Análisis de Sentimiento de lugares históricos del Paraguay basado en Técnicas de Procesamiento del Lenguaje Natural.

Andres Villamayor - Edgar Mencia | UCOM | https://github.com/edmenciab733/ucom-project3

Objetivo

Investigar técnicas de procesamiento del lenguaje natural(NLP) para la clasificación de sentimientos en texto acerca de lugares de Paraguay

Justificación

- Encontrar soluciones a nuestra para nuestro idioma, ya que la gran mayoría de los ejemplos están en el idioma ingles.
- Aprovechar las opiniones que tenían los turistas sobre nuestro lugares históricos, para determinar calificaciones probables.

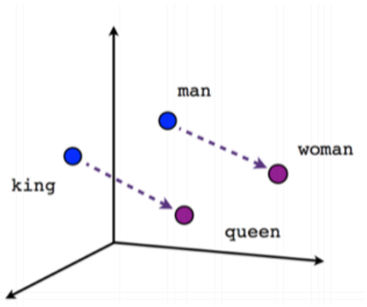
Descripción del proyecto

Este proyecto conocido como Análisis de Sentimiento de lugares históricos del Paraguay basado en Técnicas de Procesamiento del Lenguaje Natural está dentro del área de estudio de la Universidad Comunera, enfocado en los esquemas y aprendizajes de Machine Learning. Con el desarrollo de nuevos algoritmos en la búsqueda del modelado del texto principalmente en inglés, nos hemos visto con la necesidad de implementar estos conceptos en el español. Hemos abarcado desde la generación de un dataset sobre opiniones que tienen turistas que llegan a nuestro país acerca de los distintos lugares que visitan hasta generar modelos de bolsas de palabras (bag of words) o de incrustaciones(embeddings) para encontrar el algoritmo adecuado para este tipo de representación.

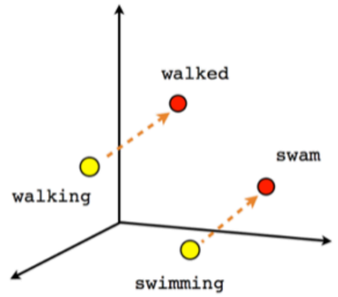
Metodología Científica

- Por el tipo de proyecto, la metodología de investigación es exploratoria, pues la misma se basa en encontrar correlaciones con los papers científicos acerca de minería de opinión, matemática lingüística e implementar en el castellano.

Embeddings



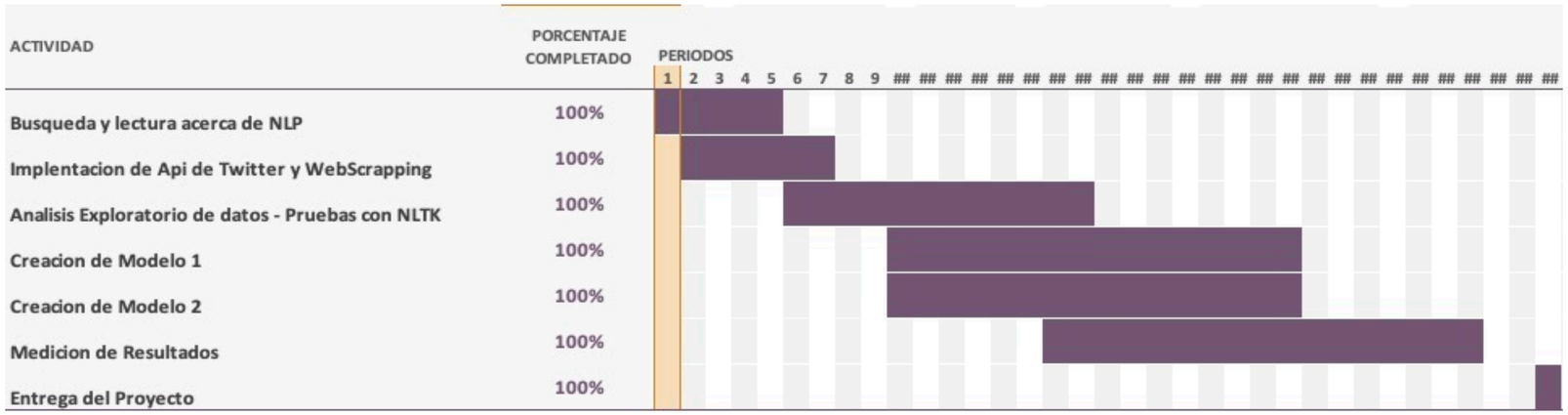
Male-Female



Verb tense

Modelo BoW(Bag of Words)

Planificación



Conclusión

- Para clasificar documentos de un tamaño, no muy grande, es recomendable utilizar el modelo de bolsa de palabras, como es este caso, ya que da mejores resultados en el momento de clasificar.
- Se da una mejor clasificación cuando los conjuntos de datos son mas grandes y mejor distribuidos, nos topamos con un dataset en la que el mayor cantidad de datos estaba entre el 4 y 5 estrellas.
- Los resultados han sido satisfactorios, pues si bien en el testing el porcentaje es muy bajo se logro detectar que es por el dataset sesgado

Bibliografia

- Introducción a Word2vec (skip gram model), Medium, URL: https://medium.com/@gruizdevilla/introducci%C3%B3n-a-word2vec-skip-gram-model-4800f72c871f
- An introduction to Bag of Words and how to code it in Python for NLP, freeCodeCamp.org, https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/