T2: Minería de Texto Básica

CIMAT, Procesamiento de Lenguaje Natural, Ciencias de la Computación

Profesor: Dr. Adrián Pastor López Monroy

Entregar: Lunes 12 Febrero de 2024 antes de las 23:59:59

1 Instrucciones

Realiza los siguientes puntos en un notebook de Python *lo mejor organizado y claro posible*. Ponga su nombre completo al archivo de entrega (e.g., adrian_pastor_lopez_monroy.ipynb) y también en la primera celda del notebook junto con el número de Tarea. Al entregar la tarea, sube al classroom el notebook como un archivo. El notebook deberá haber sido ejecutado en tú máquina y mostrar el resultado en las celdas. Puede usar libremente el código de la clase *"BoW"* para completar esta actividad. Puede usar su propio código también, pero no debe usar librerías para hacer BoW y derivados (e.g., sklearn, gensim, etc.).

- El numero de palabras para toda la tarea puede fijarse en las 5000 más frecuentes, aunque pueden ser más, y usar algún otro filtrado básico basado en frecuencia o tfidf.
- Sí una palabra del dataset no está en los recursos léxicos, diseñe algo básico para lidiar con ello, o podría simplemente ignorarla en esa representación.

1.1 Se vale pedir ayuda y "copiar"

Se vale pedir ayuda y/o copiar con atribución entre los miembros de la clase y apegándose estrictamente a los siguientes puntos:

1. Del total de actividades que se solicitan hacer (15 para esta tarea) solo puedes pedir ayuda/copiar en un total de **dos**.
2. Para los puntos dónde se pide ayuda, brevemente escribe en qué pediste ayuda y a quién.
3. Si tuviste que reusar alguna parte de código que no es tuyo, deja claro dos cosas: 1) brevemente porque tuviste dificultad para hacerlo, 2) cómo lo solucionó tu compañero.

2 Bolsas de Palabras, Bigramas y Emociones (40pts)

Representa los documentos y clasifica con SVM similar a la Práctica 3, pero con diferentes pesados de términos.

1. Evalué BoW con pesado binario.
2. Evalué BoW con pesado frecuencia.
3. Evalué BoW con pesado tfidf.

4. Evalué BoW con pesado binario normalizado l2 (no use sklearn).
5. Evalué BoW con pesado frecuencia normalizado l2 (no use sklearn).
6. Evalué BoW con pesado tfidf normalizado l2 (no use sklearn).
7. Ponga una tabla comparativa a modo de resumen con las seis entradas anteriores.
8. De las configuraciones anteriores elija la mejor y evalúela con más y menos términos (e.g., 1000 y 7000). Ponga una tabla dónde compare las tres configuraciones.
9. 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 (BoE).
10. Evalúa tú BoE clasificando con SVM. Ponga una tabla comparativa a modo de resumen con los tres pesados, normalize cada uno si lo cree conveniente.

3 Recurso Línguistico de Emociones Mexicano (30 pts)

1. Utilice el recurso léxico llamado "Spanish Emotion Lexicon (SEL)" del Dr. Grigori Sidorov, profesor del Centro de Investigación en Computación (CIC) del Instituto Politécnico 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). Proponga alguna estrategia para incorporar el "valor" del "Probability Factor of Affective use" en su representación vectorial del documento. Evalúa y escribe una tabla comparativa a modo de resumen con al menos tres pesados: binario, frecuencia, tfidf. Normalize cada pesado según lo crea conveniente.
2. En un comentario aparte, discuta sobre la estrategía que utilizó para incorporar el "Probability Factor of Affective use". No más de 5 renglones.

4 ¿Podemos mejorar con Bigramas? (30 pts)

1. Hacer un experimento dónde concatene una buena BoW según sus experimentos anteriores con otra BoW construida a partir de los 1000 bigramas más frecuentes.
2. Hacer un experimento con las Bolsas de Emociones, Bolsa de Palabras y Bolsa de Bigramas; usted elige las dimensionalidades. Para construir la representación final del documento utilice la concatenación de las representaciones según sus observaciones (e.g., Bolsa de Palabras + Bolsa de Bigramas + Bolsa de Sentimientos de Canadá + Bolsa de Sentimientos de Grigori), y aliméntelas a un SVM.
3. Elabore conclusiones sobre toda esta Tarea, incluyendo observaciones, comentarios y posibles mejoras futuras. Discuta el comportamiento de la BoW de usar solo palabras a integrar bigramas, y luego a integrar todo ¿ayudó? o ¿empeoró?. Discuta también brevemente el costo computacional de los experimentos ¿Valió la Pena tener todo?. Sea breve: todo en NO más de dos párrafos.