Módulo

Text Mining

Autor: Luis Gascó Sánchez

Actualizado Marzo 2021

Ejercicios prácticos (Tiempo estimado de realización: ~6-8 horas)

Se propone a los alumnos la realización de tres ejercicios prácticos que cubren las tres áreas principales expuestas en la formación teórica.

La entrega de los ejercicios se realizará enviando al profesor los archivos *.ipynb" generados en Google Collab o bien los links de acceso a los notebook (comprobar que las propiedades de compartición son las correctas). Cada ejercicio se debe entregar en un *notebook* independiente, que debe estar correctamente comentado para la obtención de la calificación máxima.

Para cada ejercicio se deberán utilizar los corpus que se descargan en la primera celda de los notebook enlazados al final de cada ejercicio. Recordar descargar los módulos de NLTK necesarios y modelos de Spacy necesarios en cada caso.

Si tenéis alguna duda, por favor no dudéis en preguntarla en el foro o en la plataforma. Estaré encantado de ayudaros a resolverla.

Ejercicio 1 (35%)

El objetivo de este ejercicio es comprobar los conocimientos adquiridos por el alumno en temas relacionados a las posibilidades de visualización de datos textuales y el proceso de preparación de los textos.

Para ello, teniendo en cuenta el conjunto de datos textuales del corpus, se solicita lo siguiente:

- **Apartado 1**: Información básica sobre el corpus: Número de documentos, número de documentos duplicados, número de elementos en cada clase. (10%)
- **Apartado 2**: Diseñar funciones para la limpieza de los textos:
    - Quitar palabras vacías. (10%)
    - Quitar símbolos de puntuación. (10%)
    - Lematización con Spacy. (10%)
    - Tokenización con NLTK. (10%)
- **Apartado 3**: Calcular y representar gráficamente en forma de distribución, como se vio en clase, las longitudes en caracteres y en tokens (después del proceso de limpieza) de los documentos del corpus. (25%)
- **Apartado 4**: Calcular y representar gráficamente en forma de histograma los 10 tokens más utilizados en cada una de las clases del corpus después del proceso de limpieza (25%)

Nota: Como se dijo en clase, para la obtención de la máxima calificación en los ejercicios es necesario que el código esté apropiadamente comentado indicando el proceso llevado a cabo por el alumno. Además, en el caso de las visualizaciones, deben incorporar todos los elementos necesarios para su correcta comprensión (leyenda, rotación de etiquetas si es necesario, etc)

Link al notebook del ejercicio 1

Ejercicio 2 (35%)

El objetivo de este ejercicio es comprobar los conocimientos adquiridos por el alumno en temas relacionados a la transformación de textos utilizando la técnica de TFIDF.

- **Apartado 1**: Preprocesar los textos del corpus con las funciones de preprocesado creadas en el ejercicio 1 (quitar palabras vacías, quitar símbolos de puntuación, lematizar con spacy) (20%)
- **Apartado 2**: Utilizar la función de scikit-learn para realizar la transformación a TF-IDF del texto considerando que:
    - Se consideren unigramas, bigramas y trigramas (20%)
    - No se tengan en cuenta los elementos que aparezcan en menos del 5% de los textos (30%)
    - Se utilice el tokenizador incorporado en Scikit-Learn. (10%)
    - Haya un máximo de 200 características. (10%)

Nota 1: Como se dijo en clase, recordad los valores por defecto de la función de scikit-learn ya que no es una buena práctica duplicar las tareas de preparación de textos. En este ejercicio se está pidiendo realizar el proceso de limpieza **antes** de introducir la cadena de caracteres a la función de scikit-learn, por lo que esta función no debería re-hacer tareas como quitar stopwords, por ejemplo. (la selección correcta de valores de las variables es el último 10% de la valoración del ejercicio).

Nota 2: Además, recordar una vez más la necesidad de comentar correctamente el ejercicio para obtener la máxima calificación (¡tengo que saber si tenéis claro el proceso que estáis llevando a cabo!)

Link al notebook del ejercicio 2

Ejercicio 3 (30%)

El objetivo de este ejercicio es comprobar los conocimientos adquiridos por el alumno en temas relacionados al entrenamiento de modelos de clasificación con características de análisis de sentimiento.

En esta ocasión teneis que entrenar un modelo de clasificación que además de considerar las características de TFIDF debe tener en cuenta al menos algunas características extras como el sentimiento y objetividad de las frases utilizando librerías externas.

En este ejercicio, muy similar al ejercicio de Twitter Classification se pide que repliquéis la estructura para generar un modelo de clasificación de los datos del dataset.

El corpus se descargó utilizando la API de Twitter, recopilando los datos que mencionaran la palabra "noise". Cada tweet tiene una etiqueta en la columna "Molestia". El valor de esa etiqueta representa lo siguiente:

- ***Valor 1***: Tweets con la palabra ruido que hacen referencia a molestias sufridas por ruido acústico proveniente de distintas fuentes (coches, vecinos, mascotas,..)
- ***Valor 0***: Tweets que contienen la palabra ruido, pero no expresan una molestia sufrida por el usuario que lo escribió. Aquí se incluyen tweets que hacen referencia a noticias sobre ruido, a una valoración del ruido acústico como algo positivo, u otras acepciones de la palabra ruido (mediatic noise).

Dado que el ejercicio es muy similar al compartido, se pide que:

- Repliquéis los pasos en la generación del modelo explicando con vuestras palabras cada uno de los procesos. (30%)
- Incorporéis características de sentimiento del texto al modelo de clasificación (utilizando librerías como Textblob) además de las características TFIDF. En el ejercicio sobre clasificación en Twitter se explica cómo incorporar estas nuevas variables (40%).
- Una vez entrenado el modelo, extraigáis el nombre de las 10 características más importantes (30%).

Nota 1: Cualquier cálculo adicional a los vistos en clase (visualizaciones, nuevas características añadidas al modelo, distribuciones por clase...) será valorado positivamente en la calificación.

Nota 2: Insisto en seguir la estructura de análisis en tres fases vista en clase. Esto os facilitará llevar un orden en el análisis, explicar los pasos de forma estructurada y clara, y conseguir una mejor calificación.

Link al notebook del ejercicio 3