Prueba de conocimiento VIVI: científico de datos no estructurado nlp

Lina María Ortiz Parra

Contenidos

Exploración de datos

- De la información que se tenía de cada tweet para este análisis escogí: el usuario y Embedded text
- Luego de cada tweet se eliminaron los objetos innecesarios que puede afectar el rendimiento de algoritmo, tales como:
  - hastags
  - links
  - puntuación

  Además, todos los textos son convertidos a minúsculas para evitar que el algoritmo interprete la misma palabra como diferente.

Técnicas usadas o exploradas

- Primero se realiza Tokenización, lematización y eliminación de palabras vacías
  - Las palabras vacías son palabras de uso común cuya presencia en una oración tiene menos peso en comparación con otras palabras. Incluyen palabras como 'y', 'o', 'tiene', etc.
  - La tokenización es el proceso de dividir una cadena en una lista de tokens.
  - La lematización consiste en reducir una palabra a su forma raíz.
- Para la clasificación de los tweets cree un conjunto de palabras por categorias.
- Se convirte el texto de los tweets y de las palabras claves en forma numérica.
- Se utilizo Jaccard similarities y se obtuvo los scores para cada categoría

Dificultades encontradas

- Una de las dificultades encontradas fue saber de que forma se querian categorizar los tweets
- No pude hacer uso de una API que me fuera útil para el desarrollo de este, por lo que generé un set de palabras por categoría

Resultados

- Las categorías en los que se clasificaron los tweets se pueden mejorar, agregando más palabras claves en cada una de las categorías, o haciendo uso de API que ayuden con este proceso.
- Se presentaron casos en los que un tweet esta en varias categorías, para resolver este caso se podría realizar un cluster.