Proyecto 1: Minería de Texto para Turismo

Dr. A. Pastor López (CIMAT), Dr. Rafael Guerrero (DCEA-Universidad de Guanajuato)

Entregar: Viernes 25 de Marzo de 2022 antes de las 23:59:59

Contexto

Realiza los siguientes puntos en un notebook de Python *lo mejor organizado y claro posible*. Ponga su nombre al notebook (e.g., adrian_pastor_lopez_monroy.ipynb) y también en la primera celda del notebook junto con el número de Tarea. Sube al classroom el notebook como un archivo, que deberá haber sido ejecutado en tú máquina y mostrar el resultado en las celdas.

Para este proyecto consideraremos el conjunto de datos recolectado por el equipo del Dr. Rafael Guerrero, profesor en la División de Ciencias Económico Administrativas de la Universidad de Guanajuato. Este conjunto de datos contiene un aproximado de diez mil opiniones de turistas en trip advisor en 10 sitios turísticos de la ciudad de Guanajuato. **El objetivo es realizar las siguientes actividades y contestar las preguntas.** Para esta tarea se puede usar cualquier librería o herramienta de Python (e.g., sklearn, keras, nltk, códigos de github de otras personas (citando), etc.). También puede reusar su código de tareas previas, o puede simplemente usar TfIdfVectorizer, CountVectorizer, etc. de sklearnt. Puede usar también SelectKBest como en el Lecture de DOR lo hizo el profesor, o usar su propio código de Chi2.[1]

Para estas actividades usted determine el número de features (palabras) de alguna forma según su intuición. Usualmente el top 10k con base en frecuencia podría ser buena elección si su hardware es suficiente para llevar a cabo las actividades. Sí su reducción es a 5k o menos términos, algo con base en Chi2, Ganancia de Información o valores TFIDFs podría venir mejor para no perder tanta información y llegar a buenas conclusiones.

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 (12 con valor para esta tarea) solo puedes pedir ayuda/copiar en un total de **tres**.
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 resolvió tu compañero.
4. Se darán hasta 5pts/100pts extra en esta tarea (no son acumulables, es decir la máxima nota es 100) si ayudaste a algún compañero y se comenta en ambas tareas. Se pueden

[1] Recomiendo ampliamente usar lo más posible las funciones de Sklearnt, para aprender a usarlas además de que son muy eficientes al llevar todo en matrices sparse. Esto hará que puedas manipular vocabularios enormes y más rápido.

ayudar mutuamente en diferentes puntos y ganar cada quién los 5pts/100pts extra. Siempre respetando el punto 1.

Actividades (50pts)

1. (2.5pts) Construya estadísticas básicas respecto a la opinión de cada lugar turístico. **Preprocese y limpie el texto según sus intuiciones y argumente brevemente sobre ello.** Considere scores de 4 a 5 como **positivos**, calificaciones de 3 como **neutros** y las de 2 a 1 como **negativos**. Es interesante ver:

   (a) Promedios de calificación por lugar, y desviaciones estándar en los scores
   (b) Basado en palabras: longitud promedio de opiniones y desviacones estándar
   (c) Histogramas de edades de opiniones por lugar
   (d) Histograma de tipo de visitantes (nacional o internacional) por lugar
   (e) Sugiere dos más interesantes para ti.

2. (2.5pts) Utilizando una estrategia de feature selection (se sugiere $\mathcal{X}^2$ o ganancia de información) visualice con *word_cloud* nubes de palabras el top k (se sugiere 50) de palabras más relevantes para cada uno de los 10 lugares. Note que serán 10 nubes, una por lugar.

3. (15pts) Para cada uno de los 10 sitios turísticos, haga un descubrimiento automático de los 3 tópicos con LSA (componentes principales) más relevantes y 10 palabras contenidas en cada tópico de cada uno de los siguientes subgrupos:

   (a) Hombres
   (b) Mujeres
   (c) Turistas Nacionales
   (d) Turistas Internacionales
   (e) Jovénes (elige un rango de edad interesante con base en sus estadísticas)
   (f) Mayores (elige un rango de edad interesante con base en sus estadísticas)

   Antes de aplicar LSA, asegúrese de hacerlo sobre una matriz lo más grande posible (para su hardware) de TFIDF Normalizada a L2. Note que para cada sitio turístico deberá saber cuales son los 3 temas de interés y sus palabras, para cada uno de estos subgrupos. Como sugerencia puede usar la función TruncatedSVD de sklearnt para obtener la descomposición de matrices como se sugiere en el siguiente video para implementar LSA: https://www.youtube.com/watch?v=hB51kkus-Rc. También podría llevar a cabo svd con numpy.

4. (5pts) Para cada uno de los 10 sitios turísticos, haga una nube de palabras que muestre las palabras más asociadas a sus opiniones negativas utilizando $\mathcal{X}^2$. Puede usar funciones de sklearn o hacerlo tal como en las diapositivas del autor Ricardo Baeza (diapositivas dónde está lo de información mutua).

5. (15pts) Para cada uno de los 10 sitios turísticos construya tres Bolsas de Palabras de la siguiente manera: i) 1000 términos con mayor peso tfidf, ii) 2000 bigramas con mayor tfidf, y iii) 1000 trigramas con mayor tfidf. Luego concatene las tres representaciones que fueron calculadas de forma independiente, con sus propios tfidfs según su espacio y su propio L2. Finalmente sobre todo ese espacio concatenado de 4000 características aplique ganancia de información o $\mathcal{X}^2$ y obtenga los 1000 features más relevantes. Muestre una nube de palabras con el top 50 features relevantes para cada lugar turístico (10 nubes en total).

6. (5pts) Muestre la ocurrencia temporal de las 10 palabras con más ganancia de información de TODO el dataset para cada sitio turístico. Aquí se sugiere usar la gráfica de estilo de los discursos de primer año de los presidentes en USA dada como ejemplo en NLTK. La idea sería tener una gráfica por sitio turístico y la ocurrencia marcada en azul del top 10 palabras de con mayor ganancia de información o chi en todo el dataset.

   Por ejemplo, si las 10 palabras con mayor ganancia de información en todo el dataset con respecto a las clases positivas, neutral y negativas son: $w_1, ...w_{10}$ entonces haga una gráfica de disperción temporal por lugar de como es la aparición de las palabras $w_1, ...w_{10}$ en cada sitio turistico.

7. (5pts) Diseñe un análisis temporal (formato libre) que muestre opiniones positivas, negativas y neutras a través de los meses y años para todos los sitios turísticos. En pocas palabras mostrar la evolución de las opiniones a través del tiempo.

1 (50pts) Preguntas: Conteste lo más detallado posible lo siguiente, dando argumentos y conclusiones claras según su análisis previo. Cada respuesta entre 150 y 300 palabras.

1. (10pts) ¿De los sitios turísticos, cual diría usted que es el más polémico y **la razón de ello**?

2. (10pts) En cuanto al sitio más polémico, ¿Como es la diferencia de opinión y temas entre turistas nacionales e internacionales?

3. (10pts) ¿Cual diría que es el sitio que le gusta más a las mujeres y por qué?

4. (10pts) ¿Cual diría que es el sitio que le gusta más a las personas jóvenes y por qué?

5. (10pts) ¿Qué otras observaciones valiosas puede obtener de su análisis? (e.g., ¿identificó de que se queja la gente? ¿qué tipo de cosas le gustó a la gente?, etc.)