Laboratorio 6.
Análisis de redes sociales

INSTRUCCIONES:

Se obtuvieron aproximadamente 5000 tweets de las cuentas de @traficogt y @BArevalodeLeon hasta el 12 de septiembre de 2024. Debe seleccionar uno de los dos conjuntos de datos y resolver los ejercicios. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios.

PROBLEMAS A RESOLVER

PROBLEMA 1. @TráficoGT

Preguntas interesantes: ¿Cómo complicó el tráfico en toda la ciudad la época de lluvia? ¿Cuáles son las áreas de la ciudad que más se congestionaron? ¿Cree usted que se mantengan esas áreas este año? ¿En qué horarios según los usuarios de X se hacen los mayores atascos?

PROBLEMA 2. @BArevalodeLeon

Preguntas interesantes: ¿Qué aceptación tenía Bernardo Arévalo como presidente de Guatemala el año? ¿Cuál es la popularidad que tiene en estos momentos?

EJERCICIOS

1. Descargue los archivos de datos (traficogt.txt, tioberny.txt)
2. Cargue los archivos de datos a R o a Python.
3. Limpie y preprocese los datos. Describa de forma detallada las actividades de preprocesamiento que llevó a cabo.
   3.1. Se pueden hacer tareas como:
      - Convertir el texto a mayúsculas o a minúsculas
      - Quitar los caracteres especiales que aparecen como “#”,”@” o los apóstrofes.
      - Quitar las url
      - Revisar si hay emoticones y quitarlos, ¿conviene quitarlos para este ejercicio?
      - Quitar los signos de puntuación
      - Quitar los artículos, preposiciones y conjunciones (stopwords)
      - Quitar números si considera que interferirán en los análisis.
   3.2. Cada registro de tweet tiene una estructura JSON que incluye metadatos como ID de usuario, texto del tweet, menciones, retweets, y favoritos. Asegúrate de extraer las menciones, respuestas y retweets para identificar las relaciones entre usuarios.
   3.3. Preprocesa los datos eliminando duplicados. Normaliza los nombres de usuario y las menciones para evitar inconsistencias.
   3.4. Crea una estructura de datos eficiente para el análisis de redes (e.g., un DataFrame o una matriz de adyacencia). Asegúrate de representar las interacciones como grafos dirigidos, donde los nodos representan usuarios y las aristas representan las interacciones entre ellos (retweets, menciones, respuestas).

4. Haga un análisis exploratorio de los datos para entenderlos mejor, documente todos los análisis. Escriba una serie de insights que se puedan seguir investigando.
   4.1. Sugerencias para Análisis Exploratorio:
      - Identifica las menciones, respuestas y retweets en cada tweet y piensa en cómo podrías utilizar esta información para analizar las interacciones entre usuarios.
      - Realiza un análisis básico de los datos, como calcular el número de tweets, usuarios únicos y menciones en cada conjunto de datos, hashtags frecuentes. Puede hacer una nube de palabras
   4.2. Redacte al menos 3 preguntas interesantes que le surjan al explorar los datos y respóndalas.
5. Análisis de la topología de la red
   5.1. **Construcción y visualización de grafos:** Utiliza herramientas para construir y visualizar grafos dirigidos de las interacciones entre usuarios en la red seleccionada. Muestra claramente los nodos más conectados y las relaciones de poder dentro de las comunidades.
   5.2. **Cálculo de métricas de red clave:** Calcula las siguientes métricas para la red seleccionada y las discute:
      - **Densidad de la red:** Mide cuántos enlaces existen en relación con el número máximo posible de enlaces.
      - **Diámetro de la red:** Mide la distancia máxima entre dos nodos más distantes.
      - **Coeficiente de agrupamiento:** Indica el grado en que los nodos tienden a formar clústeres o grupos dentro de la red.
6. Identificación y análisis de comunidades
   6.1. **Aplicación de algoritmos de detección de comunidades:** Investigue cuál es el algoritmo más usado para detección de comunidades y úselo para la red seleccionada.
   6.2. **Visualización y caracterización de comunidades:**
      - Visualiza las comunidades detectadas, resaltando los grupos más grandes e influyentes. Caracteriza cada comunidad en términos de su tamaño, interacciones, y temas principales de conversación.
      - Haga un gráfico con las 3 comunidades más grandes.
7. Análisis de influencers y nodos clave
   7.1. **Identificación de usuarios influyentes:** Utiliza métricas de centralidad para identificar los usuarios más influyentes en la red seleccionada.
      - **Centralidad de grado:** Usuarios con más conexiones directas (menciones o retweets).
      - **Centralidad de intermediación:** Usuarios que actúan como puentes en la red, conectando comunidades.
      - **Centralidad de cercanía:** Usuarios que pueden llegar a todos los demás nodos con la menor cantidad de saltos.

8. Detección y análisis de grupos aislados
   8.1. **Análisis de subredes y nodos aislados:** Identifica grupos o subredes aisladas dentro de red seleccionada, que interactúan muy poco con el resto. Realiza un análisis de estos grupos para entender su dinámica y si representan nichos específicos dentro de la red.
9. Análisis de contenido y sentimiento
   9.1. **Análisis de sentimiento:** Realiza un análisis de sentimiento de los tweets en la red seleccionada utilizando técnicas de procesamiento de lenguaje natural (NLP). Identifica si los usuarios están discutiendo temas de manera positiva, negativa o neutral.
   9.2. **Identificación de temas:** Utiliza un análisis de tópicos para identificar los temas principales en cada red. ¿Qué temas son más recurrentes en @traficogt o @bernardoarevalodeleon? ¿Cómo se relacionan estos temas con las comunidades detectadas?
10. Interpretación y contexto
    10.1. **Contextualización de los hallazgos:** Explica los hallazgos en un contexto más amplio. ¿Cómo influyen los influencers y las comunidades en la formación de opiniones públicas?

HERRAMIENTAS SUGERIDAS

Nota: Estas herramientas fueron sugeridas por ChatGPT, no es obligatorio su uso, si conoce o se siente cómodo usando otras herramientas hágalo.

Python:

- **Análisis de la topología de la red:**
  - **networkx**: Biblioteca clave para el análisis de redes y grafos en Python. Permite construir, manipular y analizar redes, además de calcular métricas como centralidad, cohesión y detectar comunidades.
  - **igraph** (para Python): Similar a la versión de R, permite realizar análisis más rápidos y escalables en redes grandes.
  - **pygraphviz**: Para trabajar con grafos y visualizaciones de redes más complejas.
- **Análisis de influencers y nodos clave:**
  - **networkx**: Calcula métricas de centralidad (grado, intermediación, cercanía, etc.), ideal para identificar los nodos clave en la red (influencers).
- **Análisis de contenido y sentimiento:**
  - **nltk**: Biblioteca fundamental para procesamiento de lenguaje natural (NLP), con herramientas para tokenización, análisis de frecuencia y extracción de características de texto.
  - **TextBlob**: Proporciona una API simple para realizar análisis de sentimiento y análisis gramatical.
  - **VADER** (de nltk.sentiment): Especialmente diseñado para analizar sentimientos en redes sociales, tiene una alta precisión con textos cortos como tweets.
  - **spaCy**: Un motor de NLP más avanzado y rápido, útil para análisis de contenido más profundos como detección de entidades nombradas (NER).

    - **transformers** (de Hugging Face): Si buscas aplicar modelos más complejos de análisis de texto, esta biblioteca incluye modelos de última generación como BERT para análisis de sentimientos y clasificación de texto.
- **Visualización avanzada:**
    - **plotly**: Para crear gráficos interactivos y explorar los datos de manera dinámica.
    - **PyVis**: Para la visualización interactiva de redes.
    - **Bokeh**: Alternativa a plotly, permite generar gráficos interactivos y dashboards complejos.
    - **Gephi**: Aunque no es Python puro, puedes exportar los datos de networkx o igraph a Gephi para visualización avanzada de redes.
    - **igraph:** Permite hacer visualizaciones estáticas de comunidades y de grafos

R:

- **Análisis de la topología de la red:**
    - **igraph**: Uno de los paquetes más populares para el análisis de redes. Permite construir, manipular y visualizar grafos, además de calcular métricas de red como la densidad, el diámetro y el coeficiente de agrupamiento.
    - **network**: Similar a igraph, pero con un enfoque más en la visualización de redes y la integración con otros paquetes como statnet.
    - **ggraph**: Paquete para la visualización de redes usando la gramática de gráficos de ggplot2.
- **Identificación y análisis de comunidades:**
    - **igraph**: También incluye algoritmos de detección de comunidades como Louvain y Girvan-Newman.
    - **clustree**: Para visualizar y comparar estructuras de clusters o comunidades.
- **Análisis de influencers y nodos clave:**
    - **igraph**: Permite calcular métricas de centralidad como grado, intermediación, y cercanía.
    - **CINNA**: Un paquete más especializado en métricas de centralidad.
- **Análisis de contenido y sentimiento:**
    - **textclean**: Herramienta útil para limpiar y normalizar el texto.
    - **tidytext**: Paquete para el análisis de texto en un formato de datos ordenado, perfecto para trabajar con grandes conjuntos de tweets.
    - **syuzhet**: Para el análisis de sentimiento, que incluye métodos como análisis basado en léxicos de emociones.
    - **sentimentr**: Otro paquete para realizar análisis de sentimiento a nivel de oración o documento.
- **Visualización avanzada:**
    - **ggraph**: Potente herramienta de visualización de datos que permite generar gráficos avanzados de redes y análisis
    - **plotly**: Para crear visualizaciones interactivas que los usuarios puedan explorar dinámicamente.

- **visNetwork**: Un paquete para crear visualizaciones interactivas de redes.
-

EVALUACIÓN

NOTA: **La evaluación de cada integrante del grupo será de acuerdo con sus contribuciones al trabajo grupal**

(12 puntos) Limpieza y preprocesamiento:

- (5 puntos) Describe de forma detallada todas las actividades que llevó a cabo para limpiar y preprocesar los datos. Incluidas las tareas de normalizar nombres de usuario y menciones para eliminar inconsistencias.
- (4 puntos) Identifica y extrae las menciones, respuestas y retweets de cada uno de los tweets para posterior análisis como red social
- (3 puntos) Crea una estructura de datos (dataframe, matriz de adyacencia) que represente las interacciones entre usuarios como grafos dirigidos (nodos: usuarios; aristas: retweets, menciones, respuestas).

(20 puntos) Análisis exploratorio:

- Se elaboró un análisis exploratorio en el que se explican los datos del conjunto seleccionado. Cómo mínimo debe haberse hecho lo siguiente:
  - (5 puntos) Análisis del número de tweets, usuarios únicos, menciones, y hashtags frecuentes. Generar una nube de palabras
- (5 puntos) Más análisis que contribuyen a entender los datos
- (10 puntos) Preguntas e insights
  - (6 puntos) Formular al menos 3 preguntas interesantes que surjan durante el análisis exploratorio.
  - (4 puntos) Responder estas preguntas con base en los datos.

(15 puntos) Análisis de la Topología de la Red:

- (5 puntos) Crear y visualizar grafos dirigidos que representen las interacciones entre usuarios, destacando los nodos más conectados y las relaciones de poder dentro de las comunidades
- (5 puntos) Explicar claramente las relaciones encontradas
- (3 puntos) Calcular la densidad de la red, diámetro de la red, y coeficiente de agrupamiento.
- (2 puntos) Discutir la relevancia de estas métricas en el contexto de la red

(15 puntos) Identificación y Análisis de Comunidades.

- (4 puntos) Utiliza un algoritmo adecuado (e.g., Louvain) para detectar comunidades en la red.
- (3 puntos) Explica la elección del algoritmo para detección de comunidades y su aplicación en el análisis.
- (5 puntos) Hace un gráfico de todas las comunidades detectadas, resaltando los grupos más grandes e influyentes y otro de las 3 comunidades más grandes.
- (3 puntos) Caracterizar las 3 comunidades más grandes en términos de tamaño, interacciones, y temas principales de conversación.

(10 puntos) Análisis de Influencers y Nodos Clave.

- (4 puntos) Calcula y explica la centralidad de grado, centralidad de intermediación y centralidad de cercanía.
- (6 puntos) Identificar y justificar quiénes son los usuarios más influyentes en función de estas métricas

(8 puntos) Detección y Análisis de Grupos Aislados.

- (4 puntos) Calcula y explica la centralidad de grado, centralidad de intermediación y centralidad de cercanía.
- (4 puntos) Identificar subredes o grupos aislados, y realizar un análisis de su dinámica e influencia dentro de la red principal.

(10 puntos) Análisis de Contenido y Sentimiento.

- (5 puntos) Realiza un análisis de tópicos para identificar los temas principales en los tweets, explicando cómo estos temas se relacionan con las comunidades detectadas. Explica los resultados del análisis
- (5 puntos) Realiza un análisis de sentimiento utilizando técnicas de NLP, identificando si los usuarios discuten temas de forma positiva, negativa o neutral. Explica los resultados del análisis

(20 puntos) Interpretación y Contexto.

- (6 puntos) Explicar cómo los influencers y las comunidades detectadas influyen en la formación de opiniones públicas, contextualizando los hallazgos dentro de un marco social más amplio.
- (6 puntos) Responder las preguntas interesantes planteadas en la sección de problemas a resolver.
- (8 puntos) Hace una sección de conclusiones en su entregable que resume los hallazgos que arrojó el análisis de la red social con el conjunto de datos seleccionado.

MATERIAL A ENTREGAR

- Informe que contenga, los resultados de los análisis y las explicaciones.
- Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó.
- Link del repositorio usado para versionar el código.

FECHAS DE ENTREGA

- **AVANCE:** Jueves 4 de septiembre de 2025 17:20hrs: Actividades de la **1 a la 5** de la sección de ejercicios
- **DOCUMENTO FINAL COMPLETO:** Domingo 7 de septiembre de 2025 a las 23:59.

NOTA: Para poder tener nota completa debe entregar las asignaciones en el tiempo adecuado. No se calificará el avance del laboratorio si no fue entregado en tiempo, aunque esté en el repositorio.

Sugerencia: El segundo día de clase de la semana tendrá un tiempo de aclaración de dudas con el profesor, se le sugiere que avance en la resolución del laboratorio en los pasos del contenido teórico visto en la clase presencial para que aclare todas sus dudas al respecto en dicho espacio.