad limitada	Número
C	Sociedades colectivas	Número
D	Sociedades comanditarias	Número
E	Comunidades de bienes y herencias yacentes	Número
F	Sociedades cooperativas	Número
G	Asociaciones	Número
H	Comunidades de propietarios en régimen de propiedad horizontal	Número
J	Sociedades civiles, con o sin personalidad jurídica	Número
P	Corporaciones locales	Letra
Q	Organismos públicos	Letra
R	Congregaciones e instituciones religiosas	Letra
S	Órganos de la Administración del Estado y de las Comunidades Autónomas	Letra
U	Uniones temporales de empresas	Número
V	Otros tipos no definidos en el resto de claves	Número
Clave	**Entidad extranjera**	Letra
N	Se usa para personas jurídicas y entidades sin personalidad jurídica que carezcan de la nacionalidad española	Letra
Clave	**Establecimiento permanente de una entidad no residente en España**	
W	Se usa para personas jurídicas y entidades sin personalidad jurídica no residentes en territorio español pero que operan en éste por medio de uno o varios establecimientos permanentes que realicen actividades claramente diferenciadas y cuya gestión se lleve de modo separado, de acuerdo a la Ley del Impuesto sobre la Renta de no Residentes aprobado por el Real Decreto Legislativo 5/2004, de 5 de marzo.	Letra

Como en los casos anteriores la normalización de este campo consistirá en verificar que el carácter de control que acompaña a los 8 dígitos anteriores a éste es el correcto. Para ello se efectuarán una serie de operaciones sobre los siete dígitos centrales, de forma que se verifique que el carácter de control es adecuado.

Tras la breve explicación teórica sobre cómo se lleva a cabo el proceso de normalización del campo identificador de persona física o jurídica, así como los motivos por los que se realiza, vamos a explicar cómo se efectúa el proceso de normalización a través de la aplicación informática *ADYN Herramienta de Normalización*.

Para ello vamos a utilizar el fichero '*Ejemplo_nif.csv*' ubicado en 'C:\adyn\ejemplos\idpersonas'. El fichero consta de dos campos: código de identificación (ID) y número de identificación fiscal (NIF). Los datos recogidos en estos campos son ficticios por lo que no existen personas físicas o jurídicas que estén identificados por estos valores.

El fichero tiene un tamaño de 5 registros y su contenido se muestra en la siguiente imagen:

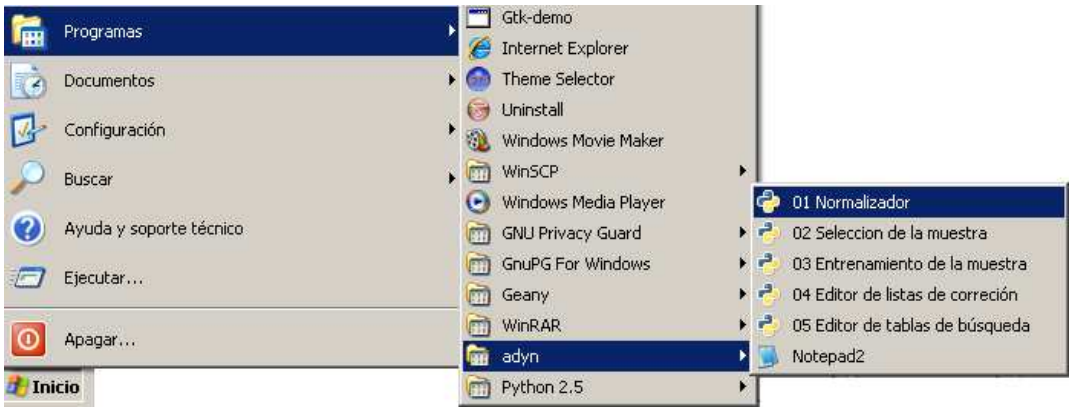
ejemplo_nif.csv

	A	B
1	id	nif
2	110103	J7609217w
3	120103	81113431s
4	130103	65091232w
5	140103	a88092546
6	150103	w76092176

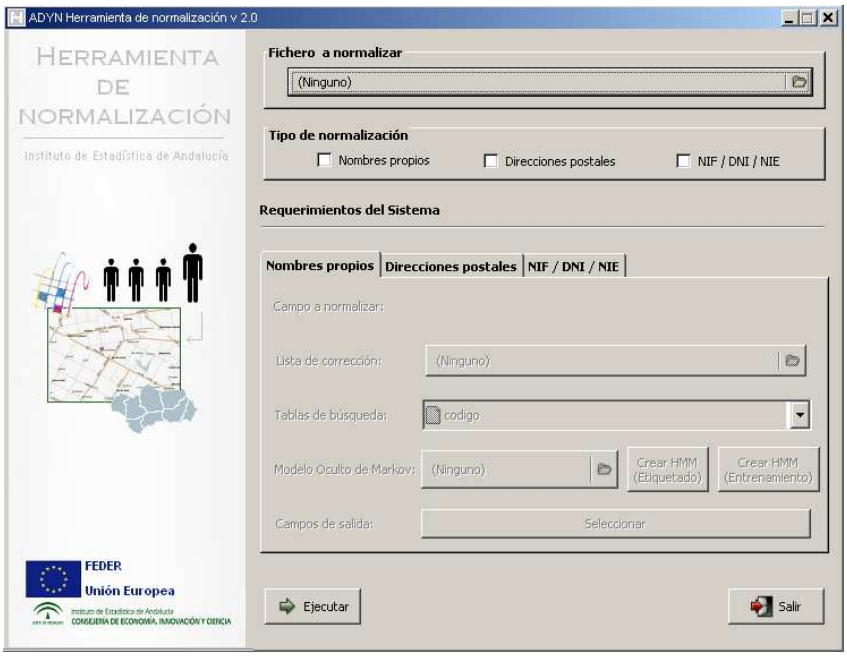
Para normalizar este fichero de datos no es necesario seleccionar ni entrenar ninguna muestra que cree el Modelo Oculto de Markov que reconozca los patrones que siguen los datos del campo '*nif*' a normalizar. El motivo es que en este caso ya se dispone de un modelo que se suministra junto a la aplicación llamado '*modelo HMM_idpersona.hmm*' y que se encuentra ubicado en el siguiente directorio:

'adyn\ejemplos\idpersonas\modelo_propuesto'

Por tanto, para efectuar el proceso nos iremos directamente a la interfaz de normalización, es decir, elegiremos la opción ***'01. Normalizador'*** tal y como se muestra en la siguiente imagen:



Y aparecerá la ventana principal de *ADYN Herramienta de Normalización*:



En ella tendremos que indicar lo siguiente:

1. Elegir el fichero que vamos a normalizar, en nuestro caso, '*ejemplo_nif*'.

2. Marcar la opción 'NIF/DNI/NIE' dentro del apartado 'Tipo de normalización', ya que vamos a normalizar un identificador de persona jurídica.

3. Seleccionar la pestaña 'NIF/DNI/NIE' dentro del apartado 'Requerimientos del Sistema', ya que vamos a normalizar el NIF. Y en ella especificar lo siguiente:

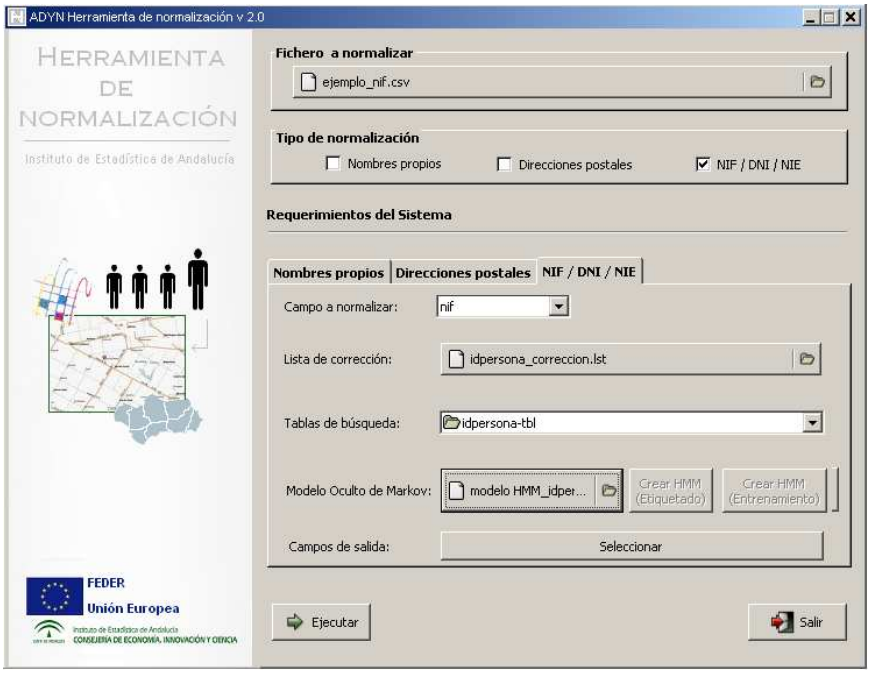
- 'Campo a normalizar': seleccionamos el campo que contiene el identificador de la persona física o jurídica a normalizar, en nuestro caso el campo '*nif*'.

- 'Lista de corrección': seleccionamos la correspondiente a identificadores de personas físicas o jurídicas, esto es, seleccionar el archivo llamado '*idpersona_correccion.lst*' que se encuentra en la siguiente ubicación '*adyn\codigo\datos\ListasDeCorreccion*'.

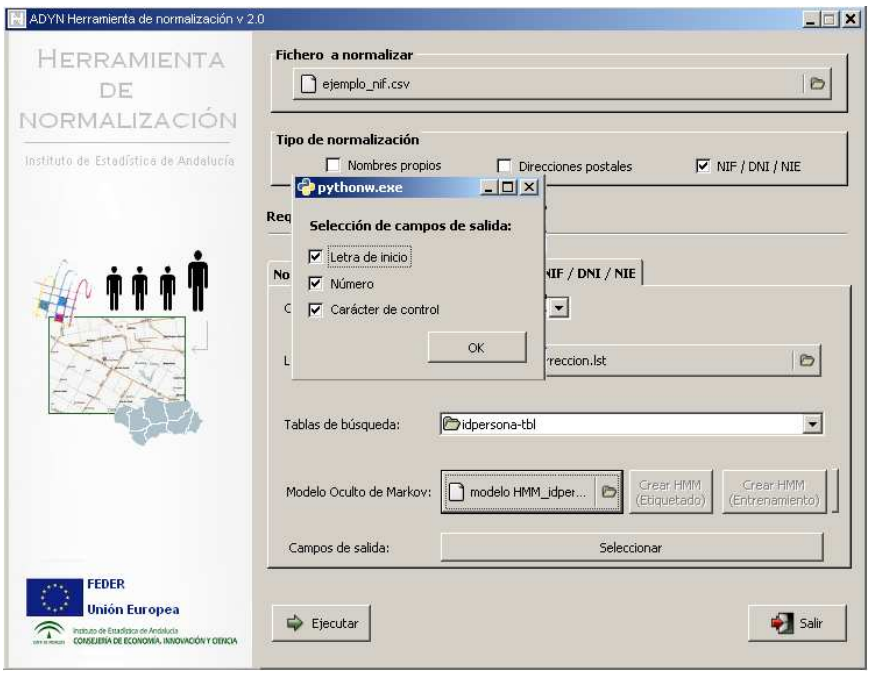
- 'Tablas de búsqueda': seleccionamos la asociada a estos identificadores de persona que se encuentra en '*adyn\codigo\datos\idpersona-tbl*'.

- 'Modelo Oculto de Markov': elegimos el modelo HMM propuesto para segmentar los identificadores de personas físicas o jurídicas llamado 'modelo HMM_idpersona.hmm' y que se encuentra ubicado en el siguiente directorio 'adyn\ejemplos\idpersonas\modelo_propuesto'.

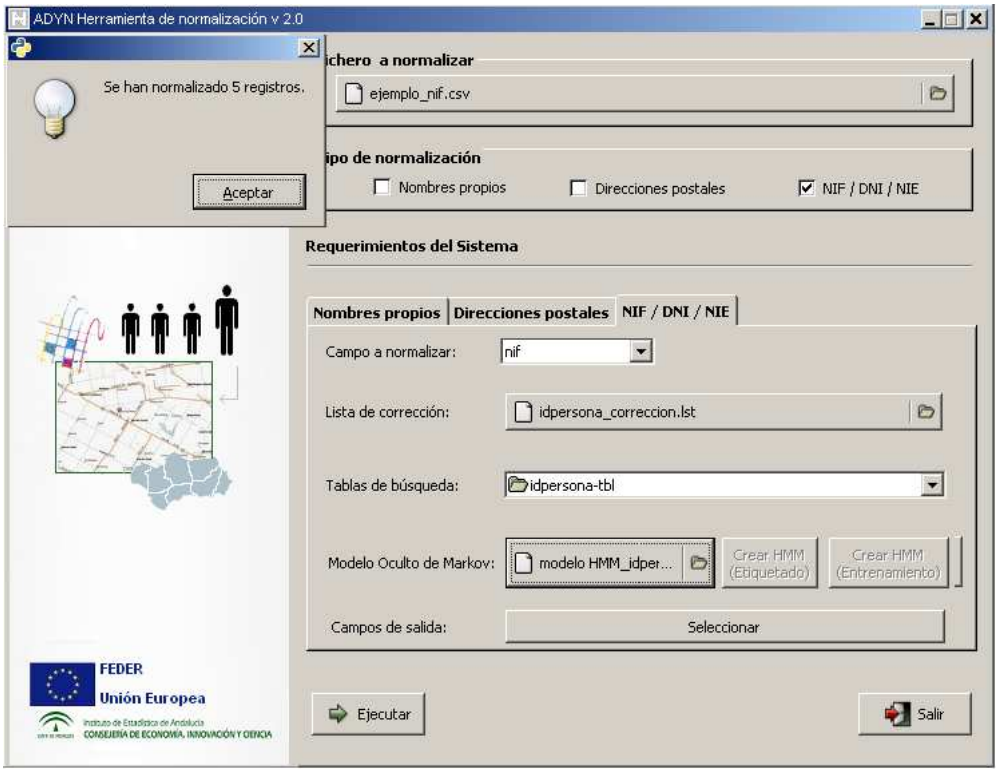
El resultado es la siguiente pantalla:



Por último seleccionamos los campos en los que se ha decidido segmentar el identificador de persona física o jurídica, estos son: letra de inicio, número de identificación y carácter de control. Marcamos, por ejemplo, los tres y pulsamos 'OK'.



Una vez hemos seleccionado los campos de salida pulsamos el botón '**Ejecutar**' para realizar el proceso de normalización. Si éste se ha realizado satisfactoriamente aparecerá una ventana de información indicándonos el número de registros que se han normalizado.



Por último observamos el fichero normalizado resultante del proceso; este fichero se ha guardado automáticamente en la misma ubicación donde se encontraba el fichero a normalizar, '*ejemplo_nif.csv*', es nuestro caso, en '*adyn\ejemplos\idpersonas*'. La denominación con la que se guarda automáticamente el fichero es: '*est_fechanormalizacion-horanormalizacion_ejemplo_nif.csv*', aunque el usuario podría decidir guardarlo con otro nombre.

est_20100223-1354_ejemplo_nif.csv - OpenOffice.org Calc

	A	B	C	D	E	F
1	id	nif	letra_inicio	numero_ID	caracter_control	validacion
2	110103	j7609217w	j	7609217	w	1
3	120103	81113431s		81113431	s	1
4	130103	65091232w		65091232	w	1
5	140103	a88092546	a	8809254	6	1
6	150103	w76092176	w	7609217	6	1

Si se observa el fichero, cada uno de los valores del campo '*nif*' ha sido segmentado correctamente según el criterio que hemos definido, esto es: letra de inicio, número de inicio y carácter de control.

Junto a estos campos aparece el campo 'validacion' que tomará los valores 1 ó 0 dependiendo de que el resultado de aplicar los algoritmos de validación a esos valores del campo '*nif*' haya dado resultados erróneos o satisfactorios respectivamente. Los algoritmos de validación variarán dependiendo de si el valor que tenemos en el campo a normalizar es un documento nacional de identidad o un número de identificación de extranjero, o de si es un número de identificación fiscal de una persona física o jurídica.

En nuestro caso aparecen todos los valores del campo validación iguales a '1'. Esto quiere decir que todos ellos están mal normalizados, lo cual no implica que nuestro proceso no funcione bien, sino que lo que ha ocurrido es que al utilizar valores ficticios con el fin de anonimizar el proceso de normalización ha dado la casualidad de que todos los caracteres de control son erróneos.

Anexo III: Ficheros de datos CSV.

a) Utilización de ficheros de datos CSV de gran tamaño.

La aplicación informática *ADYN* permite la normalización de cualquier fichero con formato CSV independientemente del tamaño que tenga, si bien es recomendable, con el fin de detectar posibles errores en la ejecución de la aplicación, dividir ese fichero de gran tamaño en otros de menor tamaño.

b) El formato de ficheros CSV.

El formato de fichero **CSV** (Valores separados por comas, del inglés *Comma Separated Value*), es de uso frecuente para el intercambio de datos entre diferentes aplicaciones. Esto le ha llevado a convertirse en un estándar de facto, si bien no es un estándar real como lo son HTML u OpenDocument. Esto provoca que podamos encontrar varias interpretaciones de este formato y necesitemos que nuestra aplicación los pueda leer todos.

A día de hoy, es muy frecuente que las aplicaciones usen otros formatos para el intercambio de información tales como XML (lenguaje de marcas ampliable, del inglés *Extensible Markup Language*), sin embargo, CSV es aún muy usado en algunos contextos por razones de compatibilidad o porque es más sencillo y fácil de leer que XML.

Características del formato CSV:

1- Cada línea contiene un registro.

2- Los campos pueden estar separados por comas (usualmente) o bien, puntos y comas.

3- Los espacios que aparecen al inicio o final de un campo son ignorados, así que si deseamos que un campo lleve espacios al final o inicio, tendremos que usar dobles comillas.

4- Los campos que contienen comas pueden ser delimitados usando dobles comas.

5- Los campos que contengan dobles comillas deben ser delimitados usando dobles comillas, y las comillas dobles deben ser encerradas entre otras comillas dobles.

6- El primer registro en un fichero puede ser opcionalmente un registro que contiene los nombres de las variables.

Se puede encontrar más información sobre este formato en el RFC 4180 (http://tools.ietf.org/html/rfc4180).

Ejemplo de un fichero CSV:

INDICE,NOMBRE,DIRECCION

1,MARIA JOSE GOMER GARCIA,C/ PENSAMIENTO 41

2,ANTONIO GOMEZ DE LA VILLA GUERRERO,CALLE MANUEL DE MOLINA 78

3,JOSE ANDRES LOPEZ JIMENEZ,AVDA DE LA CONSTITUCIÓN 34 PUERTA 1

4,GARCIA GARCIA JOSE,AVDA. DE ANDALUCIA 45

5,MANOLO JIMEPEZ DEL BOMBO,CARRETERA ALCARACEJOS S/N

6,ANA PEREZ GARCIA,CL JULIAN NAVARRO FLORES 3

7,JOSE FRANCISCO DEL OLMO PIRIR,C/ POZO 46

8,PEPI GOMEZ CRANE,C/ POZO NUEVO 46

9,DENISE ANA GOMEZ GARCIA,CL JOSÉ LAGUILLO 28

10,JULIA PEREZ HINIESTA,AV SAN ANTON 118

Esto podría representarse fácilmente en una tabla como la siguiente:

INDICE	NOMBRE	DIRECCION
1	MARIA JOSE GOMER GARCIA	C/ PENSAMIENTO 41
2	ANTONIO GOMEZ DE LA VILLA GUERRERO	CALLE MANUEL DE MOLINA 78
3	JOSE ANDRES LOPEZ JIMENEZ	AVDA DE LA CONSTITUCIÓN 34 PUERTA 1
4	GARCIA GARCIA JOSE	AVDA. DE ANDALUCIA 45
5	MANOLO JIMEPEZ DEL BOMBO	CARRETERA ALCARACEJOS S/N
6	ANA PEREZ GARCIA	CL JULIAN NAVARRO FLORES 3
7	JOSE FRANCISCO DEL OLMO PIRIR	C/ POZO 46
8	PEPI GOMEZ CRANE	C/ POZO NUEVO 46
9	DENISE ANA GOMEZ GARCIA	CL JOSÉ LAGUILLO 28
10	JULIA PEREZ HINIESTA	AV SAN ANTON 118

Puesto que nosotros trabajaremos con ficheros CSV de todo tipo, se ha impuesto que *ADYN Herramienta de Normalización* detecte automáticamente con qué tipo de fichero CSV estamos trabajando analizando que tipo de separador se usa en el fichero de entrada y si usa comillas dobles o simples. La aplicación convertirá esto a un formato donde se separan los campos con comas y se rodean de comillas dobles, de forma totalmente transparente para el usuario y

trabajará con este archivo.

Anexo IV: Etiquetas y estados.

Para **direcciones postales** tenemos el siguiente conjunto de etiquetas:

Etiqueta	Identificación	Tabla de búsqueda
AP	Apartado de correos	Kcod_postal.tbl
BD	Barriada	kbarriada.tbl
BL	Bloque	kbloque.tbl
CJ	Complejo	kcomplejo.tbl
CM	Comercio	kcomercio.tbl
ED	Edificio	kedificio.tbl
EG	Entidad singular	Kentidad_singular.tbl
ER	Edificio singular	kedificio_singular.tbl
ES	Escalera	kescalera.tbl
KM	Kilómetro	kkilomento.tbl
LE	Una sola letra	kletra.tbl
LN	Localidad	klocalidad.tbl
MZ	Manzana	kmanzana.tbl
N5	Numero de apartado de correos	
NM	Numero local	knumero_local.tbl
NP	Número de planta	kplanta_numero.tbl
NU	Valor numérico	
NV	Nave industrial	knave.tbl
PA	Parcela	kparcela.tbl
PL	Planta	kplanta.tbl
PR	Provincia	kprovicia.tbl
PT	Portal	kportal.tbl
PU	Puerta	kpuerta.tbl
ST	Sector	ksector.tbl
TV	Tipo de vía	kvia.tbl
UN	Desconocido	
ZO	Zona	kzona.tbl

Y el siguiente conjunto de estados:

Estado	Descripción del Estado
tipo_de_via	Identificador del tipo de vía (calle, avenida, etc.)
nombre_de_via	Nombre de la vía
identificador_de_numero	Identifica caracteres relacionados con el número de la vía, por ejemplo: s/n (sin número), nº…
numero	Número del local
identificador_de_bloque	Identifica caracteres relacionados con el bloque o edificio (bloq., edif,...)
bloque	Nombre del bloque o edificio
identificador_de_portal	Identifica caracteres relacionados con el portal (pol, portal,…)
portal	Nombre o número del portal
identificador_de_escalera	Identifica caracteres relacionados con la escalera (esc,…)
escalera	Nombre o número de la escalera
identificador_de_planta	Identifica caracteres relacionados con la planta (plt,…)
planta	Nombre o número de la planta
identificador_de_puerta	Identifica caracteres relacionados con la puerta (puerta,…)
puerta	Nombre o número de la puerta
identificador_de_letra	Identifica caracteres relacionados con la letra de la puerta (ltr,…).
letra	Carácter asociado a la letra
identificador_de_barriada	Identifica caracteres relacionados con la barriada (bda,…).
barriada	Nombre de la barriada
identificador_de_sector	Identifica caracteres relacionados con un sector dentro de un polígono, parque empresarial, etc.
sector	Nombre o acrónimo del sector
identificador_edificio_singular	Identifica caracteres relacionados con edificios singulares. Por singulares entendemos centros médicos, colegios, mercados, etc.
edificio_singular	Nombre del edificio singular
identificador_de_codigo_postal	Identifica caracteres relacionados con el código postal
codigo_postal	Número del código postal
localidad	Nombre de localidad
provincia	Nombre de provincia
entidad_singular	Nombre de la entidad singular (pedanías, aldeas, etc.)
identificador_de_zona	Identifica caracteres relacionados con zonas, entendiendo por zona urbanizaciones, polígonos industriales, parques empresariales, etc.

Estado	Descripción del Estado
zona	Nombre de la zona
identificador_de_complejo	Identifica caracteres relacionados con un complejo que forme parte de un polígono industrial, parque tecnológico, etc.
complejo	Nombre del complejo
identificador_de_manzana	Identifica caracteres relacionados con una manzana (mzn, etc.)
manzana	Nombre de la manzana
identificador_de_parcela	Identifica caracteres relacionados con una parcela (pzla, etc.)
parcela	Nombre de la parcela
identificador_kilometro	Identifica caracteres relacionados con un punto kilométrico
kilometro	Número del kilómetro
identificador_de_nave	Identifica caracteres relacionados con una nave industrial
nave	Nombre de la nave
tipo_de_comercio	Tipo de comercio (bar, supermercado, mercería, etc.)
nombre_de_comercio	Nombre del comercio
informacion_adicional	Información que no se sabe cómo clasificar
informacion_adicional_parentesis	Información contenida entre paréntesis

Para **nombres de personas** tenemos el siguiente conjunto de etiquetas:

Etiqueta	Identificación	Tabla de búsqueda
NF	Nombre femenino	knombres_femeninos.tbl
NM	Nombre masculino	knombres_masculinos.tbl
NN	Nombre neutro	knombres_neutros.tbl
PS	Partículas ligadas a nombres y/o apellidos	kparticulas.tbl
LE	Una sola letra	
UN	Desconocido	

Y el siguiente conjunto de estados:

Estado	Descripción
nombre1	Primer nombre
particula_nombre1	Partículas que siguen al primer nombre
nombre2	Segundo nombre
particula_nombre2	Partículas que siguen al segundo nombre
nombre3	Resto de nombres
preparticula_apellido1	Partículas que preceden al primer apellido
apellido1	Primer apellido
particula_apellido1	Partículas que forman parte del primer apellido si es compuesto
preparticula_apellido2	Partículas que preceden al segundo apellido
apellido2	Segundo apellido
particula_apellido2	Partículas que forman parte del segundo apellido si es compuesto

Para **identificadores de personas físicas y jurídicas** tenemos una etiqueta que está asociada a la única tabla de búsqueda que se utilizará en el proceso de normalización de este tipo de datos.

Etiqueta	Identificación	Tabla de búsqueda
TC	Tipo de clave	kidpersona.tbl

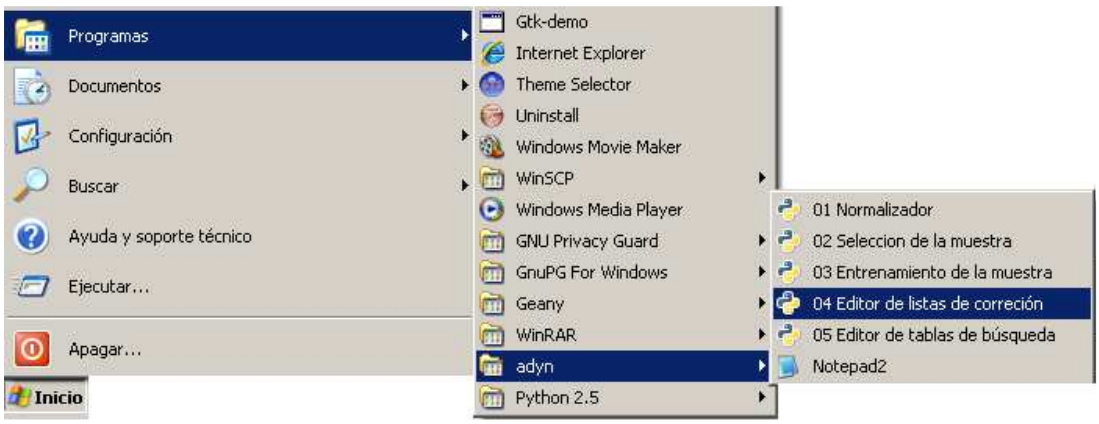
Los estados asociados a estos identificadores de personas físicas y jurídicas son los siguientes:

Estado	Descripción
letra_inicio	Letra inicial al número de identificación
numero_ID	**Número de identificación**
caracter_control	Carácter de control del identificador

Anexo V: Listas de corrección y tablas de búsqueda.

Las listas de corrección y las tablas de búsqueda son ficheros que por su función deben actualizarse de forma continua ya que a medida que se van realizando procesos de normalización irán apareciendo nuevos elementos que no se hayan recogido en ellas. Así pues, tendremos listas de corrección y tablas de búsqueda para nombres de personas, direcciones postales e identificadores de personas físicas y jurídicas que podrán ser personalizadas y modificadas por el usuario.

Para visualizar y modificar estos elementos se han desarrollado unos editores a los que se accede a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) tal y como se muestra en la siguiente imagen:

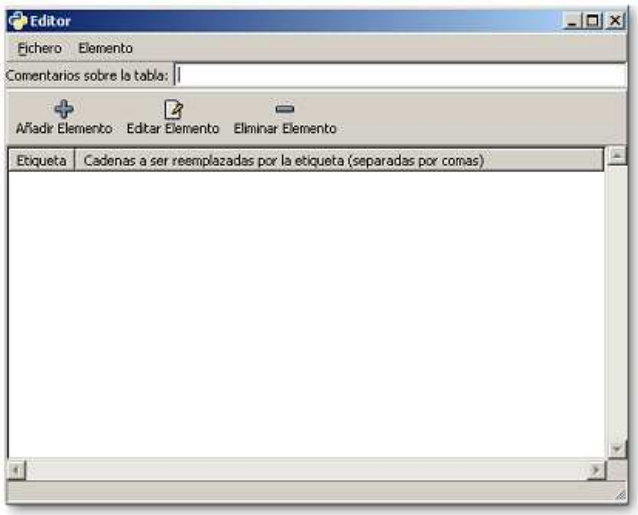


Detenidamente tenemos:

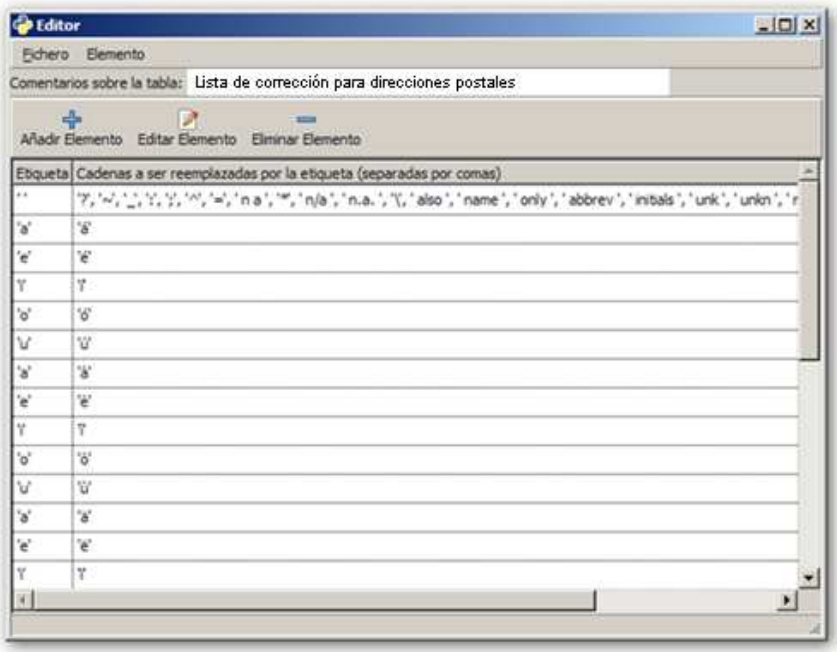
a) Las listas de corrección.

Como hemos comentado anteriormente permiten limpiar y corregir el fichero de datos. Son ficheros de extensión '*.lst*' que contienen los caracteres que el usuario desea eliminar o sustituir en los ficheros, por ejemplo, eliminar los caracteres extraños ('%', '?', …) y sustituir las vocales con tildes por las vocales sin tildes.

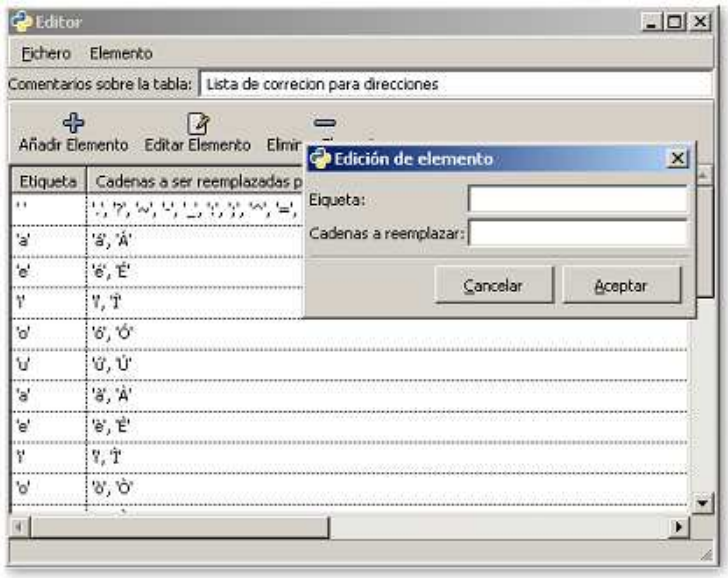
Existen listas de corrección para direcciones postales, nombres de personas e identificadores de personas físicas y jurídicas. Para visualizarlas hacemos click sobre ***'04. Editor de listas de corrección'*** y aparecerá una pantalla del tipo:



A continuación pulsamos en el menú **'Fichero'** y elegimos la opción 'Abrir'. Iremos navegando por los directorios de la aplicación hasta encontrar la carpeta '*ListasDeCorreccion*' que se encuentra en la ruta 'C:\adyn\codigo\datos\ListasDeCorreccion'. En ella podremos seleccionar la lista de corrección de nombres de personas, de direcciones postales o la correspondiente a identificadores de personas físicas y jurídicas. Si elegimos abrir la referida a direcciones postales el resultado será el siguiente:



Para añadir un nuevo elemento a sustituir tendremos que pulsar el botón 'Añadir Elemento', apareciendo la siguiente pantalla:



Así en el campo correspondiente a 'Etiqueta' pondremos el elemento que aparecerá cuando sustituyamos alguno de los elementos que incluyamos en 'Cadenas a reemplazar.'

De igual forma, si queremos editar alguno de los elementos que ya se encuentren disponibles, nos situamos sobre él y pulsamos 'Editar Elemento'.

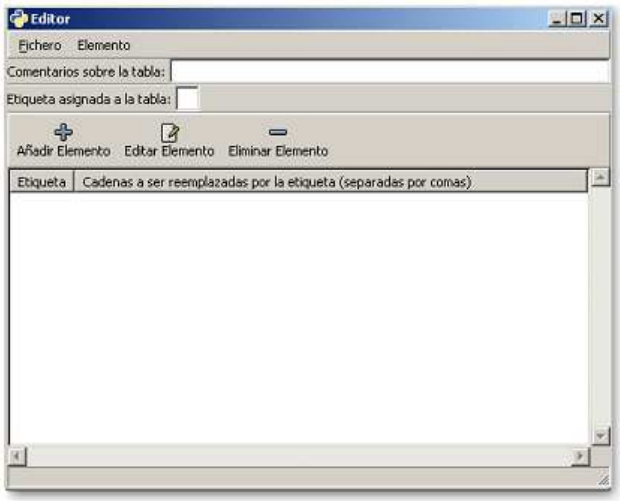
Finalmente, si queremos eliminar algún otro elemento nos situamos sobre él y pulsamos 'Eliminar Elemento'.

Tras realizar estas operaciones **es imprescindible que guardemos los cambios** ('Fichero/Guardar') antes de cerrar el archivo.

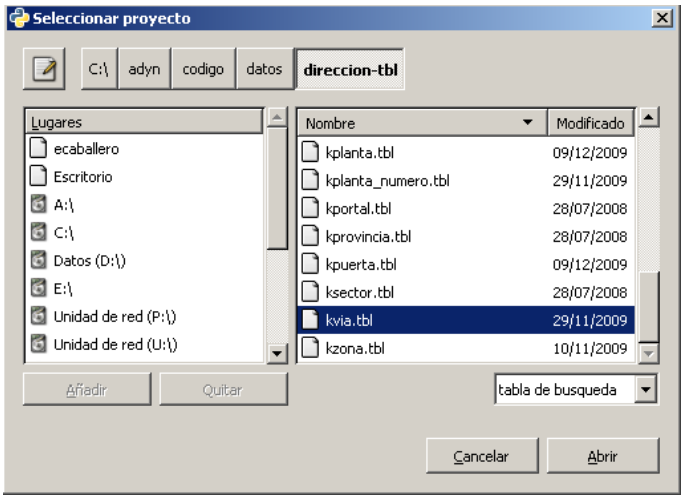
b) Las tablas de búsqueda.

Son otro tipo de ficheros con extensión '*.tbl*', que además de sustituir un elemento del fichero de datos por su valor estandarizado, le asigna una etiqueta. Al igual que las listas de corrección, tendremos tablas de búsqueda para nombres de personas, para direcciones postales y para identificadores de personas físicas y jurídicas.

Por ejemplo en el caso de que estemos normalizando un campo dirección postal, si la aplicación encuentra el elemento 'polg' lo sustituye por 'polígono industrial' y le asigna la etiqueta 'ZO' que significa Zona. Para visualizar las tablas de búsqueda hacemos click sobre ***'05. Editor de tablas de búsqueda'*** y aparecerá una pantalla del tipo:



A continuación pulsamos en el menú **'Fichero'** y elegimos la opción 'Abrir'. Iremos navegando por los directorios de la aplicación hasta encontrar las carpetas '*direccion-tbl*', '*nombre-tbl*' o '*idpersona-tbl*' que contienen las tablas de búsqueda para direcciones postales, para nombres de personas y para identificadores de personas físicas y jurídicas respectivamente. Estas carpetas se encuentran en la ruta 'C:\adyn\codigo\datos'. Si elegimos abrir las relativas a direcciones postales el resultado se muestra en la siguiente imagen:



Hemos de indicar que todos los archivos con extensión '.tbl' que se muestran en la imagen anterior son las diferentes tablas de búsqueda que utilizaremos en el proceso de etiquetado de los campos. Las etiquetas y tablas de búsqueda asociadas a direcciones postales se muestran en la siguiente tabla:

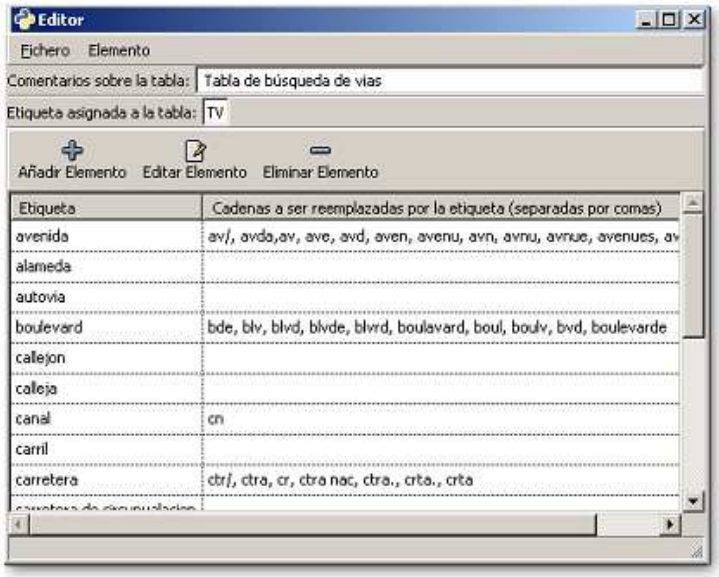
Etiqueta	Identificación	Tabla de búsqueda
AP	Apartado de correos	Kcod_postal.tbl
BD	Barriada	kbarriada.tbl
BL	Bloque	kbloque.tbl
CJ	Complejo	kcomplejo.tbl
CM	Comercio	kcomercio.tbl
ED	Edificio	kedificio.tbl
EG	Entidad singular	Kentidad_singular.tbl
ER	Edificio singular	kedificio_singular.tbl
ES	Escalera	kescalera.tbl
KM	Kilómetro	kkilomento.tbl
LE	Una sola letra	kletra.tbl
LN	Localidad	klocalidad.tbl
MZ	Manzana	kmanzana.tbl
N5	Numero de apartado de correos	
NM	Numero local	knumero_local.tbl
NP	Número de planta	kplanta_numero.tbl
NU	Valor numérico	
NV	Nave industrial	knave.tbl
PA	Parcela	kparcela.tbl
PL	Planta	kplanta.tbl
PR	Provincia	kprovicia.tbl
PT	Portal	kportal.tbl
PU	Puerta	kpuerta.tbl
ST	Sector	ksector.tbl
TV	Tipo de vía	kvia.tbl
UN	Desconocido	
ZO	Zona	kzona.tbl

Notar que para algunos casos no es necesario usar tablas de búsqueda:

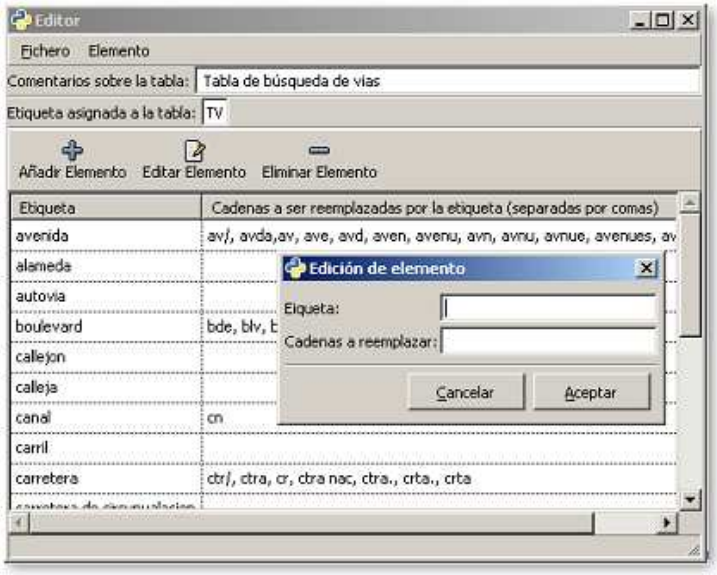
- Si en el campo dirección postal tenemos información entre paréntesis, esa información se incluirá directamente en el campo ‘información_adicional_parentesis’. En una segunda fase puede resultar útil normalizar este campo.

- Respecto a los nombres de las vías no utilizamos las tablas de búsqueda debido a la gran complejidad y heterogeneidad que presentan y etiquetaremos la información como 'UN' (*unknown* en inglés).
- En el caso de que aparezca un número se etiquetará automáticamente como 'NU'.
- En el caso de que aparezcan cinco números seguidos se etiquetará con 'N5' y será útil para su posterior asignación de estado como 'codigo_postal'.

Para visualizar, por ejemplo, los datos recogidos en la tabla de búsqueda de tipos de vía (kvia.tbl) haremos doble click sobre ella y aparecerá la siguiente pantalla:



Para añadir algún nuevo elemento habrá que pulsar el botón 'Añadir Elemento' apareciendo la siguiente pantalla:



Así, si en el campo 'Etiqueta' especificamos la cadena 'parque' y en el campo 'Cadenas a reemplazar' incluimos 'parq, paque' cuando en el fichero de datos aparezca alguna de las cadenas a reemplazar se sustituirá por la cadena 'parque'.

Para editar o eliminar un elemento pulsaremos sus botones correspondientes y actuaremos de forma similar.

En cuanto a los nombres de personas y a los identificadores de personas físicas y jurídicas, *ADYN Herramienta de Normalización* pone a disposición del usuario los siguientes conjuntos de etiquetas y tablas de búsqueda que se muestran a continuación:

- Nombres de personas

Etiqueta	Identificación	Tabla de búsqueda
NF	Nombre femenino	knombres_femeninos.tbl
NM	Nombre masculino	knombres_masculinos.tbl
NN	Nombre neutro	knombres_neutros.tbl
PS	Partículas ligadas a nombres y/o apellidos	kparticulas.tbl
LE	Una sola letra	
UN	Desconocido	

Nótese que si en un nombre de persona aparece una sola letra de forma independiente se etiquetará con 'LE' a pesar de no tener asociada ninguna tabla de búsqueda. El resto de elementos que no sean letras y que no se encuentren en las tablas de búsqueda se etiquetarán como 'UN' (desconocido en inglés). Un caso de este tipo son la mayoría de los apellidos que al no tener tablas de búsqueda asociadas, por la heterogeneidad de los datos, se etiquetarán como 'UN'.

Por otro lado, antes de finalizar este apartado es necesario comentar que en la tabla de búsqueda relativa a nombres de personas femeninos se han incluido unas etiquetas especiales para corregir los siguientes casos que nos hemos encontrado en algunos ficheros de datos: 'mariangeles', 'marisabel', etc. En estos casos lo que se he hecho es incluir las etiquetas ''Maria Angeles' o 'Maria Isabel' que van a sustituir a las cadenas 'mariangeles' o 'marisabel' cuando aparezcan en el campo nombre del fichero original de datos que queramos normalizar.

- Identificador de persona física o jurídica:

Etiqueta	Identificación	Tabla de búsqueda
TC	Tipo de clave	kidpersona.tbl

Anexo VI: Modelos Ocultos de Markov.

a) Desarrollo teórico.

Los Modelos Ocultos de Markov (en inglés, *Hidden Markov Models* o HMM) reconocen ciertos patrones de comportamiento que siguen los datos contenidos en un fichero de datos, permitiéndonos estandarizar y segmentar dichos datos.

El objetivo de la segmentación es separar las entidades presentes en un campo de un fichero para facilitar posteriormente otros tipos de análisis con dicho fichero, como pueden ser las comparaciones. Por ejemplo, un campo que contiene el nombre y apellidos puede ser separado en tres nuevos campos, nombre, primer apellido y segundo apellido. No siempre es evidente cómo aislar la descripción clara de una dirección o un nombre. Para extraer los distintos descriptores se pueden emplear los Modelos Ocultos de Markov.

Un HMM es un modelo estadístico donde el sistema subyacente puede presentar varios estados, y para cada estado se pueden observar distintas etiquetas. Los HMM fueron propuestos en la segunda década de los años 60, y se han utilizado tradicionalmente en sistemas de reconocimiento de la voz. Una descripción muy completa de estos modelos se puede consultar en el artículo de Rabiner (1989).

Un modelo HMM viene definido por los siguientes elementos fundamentales:

1. N, el número finito de posibles estados del sistema, son ocultos (no observables directamente). Denotaremos el conjunto de estados por $E=\{e_1,\ldots,e_N\}$ y el estado en el momento t por q_t.
2. La matriz de probabilidades de transición entre estados $A=\{a_{ij}\}$, donde $a_{ij}=P[q_{t+1}=E_j/\ q_t=E_i]$, $1\leq i,j\leq N$.
3. L, el número finito de posibles etiquetas por cada estado. Las etiquetas son las observaciones que pueden realizarse, sus probabilidades dependen del estado. El conjunto de etiquetas será denotado por $S=\{s_1,\ldots,s_L\}$.
4. La función de probabilidad de las etiquetas en el estado j, definida por la matriz $B=\{b_j(k)\}$, donde $b_j(k)=P[s_k \text{ en } t/\ q_t=e_j]$, $1\leq j\leq N$, $1\leq k\leq L$.
5. La distribución inicial de probabilidad de los estados $\pi=\{\pi_i\}$, siendo $\pi_i=P[q_1=e_i]$, $1\leq i\leq N$.

Por tanto la especificación completa de un modelo HMM requiere conocer N, L y las tres medidas de probabilidad $\lambda=(A,B,\pi)$. Los tres problemas fundamentales en la construcción de los modelos HMM son los siguientes:

1. Dada la secuencia observada de etiquetas $O=O_1O_2\ldots O_T$ y un modelo HMM con vector de parámetros λ, ¿cómo calcular de forma eficiente en términos computacionales $P(O/\ \lambda)$?
2. Dada la secuencia observada de etiquetas $O=O_1O_2\ldots O_T$ y un modelo HMM con vector de parámetros λ, ¿cómo elegir una secuencia de estados $Q= q_1q_2\ldots q_T$ que mejor explique la secuencia observada de etiquetas?
3. ¿Cómo estimar λ para que se maximice $P(O/\ \lambda)$?

El primer problema suele ser conocido con el nombre de problema de evaluación. En principio la dificultad fundamental surge del hecho de que no se conoce la secuencia de estados subyacente, por lo que habría que aplicar el Teorema de la Probabilidad Total:

$$P(O/\lambda) = \sum_{Q} P(O/Q,\lambda)P(Q/\lambda) = \sum_{q_1 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_T}(O_T)$$

El cálculo directo de esta probabilidad no es factible computacionalmente, por ejemplo para N=5 y T=100 se requieren aproximadamente 10^{72} cálculos. Afortunadamente la etapa adelante del algoritmo Adelante-Detrás (Forward-backward) permite calcular de forma eficiente estas probabilidades. Por ejemplo para N=5, T=100, el número de cálculos se reduce a 3000.

El problema 2, a diferencia del anterior, admite varias soluciones, dependiendo del criterio de optimalidad que se defina. Un primer criterio podría consistir en elegir aquellos estados que sean individualmente más probables. Definiendo $\gamma_t(i)=P[q_t=e^i/O,\lambda]$ que puede ser calculado fácilmente en función de términos que intervienen en el procedimiento adelante-detrás, se podría elegir q_t como aquel estado que maximiza $\gamma_t(i)$ entre los N estados posibles. Si bien este procedimiento maximiza el número esperado de estados correctamente identificados, las secuencias resultantes pueden incluso ser no válidas. Una alternativa sería elegir secuencias de estados que maximizan el número esperado de pares o tripletas de estados correctamente identificados. Sin embargo el criterio más usado consiste en encontrar la mejor secuencia (camino) de estados, es decir, maximizar P(Q/O,λ), lo que equivale a maximizar P(Q,O/λ). El algoritmo de Viterbi, basado en métodos de programación dinámica, resuelve de forma eficiente este problema.

El tercer problema es el más difícil, la idea es estimar por máxima verosimilitud los parámetros del modelo λ=(A,B,π). Sin embargo no existe solución analítica a este problema. A cambio pueden buscarse máximos locales mediante procedimientos iterativos como el de Baum-Welch, basado en el algoritmo EM, o bien técnicas basadas en el gradiente.

En trabajos como el de Churches et al. (2002), el cálculo de las estimaciones de máxima verosimilitud se obtienen a partir de las frecuencias observadas de los distintos estados, transiciones entre estados y etiquetas observadas para cada estado. Para ello se debe disponer de un conjunto amplio de registros, suficientemente representativo. A partir del modelo HMM y mediante las técnicas correspondientes a los problemas 1 y 2 se puede calcular la secuencia de estados que con más probabilidad ha originado la secuencia de etiquetas observada, obteniendo finalmente la segmentación del texto considerado.

b) Otros usos de los Modelos Ocultos de Markov.

En el caso de que deseemos normalizar otro fichero de datos cuyo campo de direcciones postales tenga una estructura similar al que hemos normalizado en el ejemplo, podemos utilizar el Modelo Oculto de Markov creado previamente. Por lo tanto para normalizar este fichero de datos utilizaremos únicamente la interfaz ***'01. Normalizador'*** donde incluiremos el Modelo Oculto de Markov anteriormente creado, '*modelo1.hmm*'.

Este Modelo también nos puede ser útil a la hora de seleccionar una muestra de entrenamiento a través de la interfaz ***'02. Selección de la muestra'***. De esta forma la muestra de entrenamiento obtenida contendrá para cada registro las etiquetas y estados que el

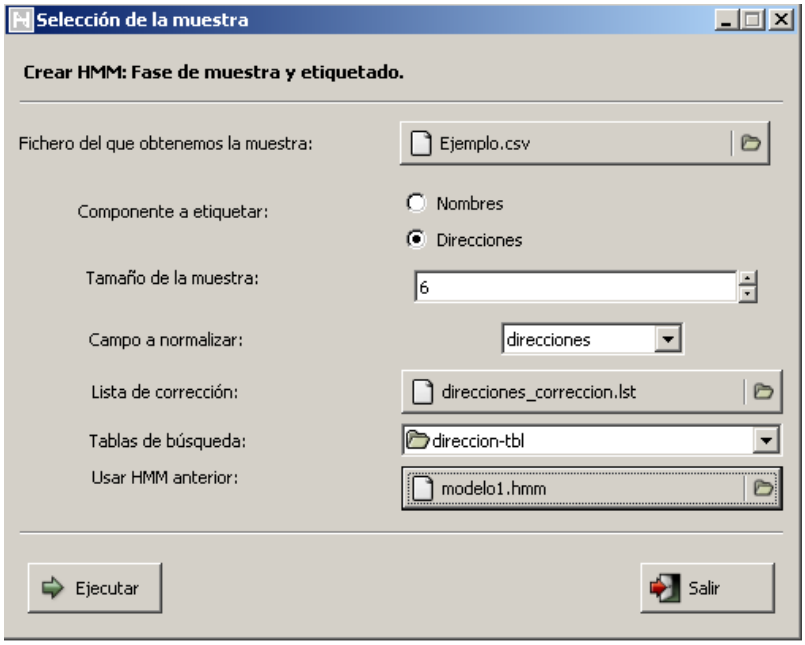
modelo HMM le haya asignado. No obstante, se aconseja realizar una revisión manual de dicho fichero con el fin de corregir posibles errores en ese proceso de asignación. Veamos cómo funciona este proceso:

Por ejemplo siguiendo con nuestro ejemplo de direcciones postales, si hubiéramos seleccionado una muestra de seis registros del campo a normalizar '*direcciones*' del fichero '*Ejemplo.csv*' sin usar un modelo HMM anterior, un posible fichero de salida sería:

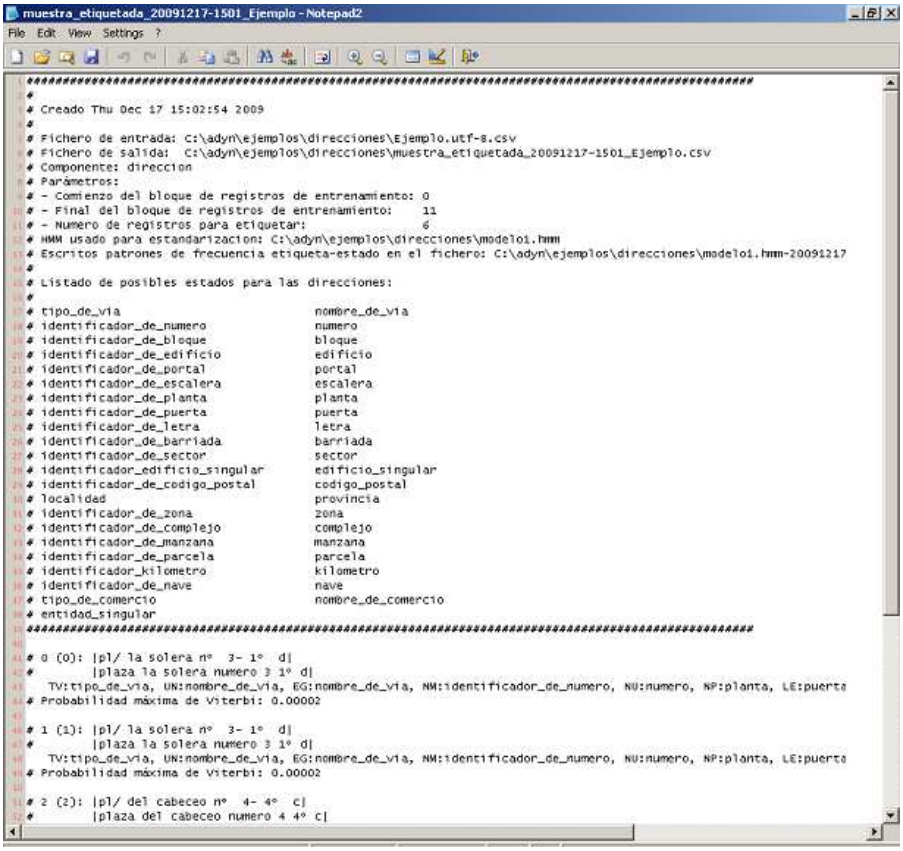
A continuación habría que asignar manualmente estados a las etiquetas que aparecen para posteriormente entrenar la muestra.

Sin embargo, si ya tenemos un modelo HMM previamente creado a partir de datos similares a los que estamos trabajando, por ejemplo, '*modelo1.hmm*', podríamos usarlo para no tener que asignar manualmente estados a las etiquetas.

En este caso la interfaz ***'02. Selección de la muestra'*** quedaría configurada como:



Un detalle del fichero de salida sería el que se muestra a continuación:



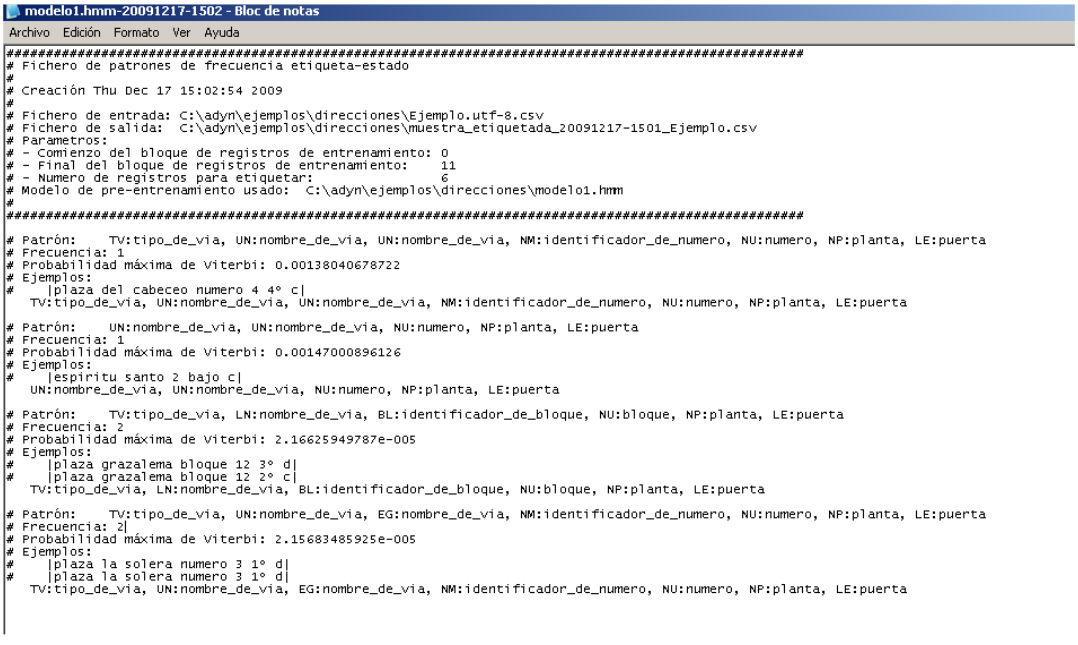
Como se puede observar al utilizar un modelo HMM ya creado se han asignado estados a las etiquetas automáticamente sin necesidad de hacerlo de forma manual. Además para cada

registro se muestra la probabilidad máxima de cada secuencia de etiquetas y estados asignados calculada mediante el algoritmo de Viterbi.

La muestra completa etiquetada y con estados asignados automáticamente se observa en la siguiente imagen:

```
# 0 (0): |pl/ la solera nº  3- 1º  d|
#        |plaza la solera numero 3 1º d|
   TV:tipo_de_via, UN:nombre_de_via, EG:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta
# Probabilidad máxima de Viterbi: 0.00002

# 1 (1): |pl/ la solera nº  3- 1º  d|
#        |plaza la solera numero 3 1º d|
   TV:tipo_de_via, UN:nombre_de_via, EG:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta
# Probabilidad máxima de Viterbi: 0.00002

# 2 (2): |pl/ del cabeceo nº  4- 4º  c|
#        |plaza del cabeceo numero 4 4º c|
   TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta
# Probabilidad máxima de Viterbi: 0.00138

# 3 (3): |pl/ grazalema     bl-12 - 3º  d|
#        |plaza grazalema bloque 12 3º d|
   TV:tipo_de_via, LN:nombre_de_via, BL:identificador_de_bloque, NU:bloque, NP:planta, LE:puerta
# Probabilidad máxima de Viterbi: 0.00002

# 4 (4): |pl/ grazalema bl-12- 2º  c|
#        |plaza grazalema bloque 12 2º c|
   TV:tipo_de_via, LN:nombre_de_via, BL:identificador_de_bloque, NU:bloque, NP:planta, LE:puerta
# Probabilidad máxima de Viterbi: 0.00002

# 6 (5): |espiritu santo 2 bajo c|
#        |espiritu santo 2 bajo c|
   UN:nombre_de_via, UN:nombre_de_via, NU:numero, NP:planta, LE:puerta
# Probabilidad máxima de Viterbi: 0.00147
```

A parte del fichero de la muestra etiquetada (fichero de extensión '*.csv*') obtenemos otro fichero de extensión '*.txt*' que contiene los patrones de frecuencia etiqueta-estado.
El contenido de fichero se muestra a continuación:


modelo1.hmm-20091217-1502 - Bloc de notas

Archivo Edición Formato Ver Ayuda

```
##############################################################################################
# Fichero de patrones de frecuencia etiqueta-estado
#
# Creación Thu Dec 17 15:02:54 2009
#
# Fichero de entrada: C:\adyn\ejemplos\direcciones\Ejemplo.utf-8.csv
# Fichero de salida:  C:\adyn\ejemplos\direcciones\muestra_etiquetada_20091217-1501_Ejemplo.csv
# Parametros:
# - Comienzo del bloque de registros de entrenamiento: 0
# - Final del bloque de registros de entrenamiento:    11
# - Numero de registros para etiquetar:                6
# Modelo de pre-entrenamiento usado:  C:\adyn\ejemplos\direcciones\modelo1.hmm
#
##############################################################################################

# Patrón:    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta
# Frecuencia: 1
# Probabilidad máxima de Viterbi: 0.00138040678722
# Ejemplos:
#    |plaza del cabeceo numero 4 4º c|
   TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta

# Patrón:    UN:nombre_de_via, UN:nombre_de_via, NU:numero, NP:planta, LE:puerta
# Frecuencia: 1
# Probabilidad máxima de Viterbi: 0.00147000896126
# Ejemplos:
#    |espiritu santo 2 bajo c|
   UN:nombre_de_via, UN:nombre_de_via, NU:numero, NP:planta, LE:puerta

# Patrón:    TV:tipo_de_via, LN:nombre_de_via, BL:identificador_de_bloque, NU:bloque, NP:planta, LE:puerta
# Frecuencia: 2
# Probabilidad máxima de Viterbi: 2.16625949787e-005
# Ejemplos:
#    |plaza grazalema bloque 12 3º d|
#    |plaza grazalema bloque 12 2º c|
   TV:tipo_de_via, LN:nombre_de_via, BL:identificador_de_bloque, NU:bloque, NP:planta, LE:puerta

# Patrón:    TV:tipo_de_via, UN:nombre_de_via, EG:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta
# Frecuencia: 2
# Probabilidad máxima de Viterbi: 2.15683485925e-005
# Ejemplos:
#    |plaza la solera numero 3 1º d|
#    |plaza la solera numero 3 1º d|
   TV:tipo_de_via, UN:nombre_de_via, EG:nombre_de_via, NM:identificador_de_numero, NU:numero, NP:planta, LE:puerta
```

En este caso para cada registro se muestra una serie de información al usuario que es adicional ya que cada línea comienza por #:

- Patrón o estructura que siguen los datos, indicándose la secuencia de etiquetas y estados asociados a ese patrón.
- Frecuencia de aparición de ese patrón en la muestra.
- Probabilidad máxima de la secuencia de etiquetas y estados asociados calculada mediante el algoritmo de Viterbi.
- Ejemplos de la muestra para los que se ha encontrado dicho patrón.

Y por último, se muestra la secuencia de etiquetas y estados asociados, que es la misma que la que aparece en el fichero '*.csv*' utilizado por la aplicación.

Anexo VII: Métodos de suavizado.

A la hora de construir el Modelo Oculto de Markov debemos tener en cuenta de que partimos de una muestra aleatoria del conjunto de datos que vamos a normalizar.

Por lo tanto podemos dejar fuera elementos cuya estructura sea diferente a las que se encuentran en la muestra y por lo tanto las probabilidades de observación de esas etiquetas y estados asociados serán nulas. Para solucionar este problema y que todas las etiquetas junto con sus estados asociados tengan una determinada probabilidad utilizamos los llamados **MÉTODOS DE SUAVIZADO**.

Los métodos de suavizado tratan de contrarrestar el efecto de los elementos compuestos por una etiqueta y un estado que no han aparecido en un conjunto de entrenamiento.

Se parte de una muestra aleatoria del conjunto de datos que vamos a analizar. Esta muestra se compone de registros donde cada uno de sus elementos o sucesos tiene asociado una etiqueta y un estado.

Dependiendo del tamaño de la muestra podemos haber dejado fuera de ella distintos elementos que también forman parte del conjunto de datos. Para asociar cierta probabilidad a los elementos no observados se han implementado dos de las técnicas que ofrecen mejores resultados, estas son, el suavizado de Laplace y la técnica de descuento absoluto (*absolute discounting*).

El **suavizado de Laplace** es un método de suavizado básico consistente en asignar una cierta probabilidad al espacio de sucesos no vistos mediante la aplicación de la *Ley de Laplace*, también conocida como *Añadir Uno* (*Adding One*, en inglés) (Jeffreys, 1948). En este método se incrementa la frecuencia de todos los sucesos en una unidad y la probabilidad de observación se define como:

$$P^{suavizada}(e_k \mid c_i) = \begin{cases} \dfrac{f(e_k, c_i)+1}{f(e_\epsilon)+|V|} & \text{si } f(e_k, c_i) \text{ existe} \\ \dfrac{1}{f(e_\epsilon)+|V|} & \text{si } f(e_k, c_i) \text{ no existe} \end{cases}$$

donde *V* es el número de etiquetas que aparecen en el corpus de entrenamiento, *f(ek,ci)* es la cantidad de veces que el estado e*k* está etiquetada con *ci* y *f(ek)* es el número total de etiquetas que tiene asociado el estado e*k* en el corpus de entrenamiento.

En la técnica de **descuento absoluto** (*absolute discounting*) se sustrae un valor pequeño, digamos *'x',* de la probabilidad de todas las etiquetas c_j vistas en el estado *j* (probabilidad ≠ 0). Entonces se distribuye la probabilidad acumulada equitativamente entre los sucesos no conocidos. Así la probabilidad de una etiqueta no conocida es:

$$\frac{c_j x}{c - c_j}$$

Donde 'c' es el número total de etiquetas, mientras que para una etiqueta conocida, su probabilidad será:

$$b_{jk} - x$$

donde b_{jk} es el cociente entre el número de veces que el estado 'k' tiene asociada la etiqueta 'j' y el total de estados asociados a la etiqueta 'j'.

No hay ninguna teoría sobre cómo seleccionar el mejor valor para *x,* por lo hemos decidido tomar el siguiente valor:

$$X = \frac{1}{f(e_k) + |V|}$$

donde *V* es el número de etiquetas que aparecen en el corpus de entrenamiento y *f(ek)* es el número total de etiquetas que tiene asociado el estado *ek* en el corpus de entrenamiento.

Anexo VIII: Manual de Instalación.

La instalación de *ADYN Herramienta de Normalización* se realizará de forma automática con un asistente que guiará al usuario en el proceso de instalación, solicitándole los directorios donde desea tener instalada la aplicación. Para un correcto funcionamiento de la herramienta se recomienda su instalación en el directorio raíz ('C:\', 'D:\', etc.) para evitar posibles errores al trabajar en directorios cuyas rutas sean muy extensas.

Otra especificación será evitar la inclusión de tildes en las rutas de acceso a los archivos de trabajo.

Hay que indicar que aunque el asistente de instalación pueda parecer algo tedioso y repetitivo es necesario ejecutarlo de forma completa para el correcto funcionamiento de la *ADYN Herramienta de Normalización*.

Esta aplicación está desarrollada en lenguaje de programación Python por lo que para poder utilizarla es necesario tener instalado el programa Python. Por este motivo al instalar *ADYN* con el asistente automáticamente queda instalado por defecto dicho programa.

Si el usuario ya tiene instalado Python puede haber un conflicto al realizar la instalación de *ADYN*, ya que esta herramienta requiere que tanto Python como los programas auxiliares que se suministran se instalen en el mismo directorio. Por lo tanto, se recomienda desinstalar la versión de Python que el usuario pueda tener instalada y realizar el proceso completo de instalación a través del asistente.

El software que se instalará de forma automática con *ADYN Herramienta de Normalización* es el siguiente:

- Python 2.5, disponible en http://www.python.org
- GTK+ 2.14, disponible en http://www.pygtk.org
- Instaladores para Windows, disponibles en http://gtk-win.sourceforge.net/

También se instalarán los siguientes módulos de Python:

- Chardet, disponible en http://chardet.feedparser.org/
- pyGTK, disponible en http://www.pygtk.org/
- pycairo, disponible en http://www.cairographics.org/pycairo/
- pygobject, disponible junto con pyGTK.

Finalmente, una vez instalada *ADYN Herramienta de Normalización* podemos acceder a sus distintas interfaces o ventanas a través de los accesos directos que se han creado en el menú de inicio. Las interfaces se corresponden a los siguientes módulos:

- **01.Normalizador**: corresponde al módulo 'python estandarizador.py'.
- **02.Seleccion de la muestra**: corresponde al módulo 'python HMM_etiquetado.py'.
- **03.Entrenamiento de la muestra**: corresponde al módulo 'python HMM_entrenamiento.py'.
- **04.Editor de listas de corrección**: corresponde al módulo 'python editor.py lst' (el módulo se encuentra en la carpeta 'editor').
- **05.Editor de tablas de búsqueda**: corresponde al módulo 'python editor.py tbl' (el módulo se encuentra en la carpeta 'editor').

Otra forma de acceder a estas interfaces o ventanas es ejecutando los módulos correspondientes a estas interfaces a través del intérprete de Python que el usuario tenga instalado o si el sistema está integrado con Python, haciendo doble click sobre esos módulos.

Hay que notar que en *Debian* y distribuciones derivadas como *Ubuntu* y *GuadaLinex*, solo necesitará instalar los siguientes paquetes: python, libgtk2.0-bin, python-gtk2 y python-chardet.

Anexo IX: Recomendación sobre la denominación del campo a normalizar.

Es imprescindible que en el fichero que vayamos a normalizar la denominación de los campos no coincidan con los nombres correspondientes a los estados asociados al Modelo Oculto de Markov ni con la denominación del campo 'validacion'.

Por ejemplo, si estamos normalizando un fichero de direcciones postales y en él aparece un campo llamado 'nombre_de_via', hemos de modificar esta denominación ya que ésta corresponde a la denominación de uno de los estados del Modelo Oculto de Markov.

Además, también es requisito necesario que las denominaciones de los campos del fichero de datos no tengan tildes para evitar errores de lectura del fichero que contiene el campo a normalizar.