Actividad 1: Etiquetado morfosintáctico

Objetivos

Esta actividad consiste en implementar y aplicar un método basado en modelos ocultos de Markov (HMM) para realizar el etiquetado morfosintáctico de una oración.

Pautas de elaboración

En esta actividad debes implementar en Python un etiquetador morfosintáctico basado en modelos ocultos de Markov (HMM) con los componentes necesarios. Se implementará, así mismo, el algoritmo de Viterbi.

Posteriormente se usará todo lo implementado para realizar el etiquetado morfosintáctico de la oración:

«Habla con el enfermo grave de trasplantes».

La actividad consta de tres partes que se describen a continuación.

Parte 1. Construcción las tablas de probabilidades para el etiquetador morfosintáctico

En esta primera parte de la actividad se tiene que implementar en Python el etiquetador morfosintáctico basado en un HMM bigrama a partir de un corpus etiquetado.

Para ello se debe usar el corpus **Corpus-tagged**, que se encuentra disponible como material adicional de la actividad.

El corpus se compone de frases en español etiquetadas con conocimiento sobre las partes de la oración (categorías gramaticales o POS *tags*). Estas frases etiquetadas han sido extraídas de algunos documentos que forman parte de Wikicorpus, un corpus trilingüe (español, catalán e inglés) compuesto por más de 750 millones de palabras. Wikicorpus fue creado por investigadores de la Universitat Politècnica de Catalunya a partir de documentos de la Wikipedia que fueron anotados con la librería *opensource* FreeLing.

La tabla 1 muestra en formato de texto plano y sin etiquetar algunos ejemplos de frases que componen el corpus. De hecho, también se indica el identificador del documento del cual han sido extraídas las frases etiquetadas.

|  |  |
| --- | --- |
| **doc id** | **texto** |
| **27315** | «(…) Tristana es una película del director español nacionalizado mexicano Luis Buñuel. Está basada en la novela del mismo nombre de Benito Pérez Galdós. Fue nominada al Oscar a la mejor película de habla no inglesa en 1970. (…)». |
| **216784** | «(…) En su primer viaje el comportamiento desadaptado y agresivo de Cole lleva a que sea apresado y recluido en un Centro Psiquiátrico, acusado de ser enfermo mental puesto que defiende venir del futuro y habla sobre un virus mortal del que nadie tiene sospechas. En este Centro Psiquiátrico conoce a la psiquiatra Kathryn Railly (Madeleine Stowe) y a un excepcional enfermo mental, Jeffrey Goines (Brad Pitt), con quienes entabla una particular relación que le permite establecer que un grupo radical probablemente ecoterrorista llamado Doce Monos podría ser responsable de la propagación del mortal virus. (…)». |

Tabla 1. Ejemplos de frases que componen el corpus. Fuente: <https://www.cs.upc.edu/~nlp/wikicorpus/>

La versión anotada la conforma el corpus anotado proporcionado para realizar esta actividad. El formato del archivo de texto que contiene el corpus es el mismo que el utilizado en Wikicorpus. Por lo tanto, cada uno de los documentos se identifica con el *tag* XML <doc> donde se indica el identificador del documento (id).

Además, cada una de las frases en el documento viene separada por una línea en blanco. La información relativa a cada palabra de la frase se representa en una nueva línea del archivo. Para cada palabra, es decir, en cada línea del archivo, se proporciona —además del *token* que representa a la propia palabra— su lema, la etiqueta gramatical (POS *tag*) asociada a la palabra y el sentido de esta.

La figura 1 muestra una captura del corpus anotado, donde se observa la frase «Tristana es una película del director español nacionalizado mexicano Luis Buñuel.» perteneciente al documento de Wikicorpus con identificador 27315 y titulado Tristana.

Si se analizan las anotaciones para la palabra «es», se observa que su lema es «ser», que la categoría gramatical a la que pertenece esa palabra es la identificada por la etiqueta gramatical «VSIP3S0» y que el sentido de la palabra es el identificado por el código «01775973175».

También se observa que la palabra «del» en la frase se representa en dos líneas y se anota con dos *tokens*, el primero «de» y el segundo «el». Esto se debe a que la palabra «del» es la contracción de la preposición «de» y el artículo «el». Por el contrario, el nombre propio «Luis Buñuel», que está formado por dos palabras (el nombre «Luis» y el apellido «Buñuel»), se anota como un único *token* «luis\_buñuel». Además, se observa que el punto final de la frase también viene anotado como un token «.».

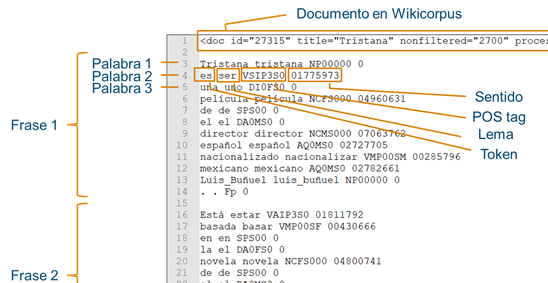


Figura 1. Muestra de parte del corpus etiquetado.

Aunque el corpus anotado proporciona más información (ver figura 1), es importante tener en cuenta que para realizar esta actividad solo será necesario el *token* y la etiqueta gramatical (POS *tag*) de cada palabra; es decir, la información contenida en la primera y la tercera cadena de cada línea que representa una palabra en el corpus anotado.

Las etiquetas gramaticales (POS *tags*) utilizadas para anotar la información morfosintáctica del corpus son las definidas en FreeLing y se basan en EAGLES, una recomendación para la anotación de la mayoría de las lenguas europeas. La definición del conjunto de etiquetas gramaticales (POS tags) utilizadas por FreeLing en el etiquetado de un corpus en español se puede consultar en la web.

Accede al recurso a través del aula virtual o desde la siguiente dirección web:

<https://freeling-user-manual.readthedocs.io/en/v4.1/tagsets/tagset-es/>

Las etiquetas gramaticales de EAGLES utilizadas por FreeLing son de longitud variable, donde cada carácter corresponde a una característica morfosintáctica. El primer carácter en la etiqueta es siempre la categoría gramatical o parte de la oración. Esa categoría gramatical determina la longitud de la etiqueta y la interpretación de cada uno del resto de caracteres en la misma.

La definición de la etiqueta para la categoría gramatical «verbo» se muestra en la tabla 2. Entonces, la etiqueta «VSIP3S0», con la que ha sido etiquetada la palabra «es» en la frase que se presentó anteriormente, se interpreta de la siguiente forma: se refiere a un verbo (V) de tipo semiauxiliar (S) en modo indicativo (I) y en tiempo presente (P) para la tercera persona (3) de (número) singular (S). Asimismo, el carácter «0» al final de la etiqueta indica que esta forma verbal no tiene género.

|  |  |  |
| --- | --- | --- |
| **Posición** | **Atributo** | **Valores** |
| 0 | categoría | **V**: verbo. |
| 1 | tipo | **M**: principal; **A**: auxiliar; **S**: semiauxiliar. |
| 2 | modo | **I**: indicativo; **S**: subjuntivo; **M**: imperativo; **P**: participio; **G**: gerundio; **N**: infinitivo. |
| 3 | tiempo | **P**: presente; **I**: imperfecto; **F**: futuro; **S**: pasado; **C**: condicional. |
| 4 | persona | **1**:1; **2**:2; **3**:3. |
| 5 | número | **S**: singular; **P**: plural. |
| 6 | género | **F**: femenino; **M**: masculino; **C**: común. |

Tabla 2. Definición de la etiqueta para la categoría gramatical «verbo».   
Fuente. <https://freeling-user-manual.readthedocs.io/en/v4.1/tagsets/tagset-es/>

Es importante destacar que para realizar la actividad se deben utilizar las etiquetas con las que se anota el corpus en formato EAGLES; por ejemplo, «VSIP3S0».

**Importante:** si se utilizan otras etiquetas, la actividad será considerada incorrecta y puntuada con cero puntos.

Para construir el etiquetador morfosintáctico a partir del corpus etiquetado con los datos de entrenamiento, se deberán seguir los siguientes pasos:

Carga del corpus para extraer la primera y tercera columna de cada registro.

Cálculo de las probabilidades que rigen el HMM bigrama, es decir:

Cálculo de las probabilidades de emisión del HMM a partir del corpus etiquetado. Construcción de la tabla de probabilidades de emisión.

Cálculo de las probabilidades de transición del HMM a partir del corpus etiquetado. Construcción de la tabla de probabilidades de transmisión.

**Nota:** se debe adjuntar en el envío de la actividad la tabla (guardada en formato de hoja de cálculo de Microsoft Excel (.xlsx) o equivalente) con las probabilidades de emisión y las de transición, calculadas para todas las etiquetas y *tokens* (palabras) que aparecen en el corpus.

Parte 2. Etiquetado morfosintáctico de una oración

En esta segunda parte de la actividad se debe implementar en Python un programa que permita calcular la mejor secuencia de etiquetas para una oración; dicho de otro modo, realizar el etiquetado morfosintáctico de la oración: «Habla con el enfermo grave de trasplantes.».

Para ello se debe utilizar el etiquetador que se ha construido en la parte 1 de esta actividad, es decir, las tablas de probabilidades calculadas. Tomando como base la información de la parte 1, se debe aplicar el algoritmo de Viterbi.

Para aplicar el algoritmo de Viterbi, se deben seguir los siguientes pasos:

Cálculo de la matriz de probabilidades de la ruta de Viterbi (matriz con los valores de Viterbi) donde se representen claramente las observaciones y los estados de la máquina de estados finitos. Se debe calcular el valor de Viterbi para cada celda de la matriz y devolver claramente los valores obtenidos.

**Nota:** para simplificar, se pueden eliminar de la matriz de Viterbi todos aquellos estados asociados a etiquetas que no aparezcan en el posible análisis de la oración, de forma que solo se tratan los estados relevantes. Además, se debe tener en cuenta la transición al estado final representado por el punto al final de la oración que se va a analizar.

Obtención de la ruta con máxima probabilidad, es decir, se debe trazar la ruta inversa para obtener la mejor secuencia de etiquetas.

Presentación de la oración etiquetada como resultado de todos los procesos llevados a cabo.

**Nota:** se debe adjuntar con el envío del informe de la actividad la tabla (guardada en formato de hoja de cálculo de Microsoft Excel (.xlsx) o equivalente) con la matriz de probabilidades de la ruta Viterbi para el etiquetado morfosintáctico de la oración «Habla con el enfermo grave de trasplantes.».

Parte 3. Análisis del etiquetador morfosintáctico

Una vez se haya creado el etiquetador morfosintáctico y se haya usado para etiquetar la oración «Habla con el enfermo grave de trasplantes.», se debe reflexionar sobre los resultados obtenidos, interpretándolos y analizando el rendimiento del etiquetador creado y sus limitaciones. Para ello, se debe responder de forma razonada a las siguientes preguntas:

¿Es correcto el etiquetado morfosintáctico que se ha obtenido? Se debe justificar la respuesta.

Indicar el resultado de etiquetar esta otra oración «El enfermo grave habla de trasplantes.» utilizando el etiquetador morfosintáctico. ¿Es correcto el etiquetado morfosintáctico que se ha obtenido? Se debe justificar la respuesta.

¿Cuáles son las limitaciones del analizador morfosintáctico creado?

¿Qué posibles mejoras se podrían aplicar para mejorar el rendimiento del etiquetador morfosintáctico creado?

**Extensión y formato**

**Extensión** máxima: no hay restricciones en la extensión.

**Rúbrica y criterios de evaluación**

|  |  |  |  |
| --- | --- | --- | --- |
| Etiquetado morfosintáctico | Descripción | Puntuación máxima sobre 10  (puntos) | Peso  % |
| Probabilidades de transición | La tabla de probabilidades de transición que rige el etiquetador morfosintáctico es correcta. | 2 | 20 % |
| Probabilidades de emisión | La tabla de probabilidades de emisión (observación) que rige el etiquetador morfosintáctico es correcta. | 2 | 20 % |
| Valores de Viterbi | La matriz de probabilidades donde se presentan los valores de Viterbi para el análisis de la oración es correcta. | 3 | 30 % |
| Etiquetado morfosintáctico | La secuencia de etiquetas obtenida, asociada a la oración propuesta al utilizar el etiquetador morfosintáctico es correcta. | 1 | 10 % |
| Análisis del etiquetador morfosintáctico | Las respuestas a las preguntas sobre el análisis de los resultados obtenidos con el etiquetador morfosintáctico, sus limitaciones y las posibles mejoras, son correctas. | 2 | 20 % |
|  |  | **10** | 1. **%** |

Debe resolverse la actividad en el Jupyter Notebook proporcionado junto con este enunciado. Si no se usa el archivo proporcionado la actividad será calificada con cero puntos.

Para la evaluación de la actividad se debe entregar el Jupyter Notebook que contenga:

El código en Python que permita resolver la actividad, debidamente comentado.

Todos los resultados que se solicita en el enunciado y que genera el código implementado.

Las respuestas a las preguntas planteadas en el enunciado.

También se debe entregar como anexos al informe los archivos de hoja de cálculo de Microsoft Excel (.xlsx) o equivalente donde se muestren las tablas de las probabilidades de emisión y las de transición y la matriz de probabilidades de la ruta Viterbi.

Se valorarán positivamente los comentarios del código claros y oportunos que permitan entender las decisiones de implementación realizadas. Si el código no está comentado de forma suficiente, se penalizará en la calificación de la actividad.

Se valorará positivamente la respuesta clara, breve y bien argumentada a las preguntas.

Si se presenta el código, pero no se muestran los resultados que este genera, no se considerarán esos resultados en la calificación de la actividad.

No se pueden utilizan recursos disponibles para el procesamiento del lenguaje natural como, por ejemplo, bibliotecas, API… Si se utiliza alguno de estos recursos para resolver la actividad y no se implementan directamente todos los pasos descritos en el enunciado, la actividad se calificará con cero puntos.

Si se detecta plagio de cualquier fuente (Internet…), ya sea en el código, en los comentarios o en las respuestas, el alumno obtendrá una calificación para la actividad de cero puntos.

Si se detecta copia entre alumnos en el código, en los comentarios o en las respuestas, todos los alumnos involucrados obtendrán una calificación para la actividad de cero puntos.