XVII JEECAS Jornadas de Estadística de las Comunidades Autónomas
Cáceres del 20 al 23 de Octubre de 2010

ADYN Herramienta de Normalización

Caballero Ruíz, Elisa Isabel
elisa.caballero.ext@juntadeandalucia.es

García González, Francisco José
franciscoj.garcia.g.ext@juntadeandalucia.es

Instituto de Estadística de Andalucía

Resumen:

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 con formatos consistentes y bien definidos y resolver las posibles inconsistencias sobre la forma en la que se representa y codifica la información. El conjunto de estas técnicas es lo que se conoce como proceso de normalización de datos.

Para realizar de forma automática y ordenada el proceso de normalización de ficheros de datos, el Instituto de Estadística de Andalucía ha creado ADYN Herramienta de Normalización, que tiene como funcionalidades la limpieza, estandarización y segmentación de nombres de personas, de direcciones postales y de identificadores de personas físicas y jurídicas.

La singularidad de ADYN Herramienta de Normalización frente a otras herramientas y métodos de normalización será el uso de técnicas de aprendizaje supervisado; lo que significa que el sistema irá 'aprendiendo' a través del conjunto de datos que el usuario le suministra. Mediante este sistema se evita tener conocimientos previos de programación ya que no necesitamos modificar el código de la aplicación para incluir nuevas reglas de normalización, siendo esto muy costoso desde el punto de vista computacional.

En este documento se realizará de forma detallada un proceso de normalización a través de un fichero de datos llamado Ejemplo.csv que contiene un único campo con datos de direcciones postales y en el que intentaremos eliminar y corregir los elementos que consideremos erróneos así como segmentar las direcciones postales en los elementos que la componen, es decir, tipo de vía, nombre de vía, portal, planta, etc.

Por último indicar que ADYN Herramienta de Normalización es de licencia libre y accesible a todos los usuarios de las Administraciones Públicas, estando disponible a través del Repositorio de Software Libre de la Junta de Andalucía.

Palabras claves: *ADYN Herramienta de Normalización, normalización, limpieza, estandarización, direcciones postales, nombres de personas e identificadores de personas físicas o jurídicas.*

1. Introducción al proceso de normalización de un fichero de datos.

Normalmente, la gran mayoría de la información con la que nos encontramos en el mundo real contiene errores, está incompleta o incorrectamente formateada. Para resolver este problema hemos fijado un conjunto de técnicas encaminadas a la obtención de datos consistentes que se engloban en el llamado *proceso de normalización de datos* y redundará en una mejor calidad y fiabilidad en posteriores análisis de esos datos.

Para realizar 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 identificadores de personas físicas y jurídicas.

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

- Las listas de corrección: estos ficheros 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 caracteres extraños ('|', '$',...) y se sustituyen las vocales con tildes por vocales sin tildes.

- Las tablas de búsqueda: son ficheros que contienen un listado de valores que permiten sustituir 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.

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

Imagen 1: Proceso general de normalización.

Caballero Ruíz, Elisa Isabel
García González, Francisco José

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 2: 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 3: Esquema general del proceso de normalización.

Caballero Ruíz, Elisa Isabel
García González, Francisco José

2. ADYN Herramienta de Normalización.

Esta aplicación informática permite 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. Concretamente, a través de este documento realizaremos la normalización del campo 'Direcciones' del fichero de datos 'Ejemplo.csv'. Visualmente queremos obtener lo siguiente:

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

est_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 4: Fichero antes y después de la normalización.

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

- **2.1 Creación del Modelo Oculto de Markov.**
- **2.2 Normalización del fichero de datos.**
- **2.3 Validación del proceso de normalización.**

Veámoslas detenidamente.

2.1 Creación del Modelo Oculto de Markov.

En esta fase del proceso de normalización el objetivo es reconocer la estructura de los datos que vamos a normalizar. Es por ello por lo que utilizamos los Modelos Ocultos de Markov que reconocen ciertos patrones de comportamiento en 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 corregimos los nombres de las vías 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.

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

2.1.1 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. Para el caso de identificadores de personas físicas o jurídicas no es necesario realizar dicho proceso ya que junto con la aplicación se proporciona un Modelo Oculto de Markov que recoge toda la casuística que podemos encontrar sobre los patrones de esos identificadores.

Una vez seleccionado el campo ‘nombre de persona’ o ‘dirección postal’, el usuario será quien especifique el tamaño de la muestra a seleccionar. Automáticamente la aplicación genera esa muestra eliminando los símbolos o caracteres extraños mediante las ‘listas de corrección’ y etiquetando 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 la almohadilla es un simple comentario y sirve de información adicional al usuario pero no es utilizada en ningún proceso.

  A continuación se muestra un número que indica la línea del fichero original en la que se encuentra el registro. En el ejemplo: 565, 594 y 640.

  Seguidamente se muestra un valor numérico entre paréntesis, el cual indica el número que tiene asignado el registro en la muestra. Dicho valor se asigna automáticamente comenzando por 0.

  Por último aparece 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 **la documentación que se suministra con ADYN Herramienta de Normalización** se pueden consultar todas las etiquetas definidas para nombres de personas, direcciones postales, e identificadores de personas físicas y jurídicas.

Desde el punto de vista práctico el proceso de selección y etiquetado de direcciones postales o de nombres de personas utilizando la herramienta *ADYN* se realiza a través de la interfaz ***'02. Selección de la muestra'***. Accedemos a ella a través del menú Inicio (si estamos trabajando con el sistema operativo de Windows) y una vez abierta nos recibirá la siguiente pantalla:

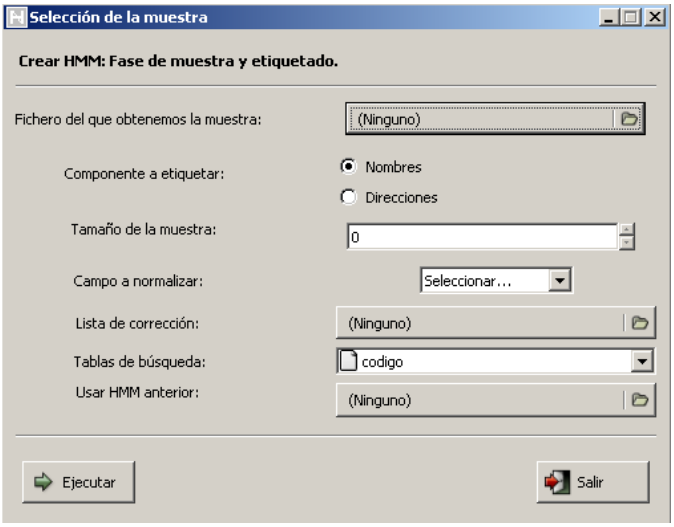


Imagen 5: Interfaz de selección de la muestra.

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*’.

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 el cuadro combinado **‘Campo a normalizar’** 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:

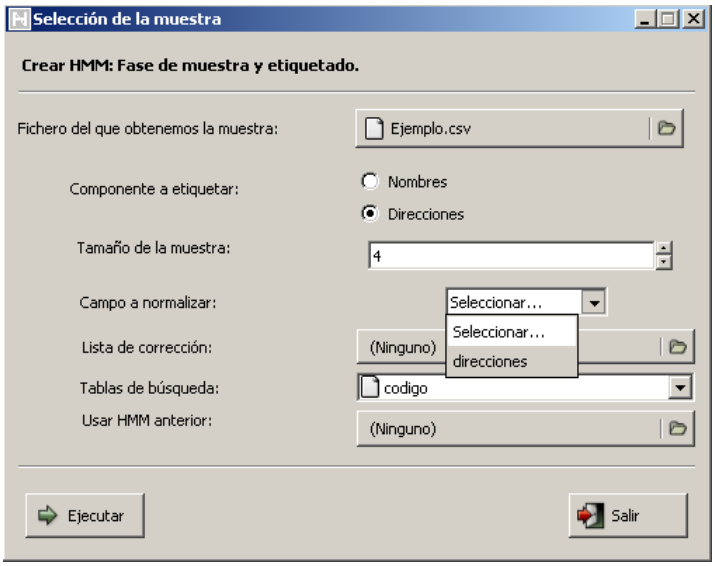


Imagen 6: Selección del campo a normalizar.

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' que se ha creado durante el proceso de instalación.

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.

En el **Anexo V de la documentación de ADYN Herramienta de Normalización** 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, con lo cual solamente nos es útil cuando ya se haya comenzado a trabajar con *ADYN*. En el **Anexo VI de la documentación de ADYN Herramienta de Normalización** 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'***:

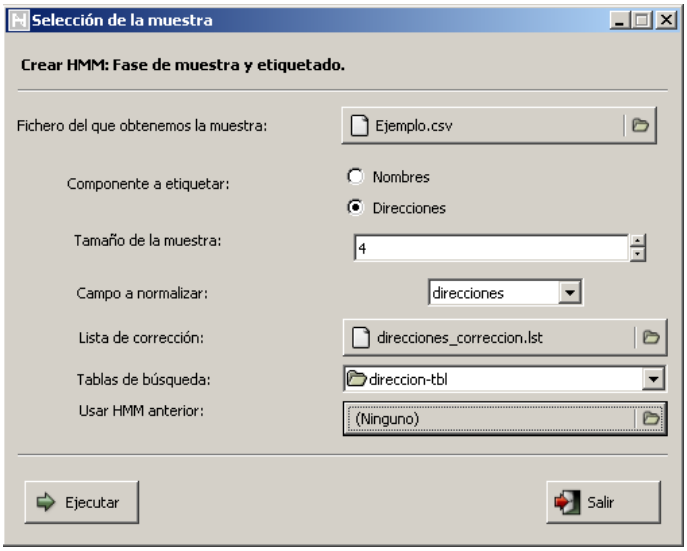


Imagen 7: Interfaz de selección de la muestra.

Hacemos click sobre el botón ‘**Ejecutar**’ y, cuando el proceso termine, nos aparecerá una pantalla confirmando que el proceso ha finalizado.

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*

El contenido del fichero será el siguiente:

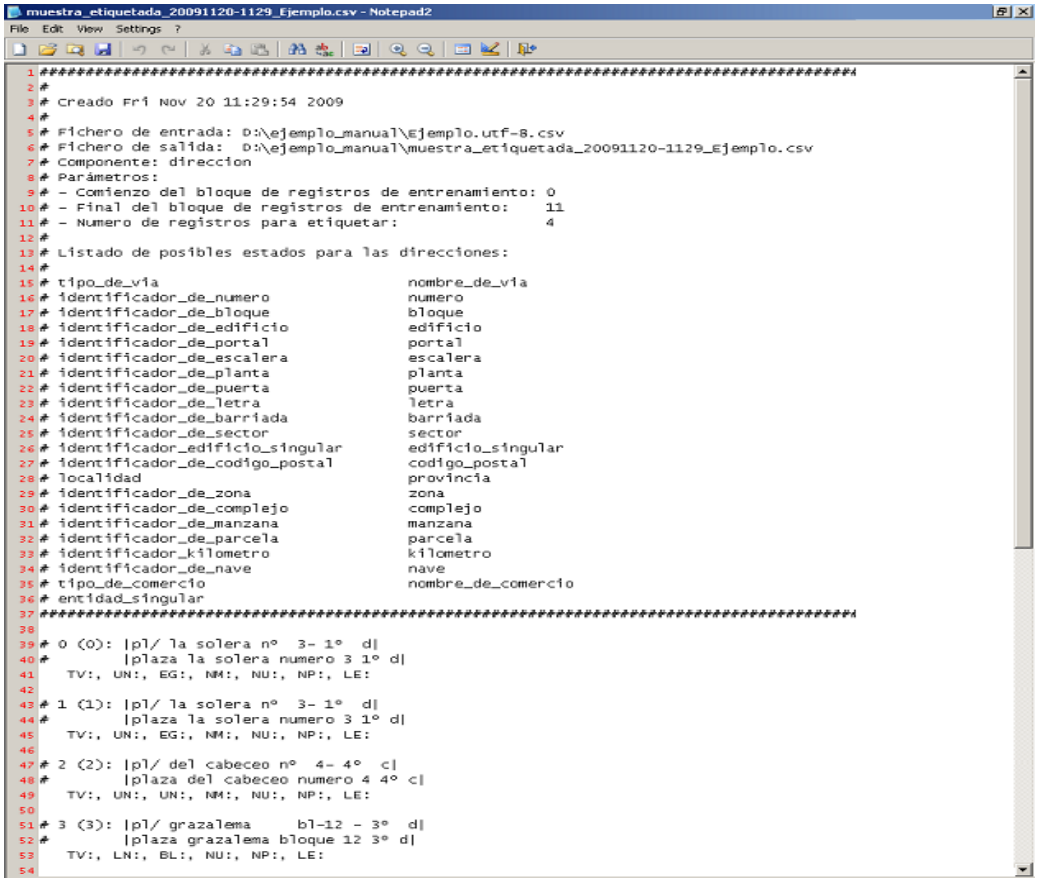


Imagen 8: Fichero resultante de la interfaz ‘Selección de la muestra’.

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 entrada: 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, ‘direccion’.

- 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. La definición de cada una de estas etiquetas y estados referidos a Direcciones se puede consultar en el **Anexo IV de la documentación de ADYN Herramienta de Normalización**.

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:
```

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

2.1.2 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 y la aparición de posibles errores en la aplicación ADYN por este motivo. 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:

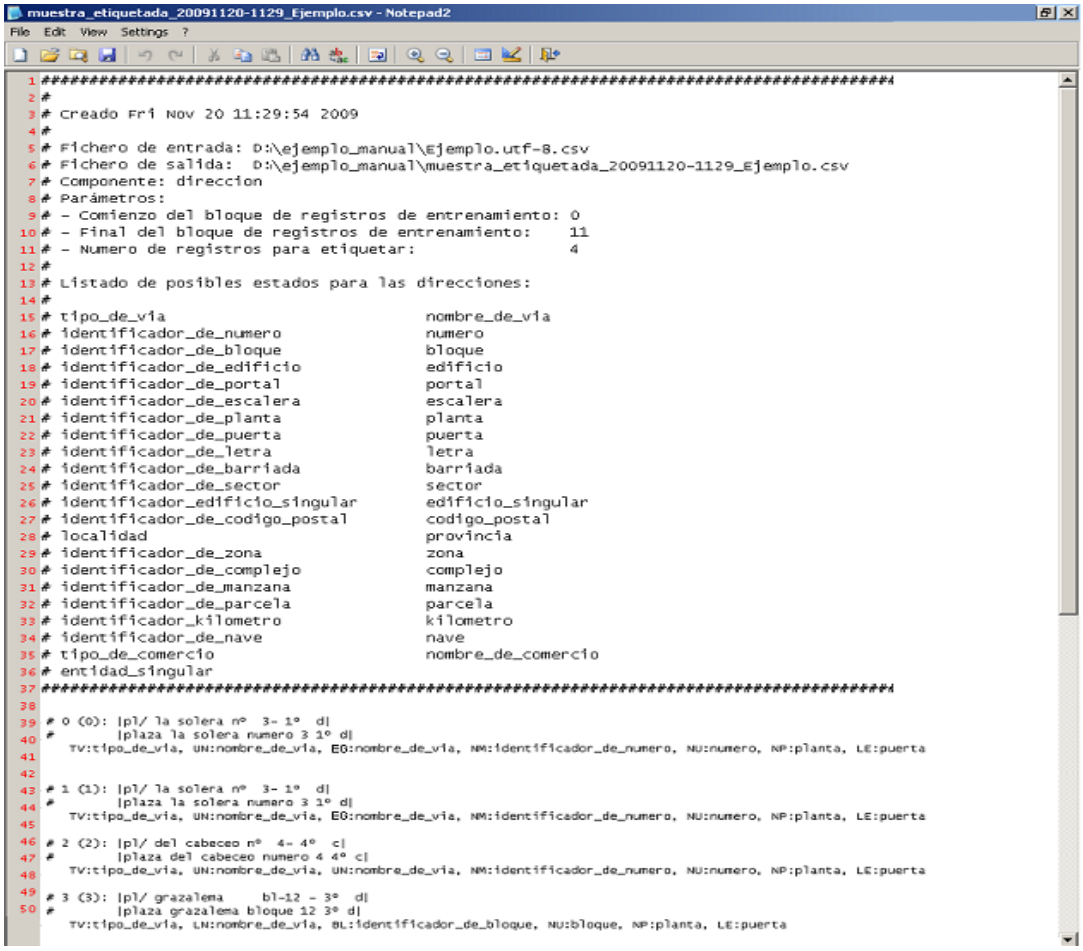


Imagen 9: Fichero con las componentes de la muestra etiquetadas y con estados asignados.

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 las coincidentes. 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.

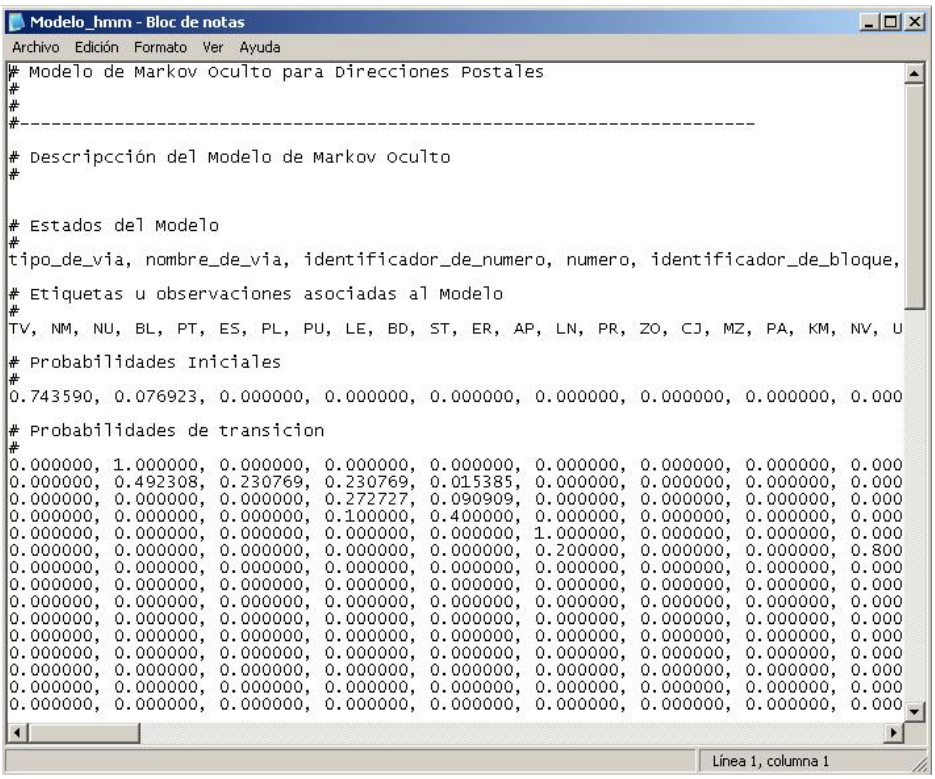
2.1.3 Paso 3: Entrenamiento de la muestra.

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_hmm - Bloc de notas

Archivo Edición Formato Ver Ayuda

```
# 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
```

Línea 1, columna 1

Imagen 10: 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 11: 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'.

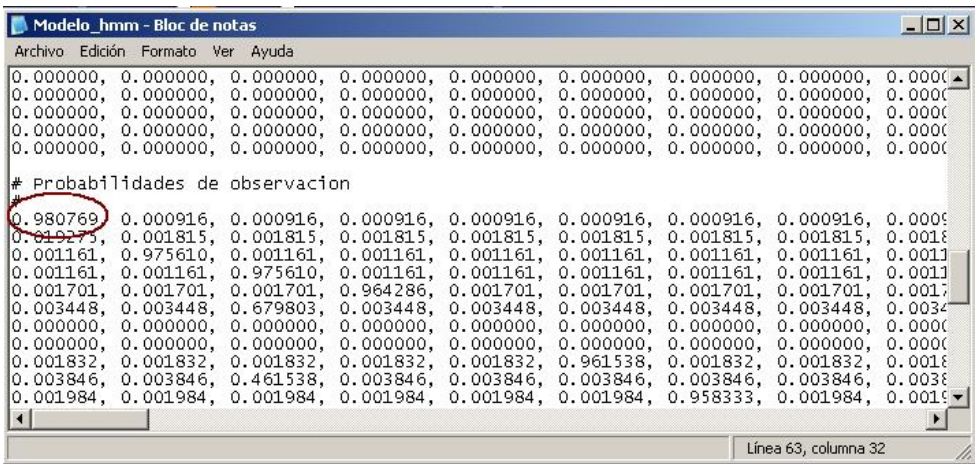
```
0.743590, 0.076923, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000

# 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
Línea 1, columna 1
```

Imagen 12: 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.


Modelo_hmm - Bloc de notas
Archivo Edición Formato Ver Ayuda
0.000000, 0.000000, 0.000000,

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 el uso 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** que se suministra con la documentación de *ADYN Herramienta de Normalización.*

A continuación vemos en la práctica 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) y al abrir '***03.Entrenamiento de la muestra***' nos recibirá la siguiente pantalla:

Imagen 14: 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.

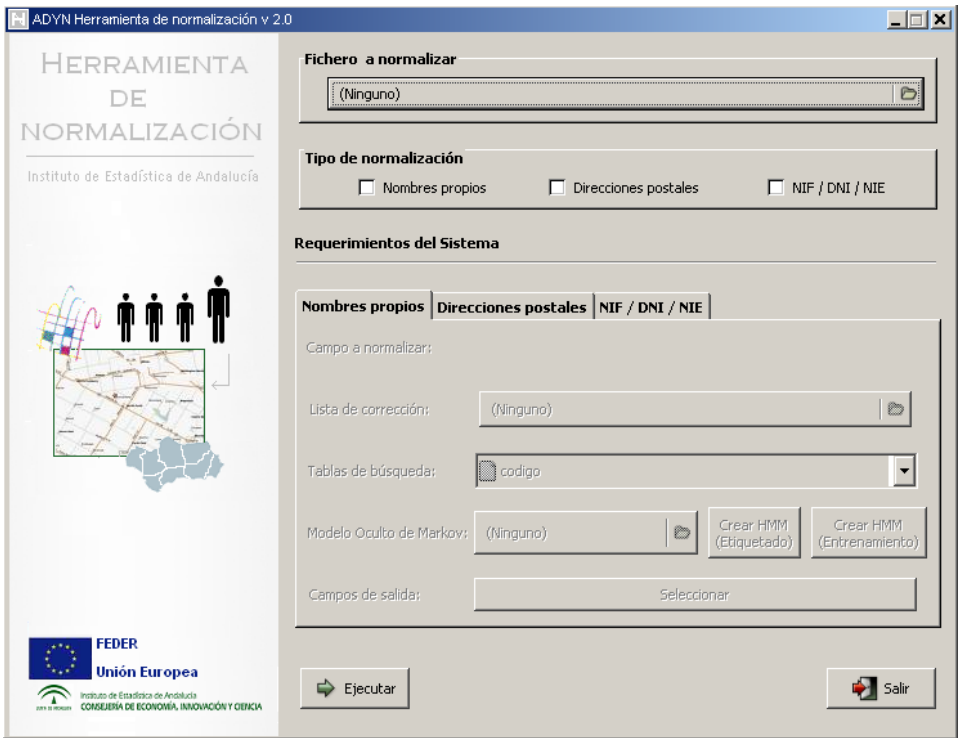
Finalmente haremos click en **'Ejecutar'** y esperaremos a que el programa nos comunique que ha terminado.

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

2.2 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) y el resultado es la siguiente



pantalla:

Imagen 15: 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*'.

Seguidamente en **'Tipo de normalización'** hemos de marcar el tipo de 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'*.

- **'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.

- **'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:

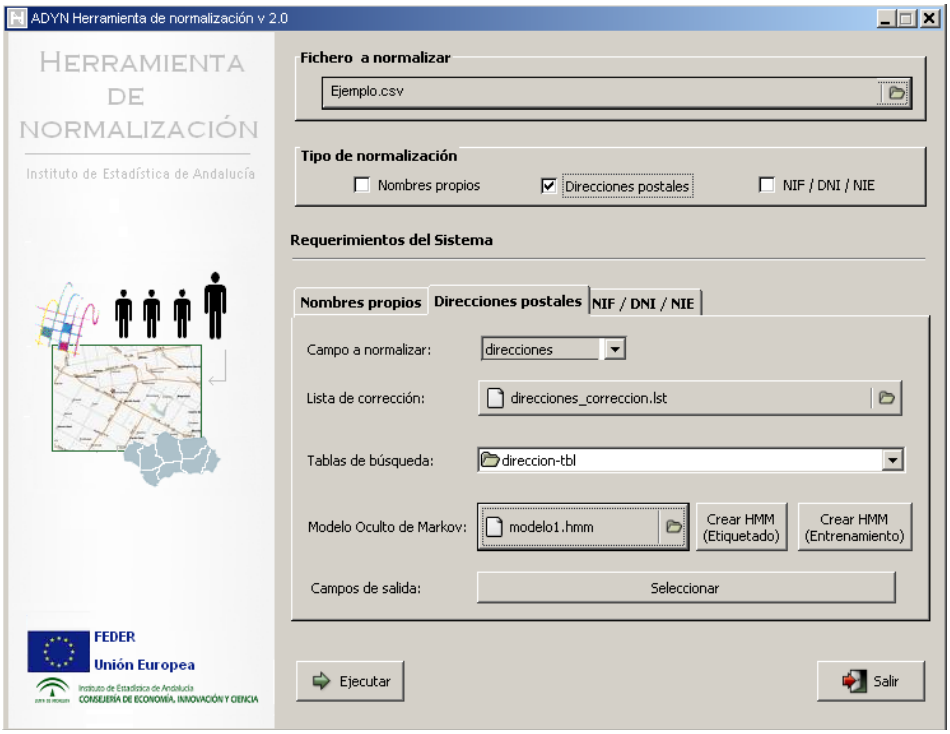


Imagen 16: Interfaz de normalización tras seleccionar la lista de corrección y las tablas de búsqueda.

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 según la forma que se ha explicado con anterioridad.

- 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 17: 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'.

Tras especificar 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 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.
- Fichero ***'proy_<fecha_creación>-<hora_creación>_<fichero_datos>.py'***: es un fichero informativo y contendrá el conjunto de parámetros con los que hemos realizado el anterior proceso de normalización, permitiendo reproducir o modificar este proceso posteriormente.

2.3 Validación del proceso de normalización.

Una vez que tengamos el fichero de datos normalizado es necesario disponer de un sistema que permita validar la bondad del proceso. Es por este motivo por el que se ha definido un algoritmo de validación cuyo funcionamiento se muestra a continuación.

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 18: 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 no se han normalizado correctamente debido a que hay registros cuyas estructuras **NO** se han introducido en la muestra con la que se generó el Modelo Oculto de Markov utilizado para normalizar el fichero original de datos. Además permite 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' que hemos obtenido anteriormente.
>
> 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) Lo más intuitivo es introducir en el fichero de la 'muestra_etiquetada' aquellas estructuras que no hayan sido bien normalizadas. Por ejemplo, si no se ha normalizado '*espiritu santo 2 bajo*' introduciremos la siguiente secuencia:

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

Seguidamente volveríamos a calcular el Modelo Oculto de Markov y normalizaremos el fichero de datos.

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. 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.

3. Conclusiones.

Por último hemos de fijar unas conclusiones respecto a la utilización de ADYN Herramienta de Normalización:

- Disponemos de una aplicación informática que permite normalizar automáticamente direcciones postales, nombres de personas e identificadores de personas físicas y

jurídicas presentes en los registros administrativos, pudiéndose a partir de ese momento tratar estadísticamente.

- El usuario no necesita tener un conocimiento previo de programación, por lo tanto resulta una herramienta intuitiva y fácil de manejar.

- Aplicación basada en una **metodología de aprendizaje supervisado** en la que el usuario podrá ir enriqueciendo el conjunto de estructuras que permitan una optimización del proceso de normalización.

- Es de licencia libre y accesible a todos los usuarios de las Administraciones Públicas, está disponible a través del Repositorio de Software Libre de la Junta de Andalucía y próximamente también a través de la página web del Instituto de Estadística de Andalucía.

- **Organismos que actualmente usan ADYN**: Consejerías de la Junta de Andalucía (C. de Obras Públicas y Vivienda, C. Agricultura y Pesca, C. Cultura, C. Igualdad y Bienestar Social, C. Salud, C. de Economía, Innovación y Ciencia), Diputaciones provinciales (Diputación de Jaén, Diputación de Málaga), Ayuntamiento de Córdoba, Escuela Andaluza de Salud pública e Instituto de Estadística de Andalucía.