PRÁCTICA OBLIGATORIA EVALUABLE

ANÁLISIS DE OPINIÓN CON NLTK

Objetivo

El objetivo es poner en práctica diferentes conceptos aprendidos durante el curso y relacionados con la Recuperación de Información y el procesamiento de texto con Python, aplicándolos a una tarea real concreta.

Normativa de entrega

La entrega de esta práctica es *obligatoria* y puede realizarse ***individualmente o por parejas***. A continuación, se detallan otros datos de interés relacionados con la normativa de entrega:

- La fecha límite de entrega de la práctica será el día indicado en el campus virtual.
- La entrega de la práctica se hará a través de Aula Virtual, empleando la actividad habilitada para ello
- Se deberá subir un único archivo comprimido con todo el código fuente que se haya implementado y se necesite para ejecutar la práctica, así como una memoria explicando qué se ha hecho y por qué.
- El archivo debe ser nombrado de la siguiente manera: UAXAYZNombre/s. Donde:
  - X es el número de la Unidad
  - Y equivale a "I" si se ha hecho Individual o "C" si se ha hecho Colectiva
  - Z equivale al número de práctica de la unidad.
  - Nombre/s es vuestro nombre/s completo/s.
  - Ejemplo: UA1AI1RafaelMuñoz

Enunciado

De toda la cantidad de información nueva disponible cada día en la Red, una buena parte se corresponde con opiniones vertidas por los usuarios. Los comentarios en Internet son una importante fuente de información, ya que permiten conocer las opiniones sobre determinados lugares o experiencias de cara a tomar nosotros una decisión al respecto. Así, los foros online y los portales colaborativos se han convertido en una de las principales fuentes de información de actualidad. Estos portales pueden ser transversales o verticales, estando especializados en temas concretos. Por ejemplo, en el portal PatientsLikeMe https://www.patientslikeme.com/ los usuarios pueden comentar sobre enfermedades, tratamientos, síntomas y demás asuntos relacionados con la medicina. Otros portales, como tripadvisor https://www.tripadvisor.es/ recopilan información de interés turístico como hoteles, restaurantes o monumentos.

Estos comentarios no suelen ser muy extensos (depende de la temática de la noticia y del contexto actual de esa temática), con una temporalidad muy marcada y una estructura similar (por lo general título, sumario y cuerpo). Estas características influyen de manera directa en la forma de procesarlas, y de ellas se nutren los trabajos que se dedican a investigar diferentes formas de procesar los documentos de noticias para llevar a cabo distintas tareas.

Una de las tareas más populares es el análisis de esos comentarios, el conocido "Sentiment Analysis"

Análisis de Opinión

El Análisis de Opinión, también conocido como Sentiment Analysis, es un campo dentro del Procesamiento de Lenguaje Natural (PNL) que construye sistemas que intentan identificar y extraer opiniones dentro del texto. Generalmente, además de identificar la opinión, estos sistemas extraen atributos de la expresión, por ejemplo:

- Polaridad: si el hablante expresa una opinión positiva o negativa.
- Asunto: de lo que se habla.
- Titular de opinión: la persona o entidad que expresa la opinión.

Actualmente, el análisis de opinión es un tema de gran interés y desarrollo, ya que tiene muchas aplicaciones prácticas. Dado que la información disponible de forma pública y privada en Internet está en constante crecimiento, una gran cantidad de textos que expresan opiniones están disponibles en sitios de revisión, foros, blogs y redes sociales.

Con la ayuda de los sistemas de análisis de opinión, esta información no estructurada podría transformarse automáticamente en datos estructurados de opiniones públicas sobre productos, servicios, marcas, políticas o cualquier tema sobre el que las personas puedan expresar opiniones. Estos datos pueden ser muy útiles para aplicaciones comerciales como análisis de mercadotecnia, relaciones públicas, revisiones de productos, puntaje de promotores netos, retroalimentación de productos y servicio al cliente.

Objetivo

El objetivo de la práctica es la realización de un sistema de análisis de opinión sobre los textos extraídos de la práctica anterior, clasificando el texto entre positivo, negativo o neutro.

Paso 1: Procesar los documentos descargados y convertirlos en texto. Habitualmente serán páginas web en formato HTML, por lo que mediante alguna librería se procesarán y se convertirán en texto. Se recomienda BeautifulSoup para ello.
Paso 2: Realizar un trabajo de preparación del texto. Por ejemplo, eliminar palabras vacías, hacer stemming, lematización, reconocimiento de Entidades Nombradas, etc. El alumno debe hacer una labor de investigación. Se recomiendoa NLTK para ello.
Paso 3: Determinar la polaridad del texto resultante. Para ello no utilizar ningún algoritmo de clustering ya que queda fuera del ámbito de la asignatura y no contamos actualmente con un corpus preparado para entrenar. Utilizar algoritmos basados en reglas ya codificados. Se recomienda TextBlob y Spacy para ello.

Procesar el texto para generar una representación común de cada documento, por ejemplo, una representación vectorial donde cada vector representa un documento. Cada componente del vector representa un rasgo del vocabulario del problema (todos aquellos rasgos únicos de todos los documentos que intervienen en el problema) y tiene un peso concreto, que mide la importancia de ese rasgo en el documento.

Entidades Nombradas

El estilo de redacción de las noticias se corresponde con el género informativo, donde lo primordial es informar. Los elementos que debe reunir una noticia se conocen en el argot periodístico con el nombre de las 6W" (What, Who, When, Where, Why, How): ¿Qué sucedió? ¿A quién le sucedió? ¿Cuándo sucedió? ¿Dónde sucedió? ¿Por qué sucedió? y ¿Cómo sucedió? Estos seis elementos no son necesarios en todas las noticias, algunos de ellos pueden faltar o se pueden unir. El orden en el que aparecerán las respuestas a estas preguntas dependerá del suceso que se relate, del redactor, o incluso de la guía de estilo del medio que publique la noticia.

Las respuestas a las diferentes preguntas que se pueden plantear ante una noticia están presentes en el contenido de la misma y varias de esas respuestas las pueden proporcionar las Entidades Nombradas, como se puede apreciar en la Figura 1, donde hay una tabla con las entidades nombradas encontradas en el texto.

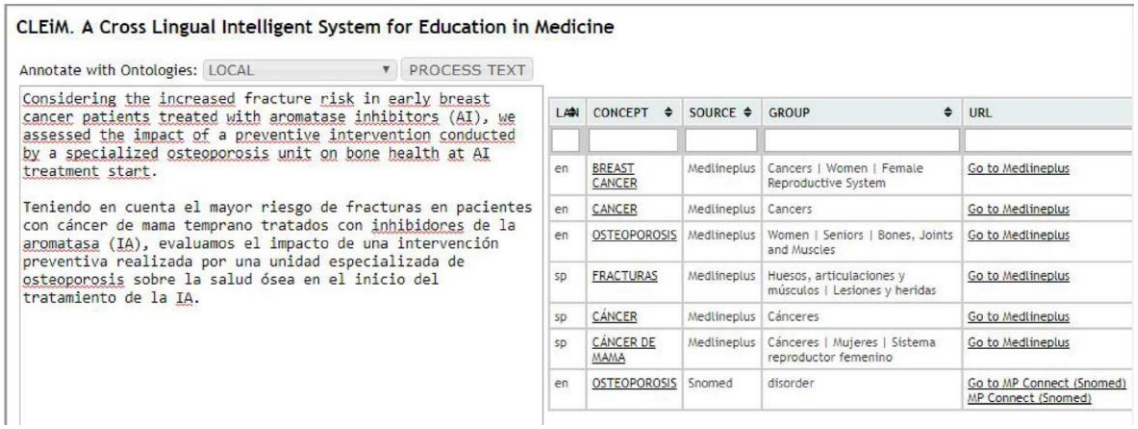

Figura 1

LIBRERÍAS ÚTILES

En la actualidad Python es uno de los lenguajes que más se utiliza en el campo de la Inteligencia Artificial. En particular, para problemas relacionados con Procesamiento del Lenguaje Natural hay diferentes librerías disponibles, donde destacan las siguientes:

- **NLTK** (http://www.nltk.org/). La librería más conocida para PLN en Python.
- TextBlob (http://textblob.readthedocs.io/en/dev/). Librería para procesar textos, basada en NLTK pero más simple e intuitiva en algunas partes.
- **Spacy** (https://spacy.io/). Librería más reciente para procesar textos, sencilla y rápida.
- Textacy (https://pypi.org/project/textacy/).
  Librería de alto nivel desarrollada sobre Spacy.
- Gensim (https://radimrehurek.com/gensim/intro.html).
  Librería diseñada para extraer automáticamente los temas semánticos de los documentos de una manera sencilla y eficiente.

ENLACES ÚTILES

- Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit http://www.nltk.org/book/

- Código

https://github.com/aylliote/senti-py
https://github.com/aylliote/senti-py/blob/master/classifier/sentimentClassifier.py
https://www.pybonacci.org/2015/11/24/como-hacer-analisis-de-sentimiento-en-espanol-2/

- Tutoriales

https://likegeeks.com/es/tutorial-de-nlp-con-python-nltk/
https://pmoracho.github.io/blog/2017/01/04/NLTK-mi-tutorial/
https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/
https://github.com/karanmilan/Automatic-Answer-Evaluation/blob/master/Python%203%20Text%20Processing%20with%20NLTK%203%20Cookbook.pdf

Anexo:
NLTK cuenta con un etiquetador PoS. Desafortunadamente, es únicamente para inglés. Afortunadamente, también tiene acceso a un etiquetador externo (de Stanford) que sí es capaz de manejar el español. Esta herramienta no la tiene por defecto, y al ser externa tampoco se descarga desde el sistema de descarga que hemos estado usando hasta ahora. https://nlp.stanford.edu/software/tagger.shtml