Guía rápida de aprendizaje de

ADYN Herramienta de 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:

Página

Introducción……1

Parte 1: Creación del Modelo Oculto de Markov……3

Paso 1: Selección y etiquetado de la muestra……3

Paso 2: Asignación manual de estados……9

Paso 3: Entrenamiento de la muestra……12

Parte 2: Normalización del fichero de datos……14

Parte 3: Validación del proceso de normalización……19

Anexo I: Manual de instalación……23

Anexo II: Listas de corrección y tablas de búsqueda……25

Anexo III: Etiquetas y estados definidos para direcciones postales……31

Anexo IV: Recomendaciones de uso……34

Introducción

El objetivo de esta Guía Rápida es proporcionar al usuario de *ADYN Herramienta de Normalización* un manual sencillo para realizar de forma eficiente el proceso de normalización de direcciones postales, de nombres de personas y de identificadores de personas físicas y jurídicas.

Para profundizar en los diferentes aspectos que conforman cada uno de los procesos de normalización anteriores sugerimos al usuario que consulte la *Guía General de Aprendizaje de ADYN Herramienta de Normalización*, que también se encuentra disponible, junto con este manual, en el Repositorio de Software Libre de la Junta de Andalucía.

Para agilizar el proceso de aprendizaje por parte del usuario realizaremos, paso a paso, un proceso de normalización de direcciones postales utilizando un fichero de ejemplo que se adjunta con la aplicación en la carpeta 'C:\adyn\ejemplos\direcciones'. La normalización de nombres de personas y de identificadores de personas físicas y jurídicas seguirá un proceso similar.

Para empezar a trabajar lo primero que haremos es instalar nuestra aplicación (**Anexo I**) y seguidamente analizaremos el fichero que vamos a utilizar a modo de ejemplo, '*Ejemplo.csv*'. Este fichero consta de un único campo llamado 'Direcciones' que contiene las direcciones postales que deseamos normalizar y el número de registros que lo componen es 11.

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'

Notar que cuando trabajemos con ficheros de datos reales, el campo que deseamos normalizar, direcciones postales, no será, como en el ejemplo, 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 'Direcciones' segmentado en tantos campos como partes componen una dirección postal con la particularidad de que estos nuevos campos ya no presentarán errores ni inconsistencias.

Por lo tanto, 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 lo tanto, la estructura de esta *Guía Rápida* queda dividida principalmente en tres partes, que corresponden a las fases seguidas en un proceso de normalización de datos a través de *ADYN Herramienta de Normalización*:

- **Parte 1: Creación del Modelo Oculto de Markov.**
- **Parte 2: Normalización del fichero de datos.**
- **Parte 3: Validación del proceso de normalización.**

A continuación se explica cada una de ellas.

Parte 1: Creación del Modelo Oculto de Markov

Los Modelos Ocultos de Markov (en inglés *Hidden Markov Models* ó HMM) permiten reconocer la estructura de los datos con los que trabajamos por lo que resultarán imprescindibles para las tareas de segmentación de campos que conforman los procesos de normalización. Será necesario construir estos modelos en los procesos de normalización de direcciones postales y de nombres de personas pero no en el de identificadores de personas físicas y jurídicas, ya que en este caso se propone un modelo que se considera representa bastante bien toda la casuística que podemos encontrar sobre esta materia.

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.

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 y etiqueta los elementos que componen dicho campo. Para este paso usaremos la interfaz ***'02. Selección de la muestra'***, a la cual accederemos a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) tal y como se muestra en la siguiente imagen:

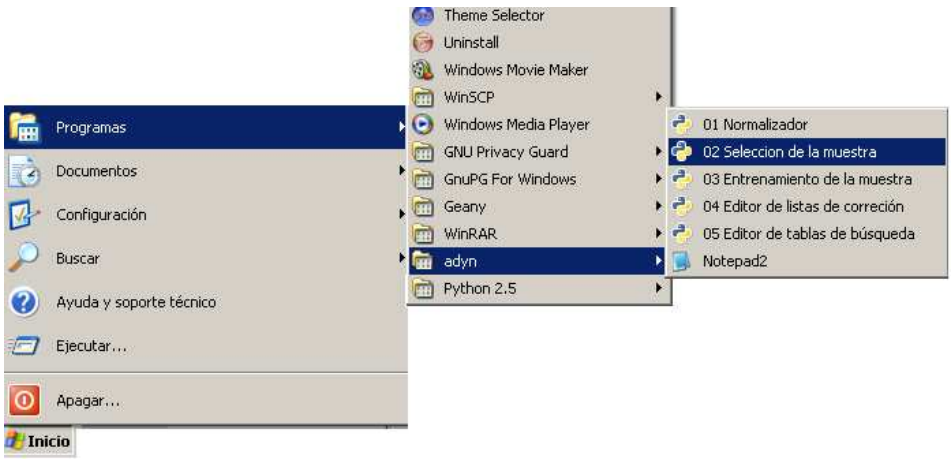

Imagen 3: Acceso a la interfaz '02. Selección de la muestra'.

Una vez abierta la interfaz nos recibirá la siguiente pantalla:

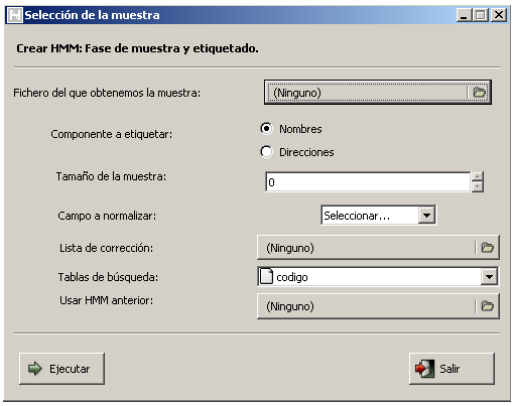

Imagen 4: Interfaz de selección de la muestra y etiquetado de componentes.

Antes de continuar hemos de indicar que esta pantalla será análoga para el caso de seleccionar una muestra y realizar el etiquetado de nombres de personas, ya que lo único que variará será el tipo de '**Componente a etiquetar**' elegida.

Continuando con nuestro ejemplo lo primero que debemos seleccionar es el fichero del que 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:

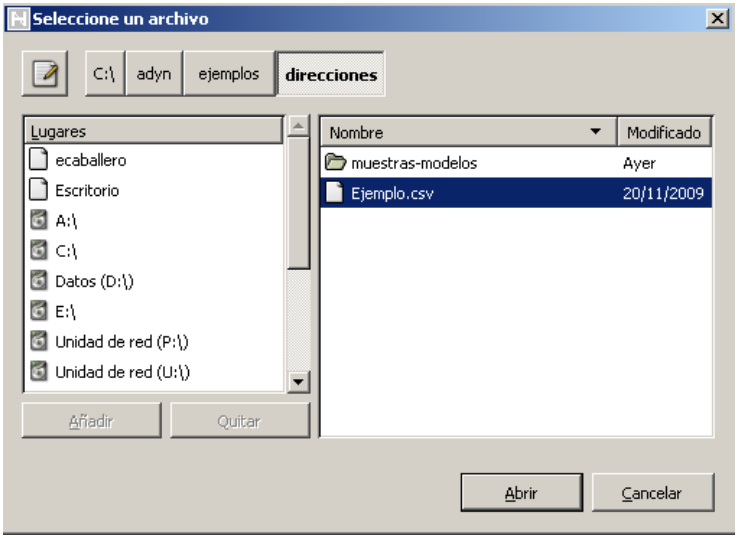

Imagen 5: Navegador de archivos.

Tras seleccionar el fichero *'Ejemplo.csv'* haremos click en 'Abrir' (o simplemente doble click sobre el nombre del fichero) y quedará seleccionado en nuestra interfaz.

Seguidamente 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 heteregoneidad 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.

A continuación en el cuadro combinado **'Campo a Estandarizar'** se listan todos los campos que tiene el fichero *'Ejemplo.csv'* y seleccionaremos el campo que contiene las direcciones postales que vamos a normalizar. En nuestro caso será el campo '*direcciones*'.

La interfaz está quedando configurada de la siguiente forma:

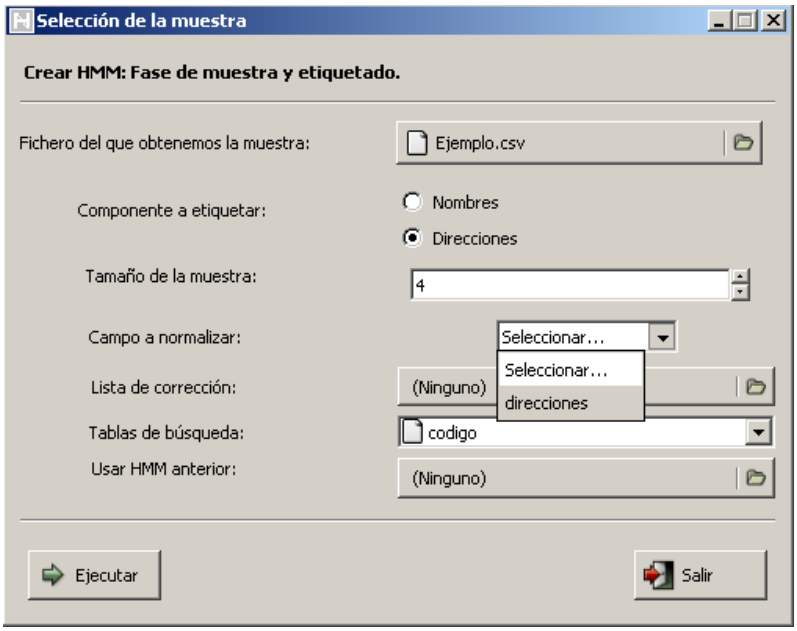

Imagen 6: Selección del campo a normalizar.

Por último, necesitamos seleccionar la lista de corrección y las tablas de búsqueda.

Las listas de corrección permiten limpiar el fichero de datos, es decir, son ficheros que contienen los caracteres que queremos eliminar o sustituir en los ficheros, por ejemplo, eliminar los caracteres extraños ('|', '$',...) y sustituir las vocales con tildes por vocales sin tildes.

Las tablas de búsqueda son otro tipo de ficheros que además de sustituir un elemento del fichero de datos por su valor normalizado, le asigna una etiqueta, por ejemplo, si la aplicación encuentra el elemento 'c/' lo sustituye por 'calle' y le asigna la etiqueta 'TV' que significa Tipo de Vía.

Tendremos listas de corrección y tablas de búsqueda de nombres de personas, de direcciones postales y de identificadores de personas físicas y jurídicas y podrán ser personalizadas y modificadas por el usuario. Estos ficheros se encuentran dentro del directorio '*datos*' de nuestra aplicación, en concreto:

- El directorio *'C:\adyn\codigo\datos\ListasDeCorreccion'* contiene la lista de corrección para direcciones postales, *direcciones_correccion.lst,* para nombres de personas, *nombres_correccion.lst* y para identificadores de personas físicas y jurídicas *idpersona_correccion.lst*, aunque ésta última sólo se utilizará en la interfaz de normalización (***01. Normalizador***).

- El directorio *'C:\adyn\codigo\datos'* contiene las tablas de búsqueda para nombres, *nombre-tbl*, para direcciones postales, *direccion-tbl*, y para identificadores de personas físicas y jurídicas, *idpersona-tbl*, aunque al igual que en el caso anterior ésta última sólo será utilizada en la interfaz de normalización.

La visualización o modificación de las listas de corrección y las tablas de búsqueda se hará según lo descrito en el **Anexo II**.

En nuestro ejemplo seleccionaremos la lista de corrección y la carpeta que contiene las tablas de búsqueda de direcciones postales:

Para la **'Lista de corrección'** abrimos el cuadro de diálogo para seleccionar el archivo '*direcciones_correccion.lst'* que se encuentra dentro de la carpeta '*ListasDeCorreccion*':

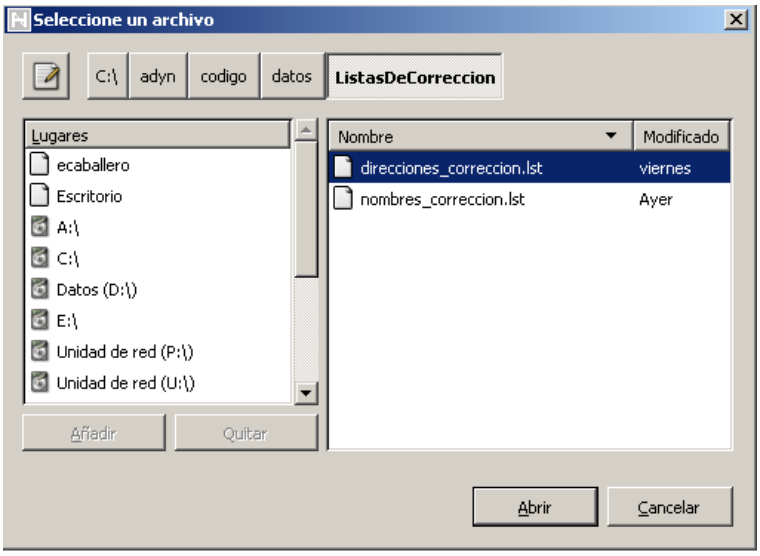

Imagen 7: Selección de la carpeta que contiene la lista de corrección de direcciones postales.

Para las **'Tablas de búsqueda**' debemos seleccionar en el desplegable la opción 'Otro' y buscamos nuestra 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.

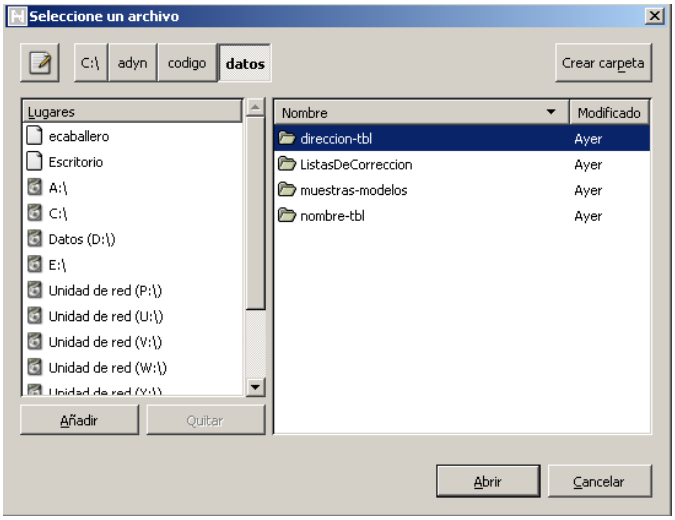

Imagen 8: Selección de las tablas de búsqueda de direcciones postales.

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 apartado 'c)' del **Anexo IV** 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 de ***'02.Selección de la muestra'***:

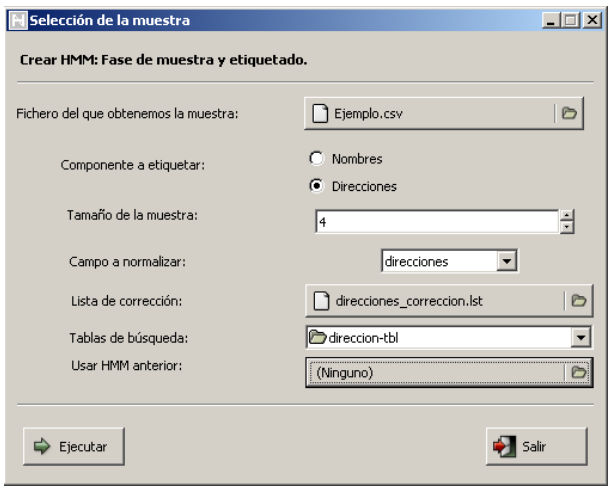

Imagen 9: Interfaz de Selección de la muestra.

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

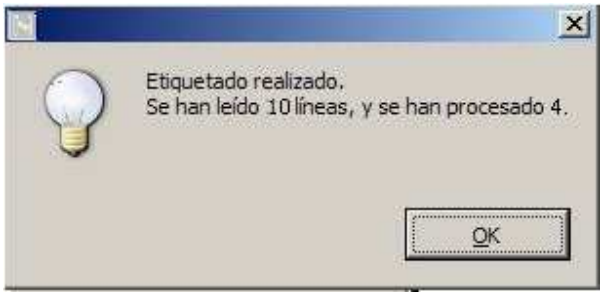

Imagen 10: Ventana de finalización del proceso de selección de la muestra.

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 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 desde el fichero *'Ejemplo.csv'*.

Paso 2: Asignación manual de estados.

En este paso vamos a trabajar con el fichero de salida del paso anterior, '*muestra_etiquetada_20091030-1241_Ejemplo.csv*'. Esta fase será **siempre manual** y requerirá intervención del usuario para asociar a cada etiqueta, del fichero de la muestra etiquetada, su estado correspondiente.

Así pues, para realizar esta asignación de estados editaremos el fichero '*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'.

El fichero con la muestra etiquetada tendrá un contenido similar a:

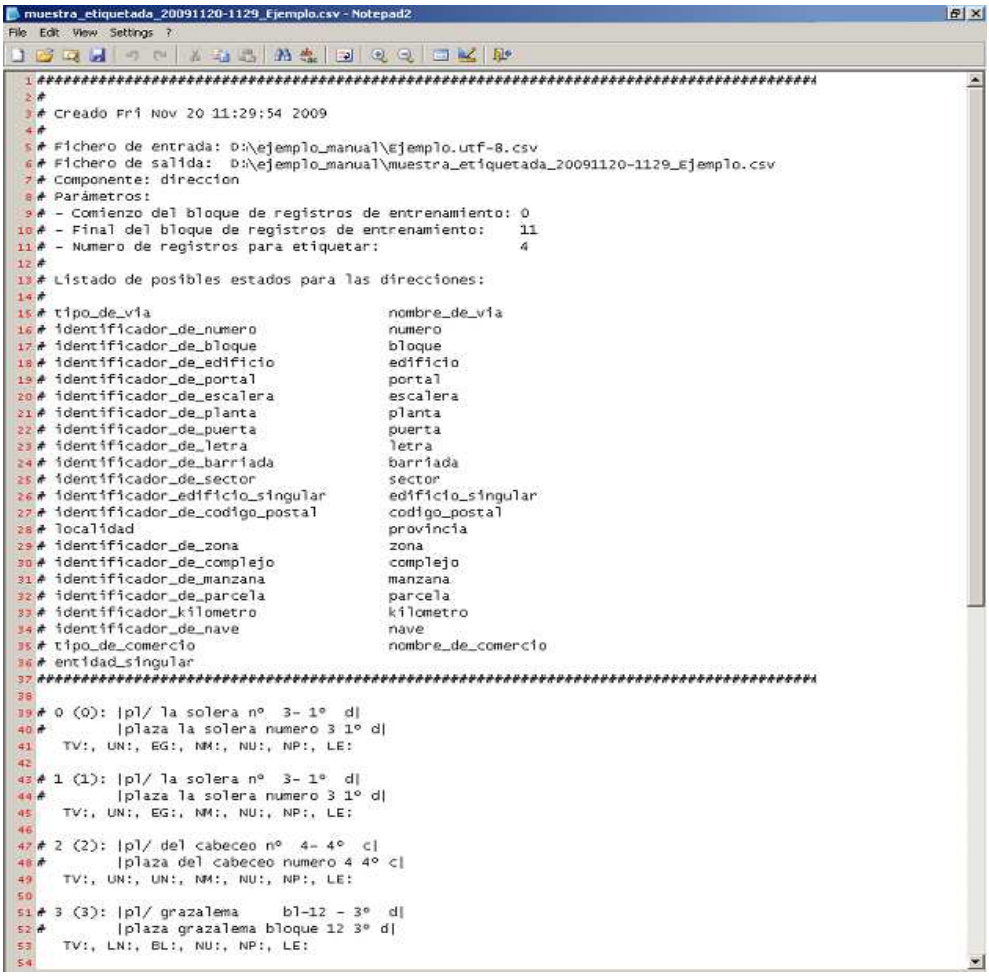

Imagen 11: Fichero con las componentes de la muestra etiquetadas.

Antes de analizar la estructura del fichero anterior, es necesario indicar que toda la información que se encuentre tras una almohadilla '#' no será leída por la aplicación porque se considera un comentario. Así pues, cuando queramos incluir información que consideremos que puede aclarar el proceso manual de asignación de estados a las etiquetas, lo que tendremos que hacer es escribir '#' y a continuación la información que consideremos aclaratoria.

Pasamos a continuación a analizar el formato del fichero con la muestra etiquetada. La primera parte del fichero rodeada de almohadillas '#' contiene información sobre el fichero de origen (aparece la ruta de ese fichero de origen al que se le ha realizado automáticamente un cambio de codificación a UTF-8), el fichero de salida (con la ruta donde se encuentra almacenado el fichero con la muestra etiquetada que hemos creado), la fecha de creación y los parámetros de selección.

Debajo y **en dos columnas**, tenemos la lista de posibles estados (45) que se pueden asignar a cada una de las etiquetas. La definición de cada una de estas etiquetas y estados se encuentra en el **Anexo III**.

Posteriormente tenemos los 4 registros de la muestra etiquetados. Nótese que al ser una muestra con reposición se ha seleccionado 2 veces el mismo registro (registros 0 y 1). Para cada registro se muestra la siguiente información:

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

La primera línea contiene el registro del fichero original, la segunda el registro después de la limpieza y estandarización y la última las etiquetas asignadas a sus componentes, resultado de la interfaz *'**02. Selección de la muestra**'* que hemos visto en el Paso 1. A cada etiqueta (en el ejemplo: TV, UN, EG, NM, UN, NP, LE) hay que asignarle el estado correspondiente. De esta 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’.

Tras asignar estos estados, 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 los cuatro registros de la muestra, el resultado es el siguiente:

```
# 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


# 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

# 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

# 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
```

Imagen 12: 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.

Hemos de indicar que es posible que para un mismo registro tengamos dos secuencias de etiquetas posibles, por ejemplo si tuviéramos la siguiente dirección postal:

C/ Luna Sevilla

El etiquetado automático de sus elementos podría quedar como:

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 lo tanto el usuario deberá eliminar aquella secuencia que considere improbable y se quedará únicamente una secuencia asociada al registro.

Finalmente, para que el usuario pueda asignar de forma adecuada los estados a las etiquetas suministramos en el **Anexo III** la relación y descripción de todos los posibles estados.

Paso 3: Entrenamiento de la muestra.

En este paso crearemos el Modelo Oculto de Markov (HMM) usando el fichero resultante del Paso 2. Para ello usaremos la interfaz ***'03. Entrenamiento de la muestra'*** que nos creará un fichero de extensión *'.hmm'.* Este fichero contendrá tres matrices que configuran el Modelo Oculto de Markov (para obtener más información sobre ellas se recomienda consultar la *Guía General de Aprendizaje de 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:

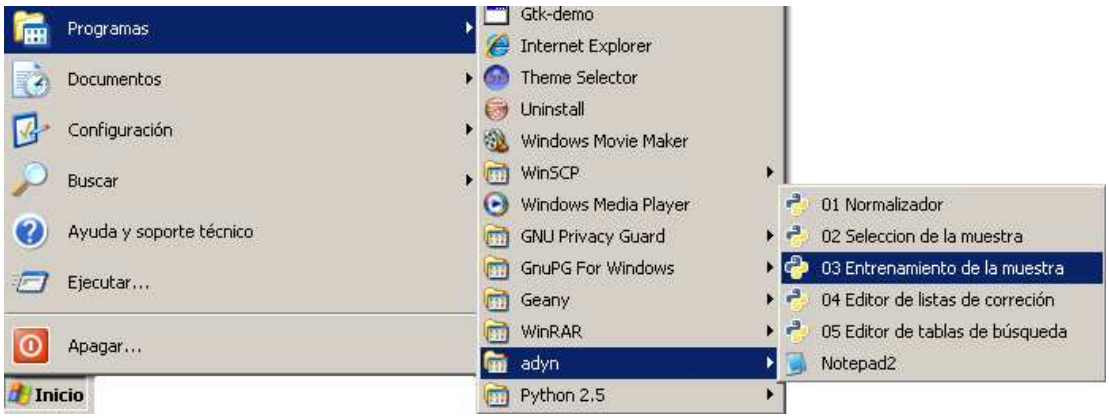

Imagen 13: Acceso a la interfaz '03. Entrenamiento de la muestra'.

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

Imagen 14: Interfaz de entrenamiento de la muestra o de creación del HMM.

Lo primero que haremos será seleccionar el fichero obtenido en el paso anterior (**'Fichero de la muestra etiquetada'**), seguidamente hemos de marcar si la componente a normalizar es un nombre o una dirección postal (**'Selecciona componente'**), en nuestro caso '**Direcciones**' y el método de suavizado que deseamos usar (**'Selecciona método de suavizado'**). En la *Guía General de Aprendizaje de ADYN Herramienta de Normalización* se explican las diferencias teóricas entre los diferentes métodos de suavizado. En nuestro ejemplo, no usaremos ninguno.

Ya solo nos queda hacer click en **'Ejecutar'** y esperar hasta que el programa nos comunique que ha terminado con la siguiente pantalla:

Imagen 15: 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 con la 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',* en nuestro caso lo hemos denominado '*modelo1.hmm*'.

Parte 2: Normalización del fichero de datos.

Ahora que ya tenemos creado un Modelo Oculto de Markov para la normalización de los datos, podemos usar la interfaz ***'01. Normalizador'***, a la cual accedemos a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) tal y como se muestra en la siguiente imagen:

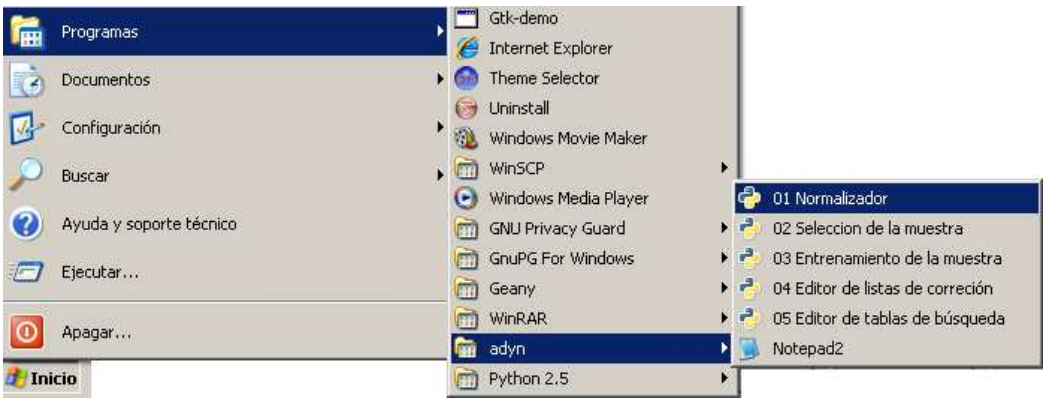

Imagen 16: Acceso a la interfaz '01. Normalizador'.

El resultado es la siguiente interfaz:

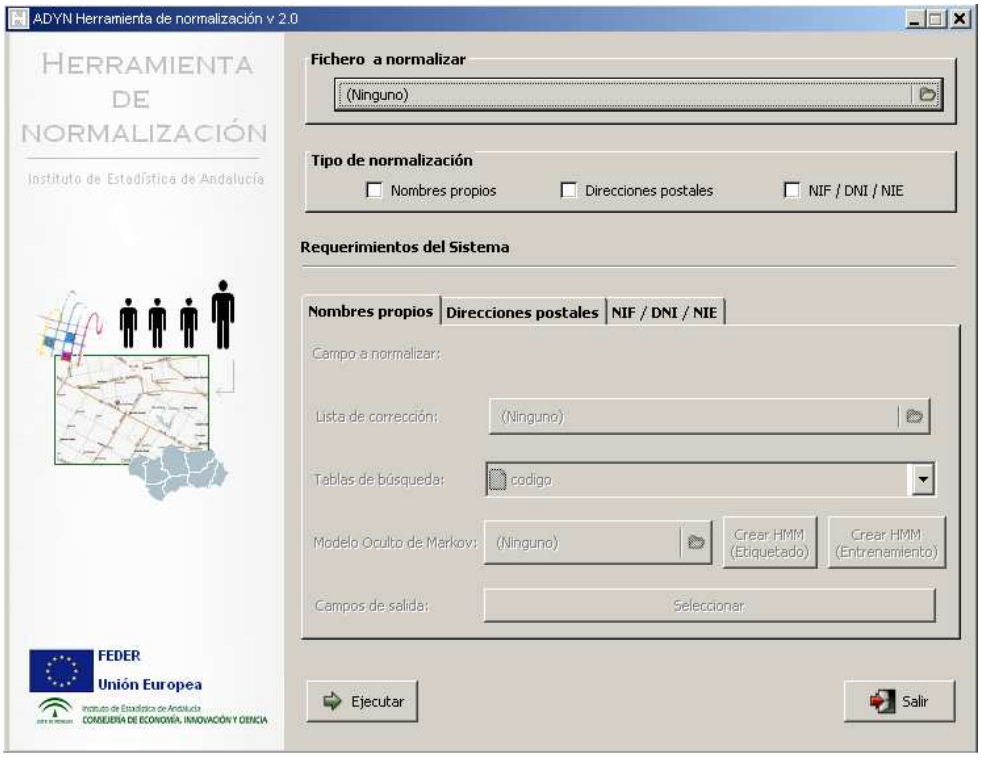

Imagen 17: ADYN Herramienta de Normalización: Interfaz de normalización.

Seleccionaremos el fichero que vamos a normalizar '*Ejemplo.csv*' y marcaremos el '**Tipo de normalización**' que queremos realizar, en nuestro caso, Direcciones postales. Como vemos en la siguiente imagen, al seleccionar el **'Tipo de normalización'** de direcciones postales se habilita la pestaña 'Direcciones postales' para que cumplimentemos los requerimientos del Sistema.

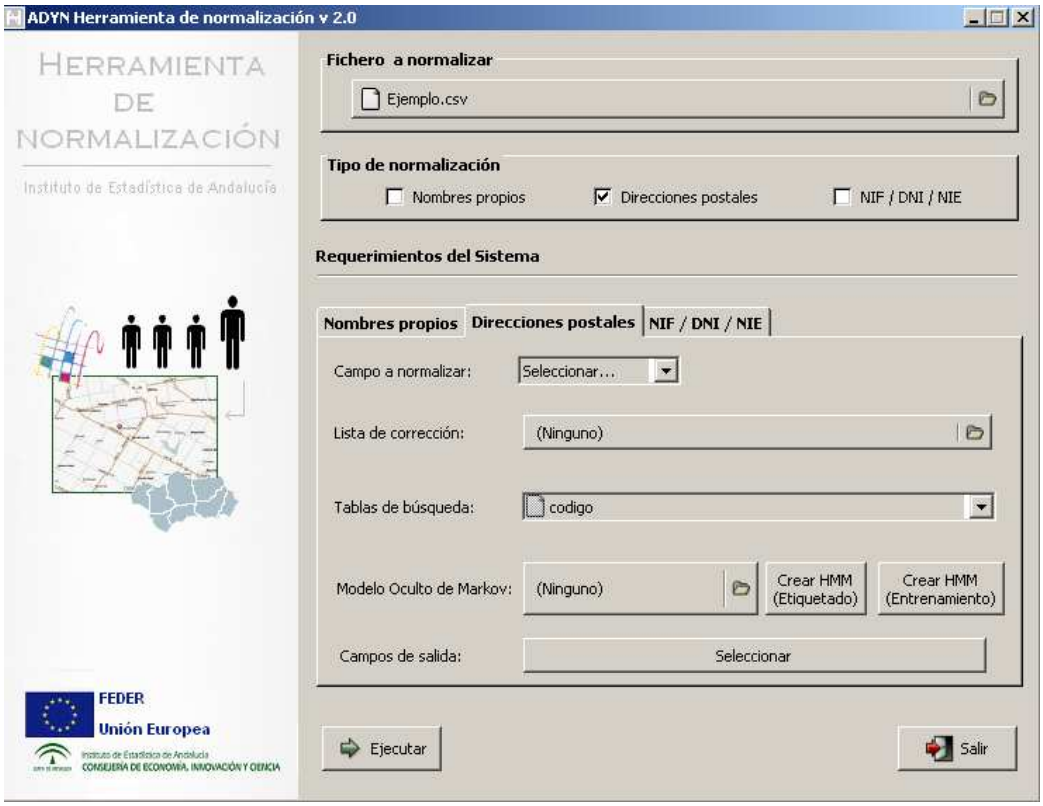

Imagen 18: Interfaz de normalización una vez seleccionado el tipo de normalización.

A continuación tendremos que especificar el campo del fichero que deseamos normalizar '*direcciones*', la lista de corrección, las tablas de búsqueda y el modelo HMM. El proceso de elección de la lista de corrección y las tablas de búsqueda es idéntico al de la interfaz de etiquetado y se han de usar los mismos ficheros. En cuanto al Modelo Oculto de Markov, habrá que incluir el que hemos creado previamente, modelo al que hemos llamado por comodidad de interpretación '*modelo1.hmm*' y nuestra interfaz quedará como:

Imagen 19: Interfaz de normalización tras seleccionar la lista de corrección y las tablas de búsqueda.

Hemos de notar que si en el caso de normalizar direcciones postales o nombres de personas no hubiéramos creado previamente el Modelo Oculto de Markov a través de las interfaces de selección y etiquetado de la muestra así como la de entrenamiento, no habría ningún problema ya que mediante la interfaz de normalización podremos acceder directamente a ellas '**Crear HMM (Etiquetado**') y '**Crear HMM (Entrenamiento)**'.

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. Indicar que estos campos son las componentes o elementos que conforman una dirección postal.

Imagen 20: Selección de los campos de salida de direcciones postales.

Podremos desmarcar aquellos campos que no queremos que se muestren en el fichero de salida pero por defecto aparecerán marcados todos. Una vez elegidos (todos o los seleccionados por el usuario) pulsaremos 'OK'.

A continuación pulsaremos **'Ejecutar',** esperaremos unos segundos (o unos minutos si el fichero es grande) y la interfaz nos avisará cuando se hayan normalizado todos los registros:

Imagen 21: 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 de direcciones estandarizado y segmentado en los campos que se han seleccionado previamente.
- 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 ejecutarse correctamente este fichero deberá guardarse en la carpeta '*codigo***'** de la aplicación *ADYN*, y hacer doble clic sobre él.

Parte 3: Validación del proceso de normalización.

Al abrir el fichero de datos normalizados, podrá apreciarse una columna extra llamada 'validacion' con 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 23: Fichero normalizado donde se muestran varios campos de salida, entre ellos el 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 habían introducido en la muestra con la que se generó el Modelo Oculto de Markov, utilizado para normalizar el fichero o bien existen valores que no están incluidos en las tablas de búsqueda y por lo tanto no pueden ser reconocidos por el Modelo Oculto de Markov.

Supongamos el siguiente ejemplo; si al realizar el proceso de normalización hemos obtenido que el registro:

'*espiritu santo 2 bajo*'

no está bien normalizado, es decir, el campo de validación tiene asignado el valor 1, entonces con el fin de ir corrigiendo esos errores y construir un Modelo Oculto de Markov más eficiente, o lo que es lo mismo, que contenga más estructuras de datos, podemos tomar dos caminos:

a) 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 con la muestra etiquetada y en la que hemos asignado sus estados correspondientes al principio del proceso:

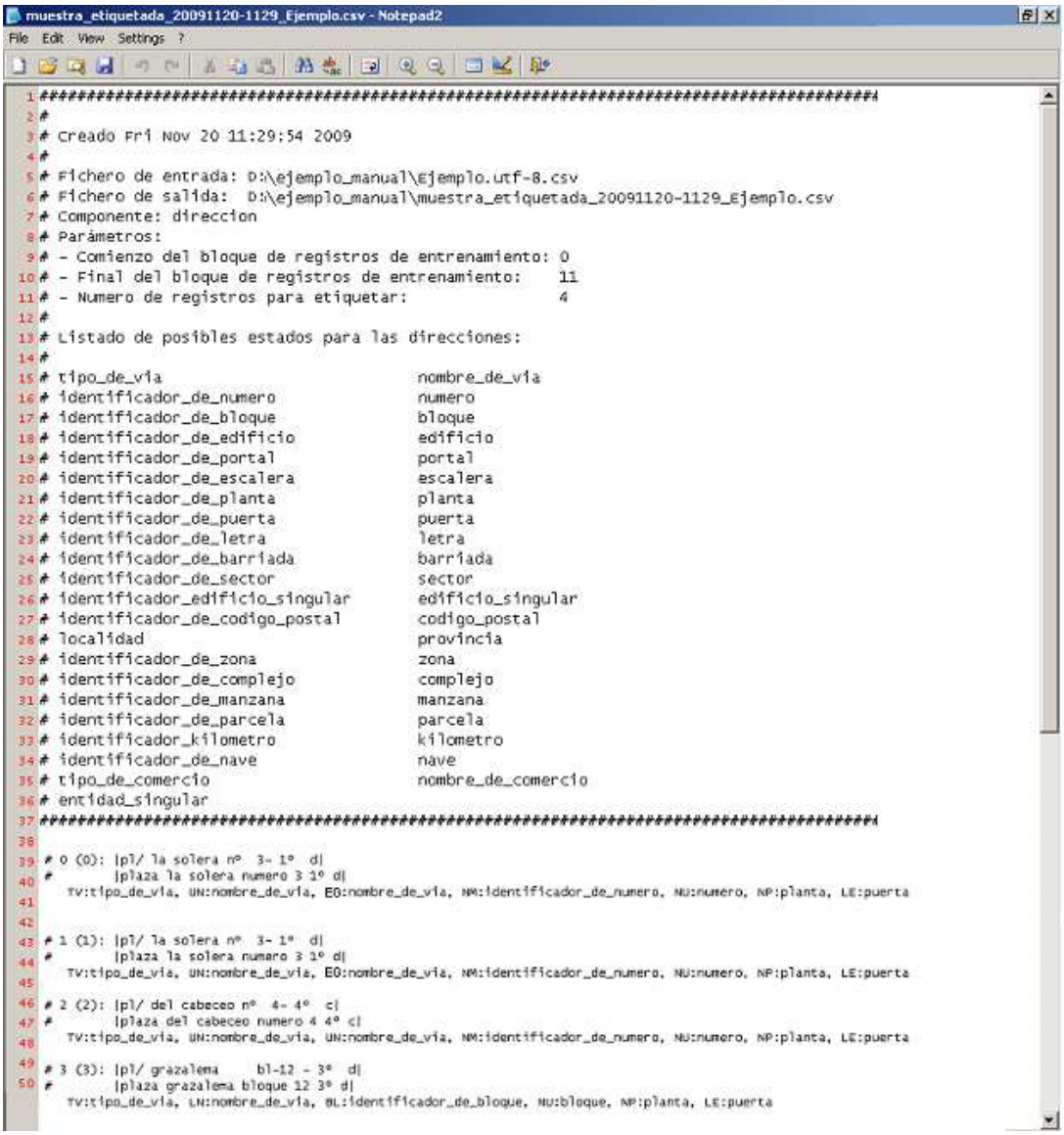

Imagen 24: Fichero de la muestra etiquetada con estados asignados.

2. Introduciremos la estructura o patrón correspondiente al registro:

'espiritu santo 2 bajo'

Y por el conocimiento que tenemos sobre los otros registros del fichero original y sobre las etiquetas y sus posibles estados, podemos asignarle las siguientes etiquetas:

espiritu	santo	2	bajo
UN: ,	UN: ,	NU: ,	PL:

Ahora añadiremos los estados a cada etiqueta:

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 en el fichero de la muestra etiquetada aparecerá de la forma:

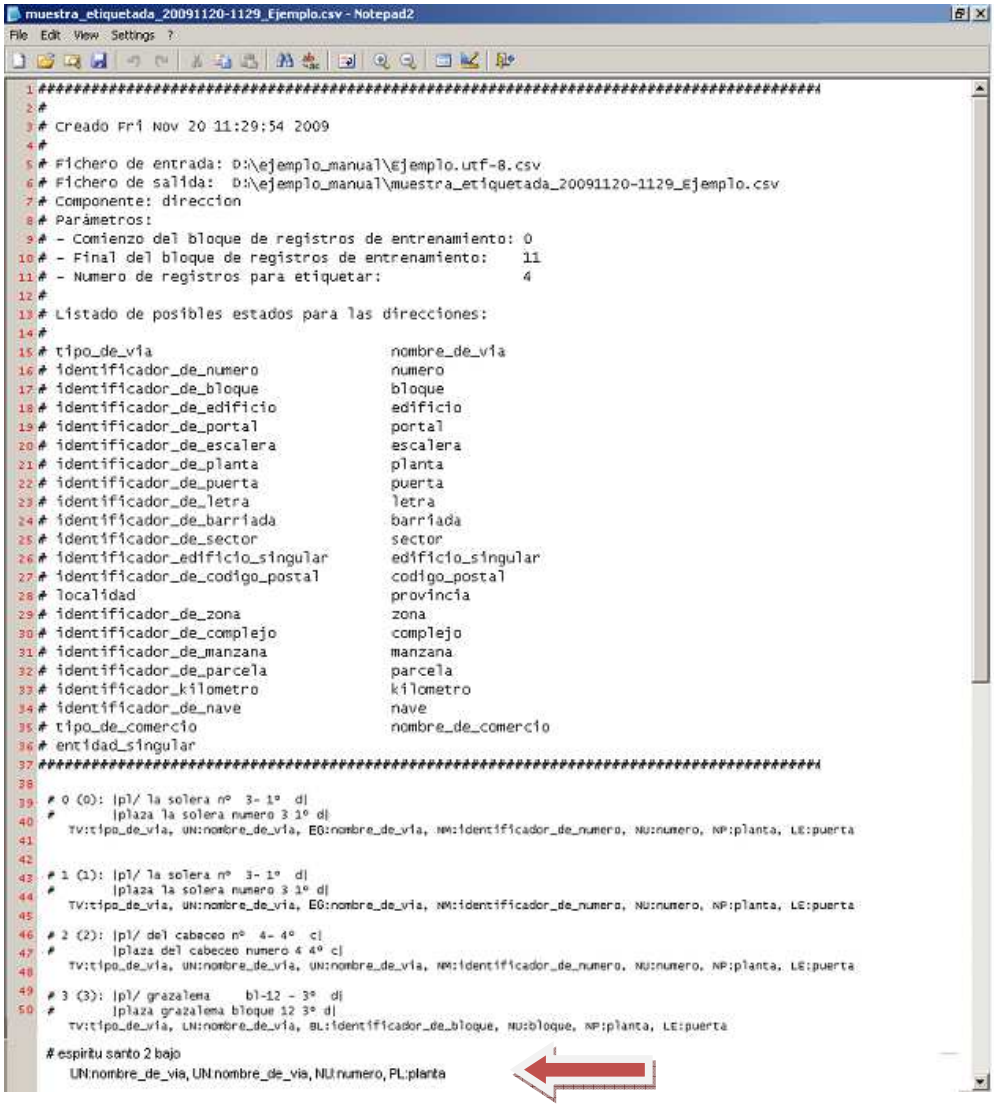

Imagen 25: Fichero de la muestra etiquetada con estados asignados y un nuevo patrón.

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, por ejemplo:

#espiritu santo 2 bajo

A continuación se volverá a construir el Modelo Oculto de Markov utilizando la interfaz ***'03. Entrenamiento de la muestra'*** y esta nueva muestra etiquetada. Seguidamente utilizando la interfaz ***'01. Normalizador'*** volveremos a realizar el proceso de normalización el fichero de datos original con este nuevo modelo.

b) Crear un fichero de datos que contenga los registros que se encuentren mal normalizados y a través de la interfaz ***'02. Selección de la muestra'*** seleccionar una muestra de esos registros. En este caso la componente a normalizar quedará etiquetada y manualmente se introducirán los estados correspondientes.

   Posteriormente, incluimos la muestra etiquetada generada en el Paso 1 (***'Selección y etiquetado de la muestra'***) en este segundo fichero. 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 muestra etiquetada el nuevo Modelo Oculto de Markov.

La experiencia determina que el camino más eficiente es el descrito en a), sin embargo dejamos al usuario la determinación del mismo según estime conveniente.

Anexo I: 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 II: 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. 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:

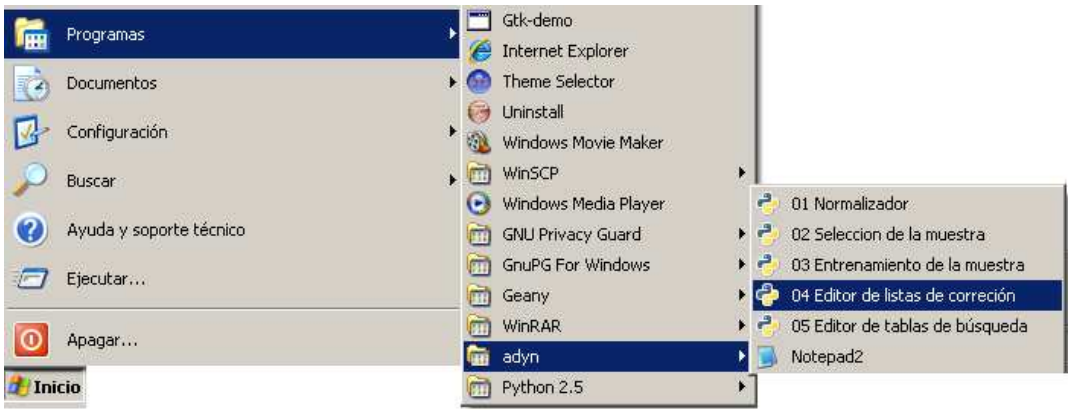

Imagen 26: Acceso a '04. Editor de listas de corrección' y '05. Editor de tablas de búsqueda'.

Detenidamente tenemos:

a) Las listas de corrección:

Como hemos comentado anteriormente permiten limpiar y corregir el fichero de datos. Son ficheros que contienen los caracteres que queremos eliminar o sustituir en los ficheros, por ejemplo, eliminar los caracteres extraños ('%', '?', etc.) y sustituir las vocales acentuadas por las vocales sin acentos. Para visualizarla hacemos click sobre ***'04. Editor de listas de corrección'*** y aparecerá una pantalla del tipo:

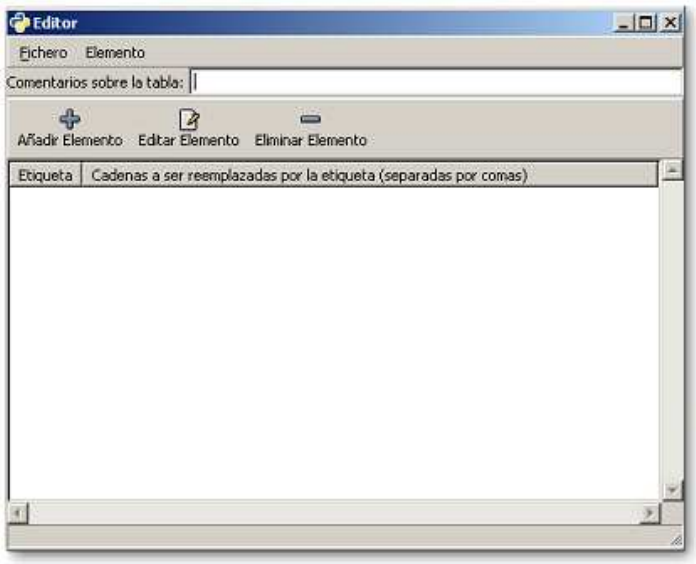

Imagen 27: Editor de listas de corrección.

Para visualizar la lista de corrección para direcciones postales habrá que hacer click en el menú 'Fichero' y entrar en el directorio 'C:\adyn\codigo\datos\listasdecorrección.lst'. Siendo el resultado:

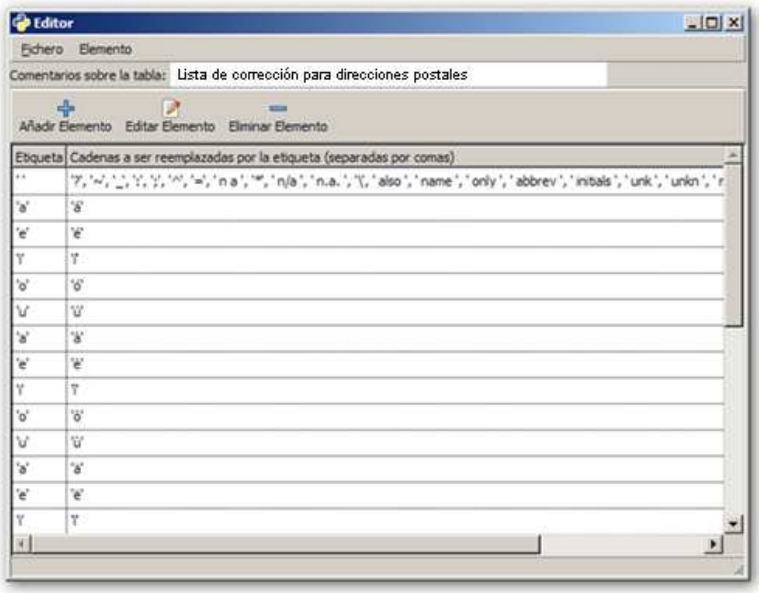

Imagen 28: Lista de corrección de direcciones postales.

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

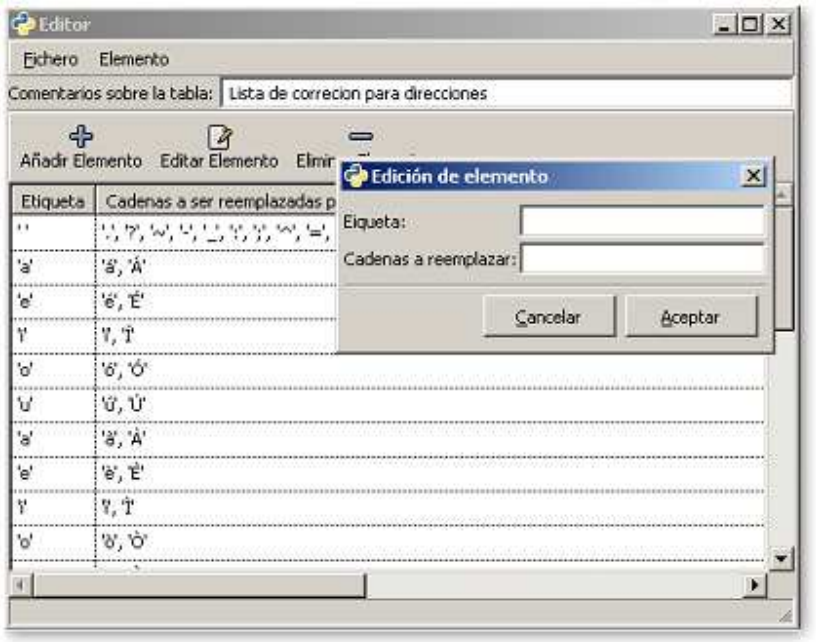

Imagen 29: Añadiendo un nuevo elemento a la lista de corrección de direcciones postales.

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 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 que además de sustituir un elemento del fichero de datos por su valor estandarizado, le asigna una etiqueta. Por ejemplo, si la aplicación encuentra en el campo direcciones el elemento 'polg' lo sustituye por 'polígono industrial' y le asigna la etiqueta 'ZO' que significa Zona. Para visualizarlas hacemos click sobre ***'05. Editor de tablas de búsqueda'*** y aparecerá una pantalla del tipo:

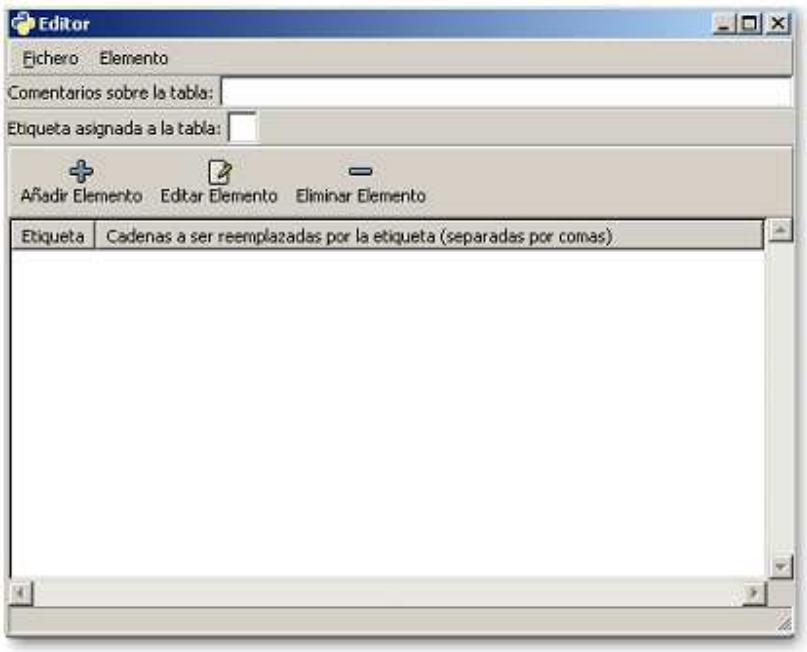

Imagen 30: Editor de tablas de búsqueda.

Para visualizar las tablas de búsqueda para direcciones postales habrá que hacer click en el menú 'Fichero' y entrar en el directorio 'C:\adyn\codigo\datos\direccion-tbl'. Siendo el resultado el que se muestra en la siguiente imagen:

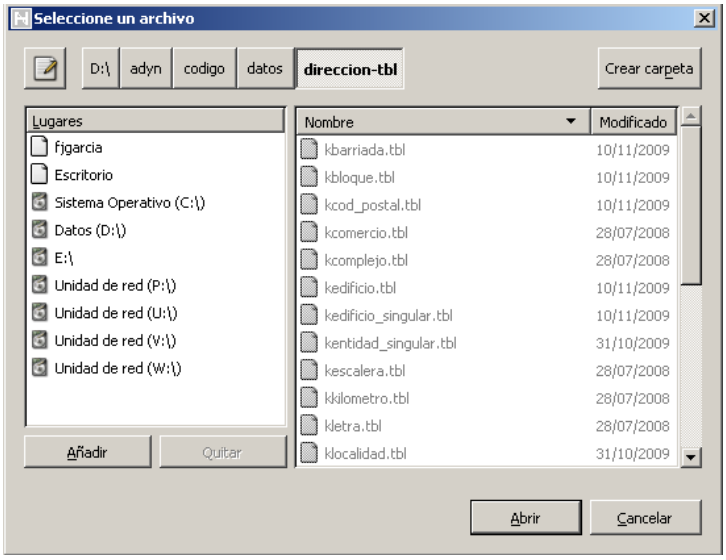

Imagen 31: Selección de la tabla de búsqueda 'kvia.tbl'.

Hemos de indicar que todos los archivos con extensión *'.tbl'* son las diferentes tablas de búsqueda que utilizaremos en el proceso de etiquetado de los campos. Las etiquetas y tablas de búsqueda asociadas se muestran en el **Anexo III.**

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 'informacion_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'.
- 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:

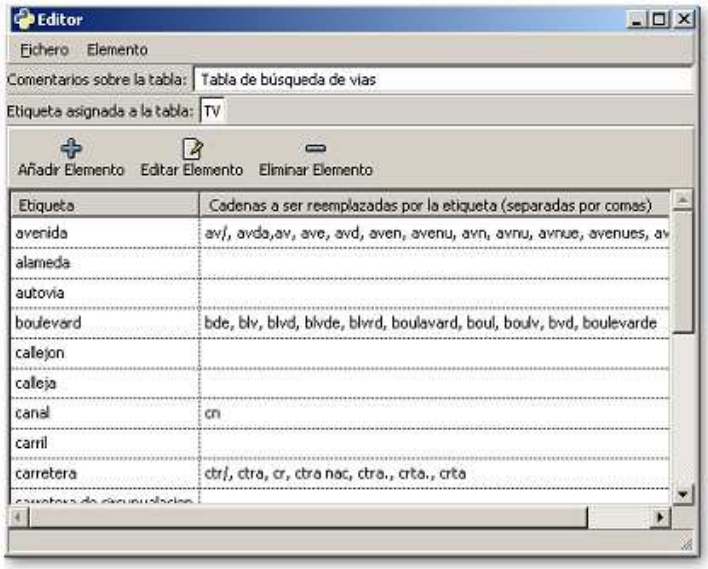

Imagen 32: Elementos de la tabla de búsqueda 'kvia.tbl'.

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

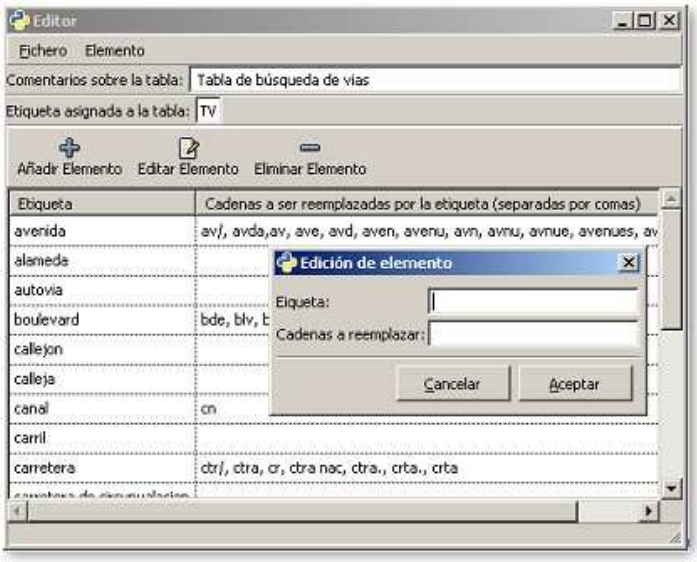

Imagen 33: Añadiendo un nuevo elemento a la tabla de búsqueda 'kvia.tbl'.

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.

Anexo III: Etiquetas y estados definidos para direcciones postales.

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 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

Estado	Descripción del Estado
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.
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

Anexo IV: Recomendaciones de uso.

a) Nombres de los campos 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 los nombres de los campos del fichero de datos no tengan tildes para evitar errores de lectura del fichero que contiene el campo a normalizar.

b) Ficheros de prueba

Para agilizar el trabajo del usuario, *ADYN Herramienta de Normalización* pone a su disposición un conjunto de ficheros de muestras etiquetadas junto con sus correspondientes Modelos Ocultos de Markov, de tal forma que el usuario pueda comenzar a trabajar sin tener que crear previamente dichos modelos de normalización. En aquellos casos en que encuentre alguna estructura o patrón no presente en los ficheros de muestras etiquetadas podrá introducirla de la forma comentada en este manual (Ver *Parte 3: Validación del proceso de normalización*) de manera que los modelos se irán enriqueciendo progresivamente.

La ubicación de estos ficheros es la siguiente:

C:\adyn\ejemplos

c) 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***' (parte 3 del proceso) 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 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. En la *Guía General de Aprendizaje de ADYN Herramienta de Normalización* se puede consultar más ampliamente cómo funciona este proceso.

d) 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.