

# T2: Minería de Texto Básica

CIMAT, Procesamiento de Lenguaje Natural, Ciencias de la Computación

Profesor: Dr. Adrián Pastor López Monroy

Entregar: Viernes 26 Febrero de 2021 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 "*Lecture 8: Práctica 3 (Sugerencia BoW)*" y "*Lecture 9: Práctica 4 (Sugerencia DOR)*" 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 al Lecture 8, pero con diferentes pesados de términos.

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

3. Evalúe BoW con pesado tfidf.
4. Evalúe BoW con pesado binario normalizado l2 (no use sklearn).
5. Evalúe BoW con pesado frecuencia normalizado l2 (no use sklearn).
6. Evalúe 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. De las configuraciones anteriores elija la mejor y concatene una bolsa-de-bigramas de tamaño 1000 con algún pesado que usted determine y normalizada con l2. Evalúela en clasificación y discuta si ve mejoría.
10. 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).
11. 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 Lingüístico 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. Normalice cada pesado según lo crea conveniente de acuerdo el experimento (1).
2. En un comentario aparte, discuta sobre la estrategia que utilizó para incorporar el "Probability Factor of Affective use". No más de 5 renglones.

### **4 ¿Le podemos ganar a BoW con Bigramas? (30 pts)**

1. Combine algo de emociones con Bolsa de Palabras y Bigramas. 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 Canada + Bolsa de Sentimientos de Grigori), y aliméntelas a un SVM.
2. Elabore conclusiones sobre toda esta Tarea, incluyendo observaciones, comentarios y posibles mejoras futuras. Discuta que tanto pudo mejorar la BoW con Bigramas, o ¿empeoró?. Discuta también brevemente el costo computacional de los experimentos. Sea breve: todo en NO más de un párrafo; máximo dos.