INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

DEPARTAMENTO DE ELECTRÓNICA, SISTEMAS E INFORMÁTICA
MAESTRÍA EN SISTEMAS COMPUTACIONALES

MSC2526A – PROGRAMACIÓN PARA ANÁLISIS DE DATOS
PRIMAVERA 2021

ITESO, Universidad Jesuita de Guadalajara

ANÁLISIS DE DISCURSOS HISTÓRICOS

PROYECTO FINAL
Presenta:

Jaime A. Santos Orozco,
747150, jaime.santos@iteso.mx

Jalisco. Mayo de 2023.

RESUMEN

El desarrollo del presente reporte tiene como fin explicar el proyecto, dividiendo los contenidos del mismo en las secciones pertinentes, donde se pretende explicar los objetivos propuestos como finalidad del desarrollo del proyecto, la obtención de los datos y la manera en la que se obtuvieron los mismos, la preparación de los datos donde se explican todas las operaciones que se realizan sobre los datasets y la comprensión de los mismos, la explicación de los modelos utilizados y la conclusión a la que se llega.

El objetivo del trabajo es analizar datasets de discursos históricos en inglés, clasificados como de alto impacto y encontrar similitudes por medio de librerías de análisis de texto con el objetivo de construir un modelo que permita clasificar otros discursos como de alto, o bajo impacto social.

Los datasets incluyen 80 archivos txt con la transcripción de 80 discursos en inglés, de diferentes contextos, tales como políticos, premios nobel, guerra, liberación, etc.

El desarrollo se realizó en Python, utilizando las librerías de nltk, con la cual se procesaron los datasets y pandas, numpy y sklearn, con las que se creó un modelo de predicción de KNN.

El modelo obtenido obtuvo resultados regulares con una precisión de 35%, lo que indica que los parámetros obtenidos a través del procesamiento de los discursos no fue suficiente para la predicción confiable.

Sin embargo los resultados obtenidos con k-Means fueron mucho mejores y se logró obtener una clasificación de los discursos mucho más confiable, misma que servirá para clasificar nuevos discursos después de procesar los mismos.

TABLA DE CONTENIDOS

1. OBJETIVO DE INVESTIGACIÓN

Resumen: *El objetivo de la investigación es conocer si, a través de técnicas de análisis de textos se puede llegar a un clasificador confiable para calificar discursos como de alto impacto o de bajo impacto, en 9 niveles para KNN y 5 para K-Means*

1.1. Introducción

En el presente trabajo se abordan las técnicas utilizadas para el procesamiento de los datos, así como la limpieza y uso, y construcción de los modelos de predicciones.

El aprendizaje de máquina (*Machine Learning*) estudia el aprendizaje automático a partir de datos (*data-driven*, gobernado por los datos) para conseguir hacer prediciones precisas a partir de observaciones con datos previos.
La clasificación automática de objetos o datos es uno de los objetivos del aprendizaje de máquina. Podemos considerar tres tipos de algoritmos:

- **Clasificación supervisada:** disponemos de un conjunto de datos (por ejemplo, imágenes de letras escritas a mano) que vamos a llamar datos de entrenamiento y cada dato está asociado a una etiqueta (a qué letra corresponde cada imagen). Construímos un modelo en la fase de entrenamiento (training) utilizando dichas etiquetas, que nos dicen si una imagen está clasificada correcta o incorrectamente por el modelo. Una vez construído el modelo podemos utilizarlo para clasificar nuevos datos que, en esta fase, ya no necesitan etiqueta para su clasificación, aunque sí la necesitan para evaluar el porcentaje de objetos bien clasificados.
- **Clasificación no supervisada:** los datos no tienen etiquetas (o no queremos utilizarlas) y estos se clasifican a partir de su estructura interna (propiedades, características).
- **Clasificación semisupervisada:** algunos datos de entrenamiento tienen etiquetas, pero no todos. Este último caso es muy típico en clasificación de imágenes, donde es habitual disponer de muchas imágenes mayormente no etiquetadas. Estos se pueden considerar algoritmos supervisados que no necesitan todas las etiquetas de los datos de entrenamiento.

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.
El algoritmo consta de tres pasos:

1. **Inicialización:** una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
2. **Asignación objetos a los centroides:** cada objeto de los datos es asignado a su centroide más cercano.
3. **Actualización centroides:** se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso. [1]

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Si bien se puede usar para problemas de regresión o clasificación, generalmente se usa como

un algoritmo de clasificación, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro. [2]

1.2. Antecedentes

En la literatura existen varios ejemplos de investigaciones de procesamiento de textos, orientados a varios fines tales como clasificar emociones, filtrar contenidos raciales o no deseados, reconocer contextos, sistemas de pregunta-respuesta.
Incluso existen librerías de machine learning bastante utilizadas y repositorios de los mismos tales como *hugging face* que permiten hacer uso de herramientas para tokenizar, clasificar y procesar textos.

1.3. Justificación

A pesar de la literatura existente y de las herramientas disponibles, si bien existe como tal software que permite ayudar a redactar, de manera clara y concisa a los usuarios, generalmente este no clasifica o califica el resultado final de un escrito completo, sino que en general, únicamente se enfocan en la gramática. En este trabajo se intenta dar un paso adelante en este aspecto, y se propone, con base en escritos públicos e históricos, clasificar redacciones.

1.4. Problema

No existen como tal clasificadores o calificadores de la calidad de los textos, con base en escritos públicos e históricos, clasificar redacciones.

1.5. Hipótesis

Por medio de técnicas de clasificación, tales como KNN y técnicas de análisis de textos básicas, es posible calificador la calidad de los textos, con base en escritos públicos e históricos.

1.6. Objetivos

1.6.1. Objetivo General:

Utilizar transcripciones de discursos históricos, propiamente clasificados en popularidad e impacto para generar clasificadores de discursos.

2. OBTENCIÓN DE LOS DATOS

Resumen: *En esta sección se presentan las técnicas utilizadas para la obtención de los discursos*

2.1. Proceso de obtención de los datos

Si bien existen bases de datos de discursos, fue difícil encontrar una que tuviera una cantidad suficiente de los mismos, y por supuesto, ya clasificados.
Los datos fueron obtenidos de la página https://www.americanrhetoric.com/top100speechesall.html, en donde se encuentran las transcripciones en formato PDF.

2.2. Conclusiones

A pesar de no contar con el formato deseado, se utilizó la base de datos antes mencionada por la calidad de las transcripciones y la cantidad de discursos presentados.

3. PREPARACIÓN DE LOS DATOS

Resumen: *En esta sección se explica cómo fue que se preparan los datos para el uso y el análisis estadístico de los mismos.*

3.1. Proceso de preparación de los datos

Desafortunadamente el formato de los discursos no fue el adecuado, por lo que primero fue necesario crear un archivo de texto plano para cada uno de los discursos.
Después se procedió a limpiar los discursos de los membretes de la página de donde se obtuvieron.
De la misma manera fue necesario quitar imágenes incluidas en los archivos, así como partes de la transcripción en las que se especificaba quién era el locutor (en algunas transcripciones se mencionaba que interrumpían al hablante y se indicaba quién tenía la palabra en dicho momento).

3.2. Conclusiones

Debido a la poca homogeneidad de la información, en trabajo de limpieza y preparación de datos tuvo que ser manual, por lo que implicó un gran porcentaje del total del tiempo dedicado para el desarrollo del trabajo (aproximadamente un 50%).

4. EXPLORACIÓN DE LOS DATOS

Resumen: *En esta sección se explica la naturaleza de los archivos de texto que contienen los discursos utilizados para el estudio.*

4.1. Proceso de exploración de los datos

Una vez realizada la limpieza de los datos no fue necesaria mucha exploración, ya que todos contienen simples transcripciones de los discursos históricos. Simplemente se comprobó que el lenguaje de los mismos coincidiera (inglés) ya que las librerías utilizadas necesitan trabajar sobre el mismo lenguaje.

4.2. Conclusiones

La mayor parte del trabajo con los datos se realizó al momento de la preparación, para asegurar su homogeneidad. No fue necesario eliminar ningún archivo, ya que todos contaban con la información requerida, y el dataset utilizado para el modelo de predicción fue realizado en su totalidad mediante los datos provenientes de el análisis de los datos.

5. CONSTRUCCIÓN DE MODELOS

Resumen: *En esta sección se presentan los pasos realizados para la construcción de los modelos, así como las librerías y lenguajes utilizados para ello.*

5.1. Proceso de construcción de modelos

Derivado de la naturaleza de los datos, y como estos se encuentra clasificados, se decidió utilizar un modelo de KNN.

Primero se procesa cada archivo de discursos, con ayuda de la librería nltk se crea un *corpus*, mismo que contiene los 80 archivos txt a utilizar. A la par se crea un diccionario con las palabras nombradas como *stop_words,* estas no contienen significado por sí mismas. Este diccionario se utilizara en los pasos siguientes para descartar palabras.
Se procede a leer los archivos con el fin de extraer, de cada uno, el top 20 de palabras más utilizadas, excluyendo aquellas que aparecen en nuestro diccionario creado en el paso anterior.

Después, todos los top 20 de palabras se unifican en una lista, misma que contiene la cantidad de repeticiones, ahora para todos los discursos. A la par se crea un diccionario que contiene estas palabras y la cantidad de discursos en los que aparecen.
Posteriormente se crea un diccionario que servirá para asignar un *score* a las palabras, este score se calcula multplicando la cantidad de veces que dicha palabra se repite en todos los discursos por la cantidad de discursos en las que aparece, esto utilizando los diccionarios mencionados en el párrafo anterior.

Toda esta información hasta ahora obtenida se guarda en un objecto, mismo que será transformado en formato JSON, después, a través del JSON se procede a construir un dataset de Pandas con las columnas Filename, totalSignificantWords, totalWords, wordScore, peopleScore, autor, y title.

La columna peopleScore se utilizará como el resultado del clasificador y las columnas totalSignificantWords, totalWords, wordScore servirán como los inputs para entrenar el modelo.

Posteriormente se realiza el modelo de K-Means, con los mismos inputs, y una clusterización de 5, el cuál logra dar resultados mucho mejores.

5.2. Conclusiones

Para la generación del modelo de predicción se utilizaron bastantes técnicas y herramientas vistas en clase, tales como el procesamiento de información utilizando datasets, y el uso de las librerías de Python específicas para cada tarea, tales como numpy, pandas, sklearn, etc.

6. PRESENTACIÓN DE RESULTADOS

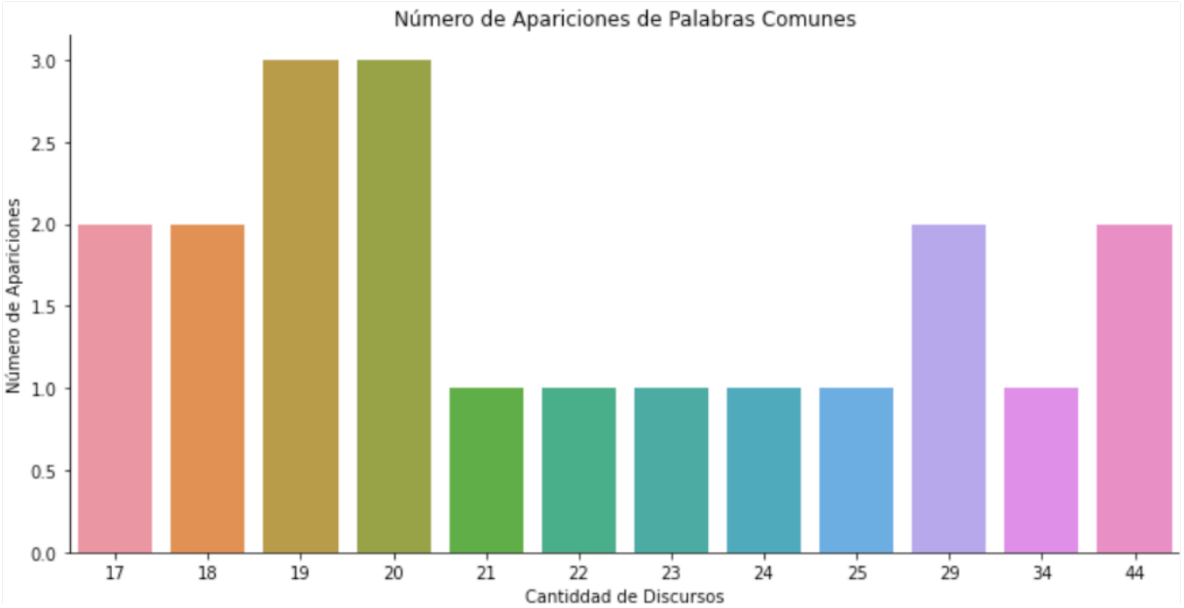
Resumen: *En esta sección se presentan los resultados obtenidos con el modelo descrito anteriormente, se presenta la matriz de confusión obtenida y las gráficas derivadas del análisis estadístico de los textos utilizados.*

6.1. Presentación de resultados

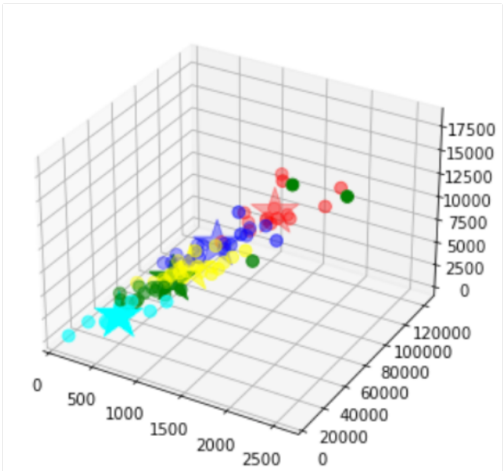
A continuación se presentan los resultados estadísticos de los textos analizados:

	Palabra	No. De Discursos en los que aparece	No. de Repeticiones totales	Score asignado
0	PEOPLE	44	721	31724
1	US	44	614	27016
2	MUST	34	496	16864
3	ONE	29	310	8990
4	WOULD	29	480	13920
5	TIME	25	274	6850
6	WORLD	24	311	7464
7	YEARS	23	204	4692
8	AMERICAN	22	254	5588
9	GREAT	21	195	4095
10	NATION	20	229	4580
11	GOVERNMENT	20	280	5600
12	AMERICA	20	222	4440
13	EVERY	19	162	3078
14	MAN	19	434	8246
15	KNOW	19	245	4655
16	SAY	18	241	4338
17	PEACE	18	223	4014
18	MEN	17	186	3162
19	COUNTRY	17	268	4556

En la siguiente gráfica se puede apreciar que la mayoría de las palabras comunes solo estuvieron entre 20 y 30.



Para la clasificación con k-Means se obtivieron 5 clusters de datos correspondientes con la popularidad del discurso, mismos que se puden observar claramente en la gráfica siguiente:



6.2. Conclusiones

Fue posible clasificar los discursos según su popularidad, con la combinación de análisis de textos y análisis estadísticos, así como la transformación de variables discretas a numéricas.

7. CONCLUSIONES

Resumen: *En esta sección se presentan las conclusiones a las que se llegó después de la realización del trabajo.*

7.1. Conclusiones

Durante la realización del trabajo y el análisis de los resultados se destacó la diferencia entre técnicas de análisis y clusterización así como la importancia de usar las herramientas correctas ya que no todas pueden llegar a resultados esperados.

Si bien, el fracaso en encontrar los resultados deseados puede llevar a la conclusión de que el objetivo no se cumplió, puede también ser un indicador que las herramientas utilizadas no fueron las más adecuadas, por lo que es importante explorar más soluciones.

Como conclusión del tema como tal, dado que los datos son en su mayoría discursos políticos de Estados Unidos durante los tiempos de las guerras mundiales y conflictos globales, es de esperar que los resultados estén orientados a calificar de mejor manera el uso de palabras patrióticas tales como *Nation, America o American, Men, Country, Peace, etc.* ya que en este contexto social americano, el patriotismo y nacionalismo eran muy importantes para sociedad.

Se encuentran a los discursos menos populares a aquellos que tienen un contexto distinto, tales como aceptación de premios nobel o resignación de presidentes.

7.2. Trabajo a Futuro

Puede resultar de especial interés la elaboración de una herramienta que utilice el modelo para analizar textos u clasificarlos automáticamente, sin necesidad de procesar los datos manualmente, así como la exploración de nuevas técnicas tales como redes neuronales, que puedan dar resultados mejores, o mínimo más interesantes.

REFERENCIAS BIBLIOGRÁFICAS

[1] “El Algoritmo K-means aplicado a clasificación y procesamiento de imágenes¶,” kmeans, https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html (accessed May 11, 2023).

[2] “¿Qué es el algoritmo de k Vecinos Más cercanos?,” IBM, https://www.ibm.com/mx-es/topics/knn (accessed May 11, 2023).