Tarea 1: Fundamentos de Minería de Texto

Dr. Adrián Pastor López Monroy 14/02/2020

Instrucciones

Las siguientes actividades deberán hacerse en un Notebook de Python debidamente comentado. Para ello use el markdown para explicar brevemente lo que hacen los bloques de código claves pedidos en cada actividad.

Actividad 1: Procesamiento básico

1. Descargue una página web del top de noticias de Google News, guárdela en un archivo de texto plano.
2. Investigue el uso de la librería BeautifulSoup de Python para limpiar el formato html de la página.
3. Utilice el objeto RegexpTokenizer de nltk para leerlo, tokenizando con alguna expresión regular para sacar solo secuencias de caracteres de la "a" a la "z". Imprima como resultado los primeros 10 tokens.
4. Cargue la lista de Tokens anteriores en un objeto Text de nltk.
5. Visualice la concordancia de alguna palabra.
6. Imprima con matplotlib las 50 palabras más frecuentes cargadas en un objeto FreqDist de nltk. Antes de este paso remueva las stopword usando un recurso léxico en español de nltk.
7. Cuente las palabras del vocabulario del documento en cuestión.
8. Imprima las palabras con logitud menor a 5 caracteres pero con frecuencia mayor a 10, usando list-comprehension de python.
9. Calcule la riqueza léxica.

1 Actividad 2: Cargando un corpus

1. Cargue el dataset de training de agresividad como un corpus de categorías en nltk. Para referencia, un corpus de categorias es como el Brown corpus visto en clase, y necesita construirse a base del objeto CategorizedPlaintextCorpusReader de nltk.

2. Tokenize cada documento del dataset usando el objeto WordPunctTokenizer de nltk. Imprima la secuencia de tokens de los diez primeros tweets del dataset de agresividad.

3. Tokenize cada documento del dataset usando el objeto TweetTokenizer de nltk. Imprima la secuencia de tokens de los diez primeros tweets del dataset de agresividad.

4. Ordene los tweets por clase, y luego dentro de los tweets agresivos y no agresivos, ordenelos alfabéticamente antes de continuar con el siguiente punto.

5. Para los primeros 50 tweets de la clase positiva, imprima por tweet la longitud promedio en palabras, longitud promedio en caracteres, y la diversidad léxica, i.e. palabras únicas en el tweet entre la longitud del tweet en palabras.

6. Elabore un comentario sobre el punto anterior y si cree que este tipo de información podría ser útil para la detección de agresividad.

7. Utilizando el objeto ConditionalFreqDist de nltk visualize de manera tabular la distribución de las 5 palabras con mayor TFIDF del corpus para cada clase. Elabore un comentario sobre esto.

8. Haga una versión fonética del corpus usando algún diccionario. Ejemplo: https://github.com/Kyubyong/pron_dictionaries

2 Actividad 3: Detección de Agresividad con Análisis de Sentimiento Básico

Use el código que sea necesario de la Práctica 3 para completar esta actividad.

2.1 Experimentos Parte 1

1. Utilice el recurso léxico del Consejo Nacional de Investigación de Canadá llamado "EmoLex" (https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm) para construir una "Bolsa de Emociones" de los Tweets de agresividad (Debe usar EmoLex en Español). Para esto, una estrategia sencilla sería enmascarar cada palabra con su emoción, y después construir la Bolsa de Emociones.

2. Representa a los documentos y clasifica con SVM como en la Practica de Clase 3. Evalúa varias representaciones, y ponga una tabla comparativa a modo de resumen (e.g., binario, frecuencia, tfidf, etc.).

2.2 Experimentos Parte 2

1. Utilice el recurso léxico llamado "Spanish Emotion Lexicon (SEL)" del Dr. Grigori Sidorov, profesor del Centro de Investigación en Computación (CIC) del Instituto Politecnico Nacional (http://www.cic.ipn.mx/∼sidorov/), para enmascarar cada palabra con su emoción,

y después construir la Bolsa de Emociones con algún pesado (e.g., binario, tf, tfidf). Considere alguna estrategia para incorporar el "valor" del "Probability Factor of Affective use" en su representación vectorial del documento. Evalúa varias representaciones, y ponga una tabla comparativa a modo de resumen (e.g., binario, frecuencia, tfidf, etc.).

2. En un comentario aparte, discuta sobre la estrategía que utilizó para incorporar el "Probability Factor of Affective use".

2.3 Experimentos Parte 3

1. Utilice el recurso léxico de la actividad de representación fonética de esta tarea para construir una Bolsa de Palabras-Fonéticas. Evalúa varias representaciones (al menos binario, tf y tfidf), y ponga una tabla comparativa a modo de resumen.

2.4 Experimentos Parte 4

1. Combine todo lo anterior en experimentos con una Bolsa de Palabras Tradicional con diferente pesado y observé si la clasificación mejora cuando se incorpora algo de lo anterior. Pruebe al menos tres pesados: binario, frecuencia normalizada y tfidf. Para construir la representación final del documento utilice la concatenación de todas representaciones anteriores (Bolsa de Palabras Normal + Bolsa de Sentimientos de Canada + Bolsa de Sentimientos de Grigori + Bolsa de PalabrasFoneticas), y aliméntelas a un SVM.

2. Elabore conclusiones sobre toda esta Tarea, incluyendo observaciones, comentarios y posibles mejoras futuras.

Notas:

- El numero de palabras para toda la Actividad 3 puede fijarse en 5000, aunque puede optarse por otro numero mayor, o algún filtrado básico basado en frecuencia o tfidf.
- Sí una palabra del dataset no este en los recursos léxicos, diseñe algo básico para lidiar con ello, por ejemplo, podría simplemente ignorarla en esa representación.