Procesamiento del Lenguaje Natural: Análisis de Sentimientos

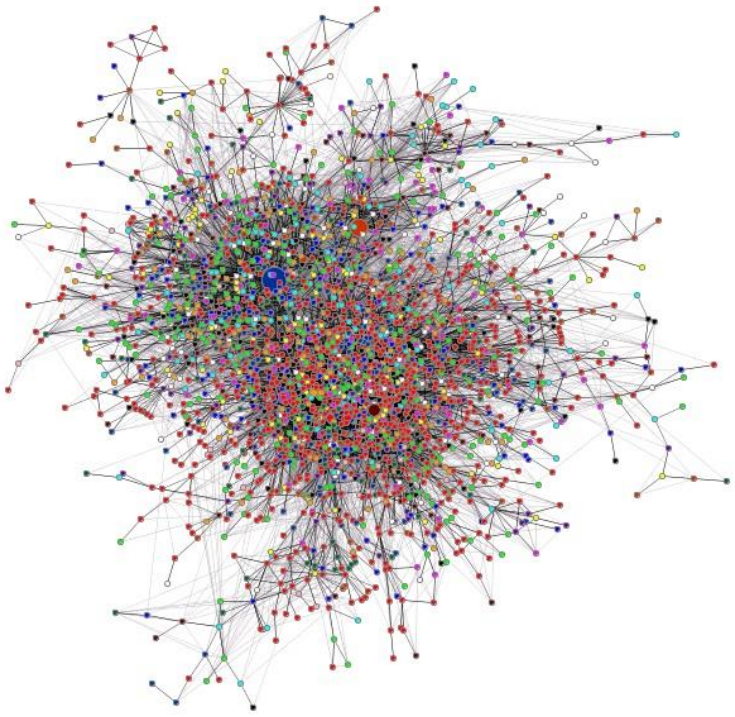
Venegas Guerrero Fátima Alejandra
León Canto Ángel Efraín
Vázquez García Palemón
Hernández Chacón Carlos Alberto

1.INTRODUCCION

El análisis de sentimientos (también conocido como minería de opiniones), es el proceso de determinar el tono emocional que hay detrás de una seria de palabras y se utiliza para intentar entender las actitudes, opiniones y emociones expresadas en un texto.
Esta tarea ha adquirido gran interés por su potencialidad de aplicación y por la gran cantidad de opiniones no estructuradas que se encuentran disponibles en distintos medios sociales. Por ejemplo, resulta interesante para las empresas conocer la opinión de sus clientes sobre productos; para los clientes conocer la opinión de otros compradores, para la sociedad conocer la opinión pública sobre alguna situación, etc.
Es extremadamente útil en la monitorización de minería de datos, opiniones del consumidor, tendencias en redes sociales, etc.

1.1. OBJETIVO

El objetivo principal de este proyecto es el análisis de textos a través del procesamiento computacional, utilizando el algoritmo de agrupamiento conocido como Clustering Jerárquico, detallando paso a paso su procedimiento.
Y simulándolo a través de una pequeña aplicación.
A su vez se darán algunas observaciones obtenidas, acerca de los diferentes métodos utilizados.



2. DESARROLO DE LA HERRAMIENTA DE ANALISIS DE SENTIMIENTOS

2.1. Procesamiento del Texto

La utilización de técnicas de procesamiento de lenguaje natural para la clasificación de textos consiste principalmente en encontrar

patrones y características del lenguaje que permitan asignar una categoría, etiqueta o clase a un documento.
Para esto pueden requerirse algunas transformaciones en el corpus de datos de entrada antes de su procedimiento
Algunas transformaciones realizadas se enlistan a continuación:

- Tokenizacion:

Separación de sentencias y palabras de un documento a partir de *tokens,* o caracteres especiales, que indican el fin de una sentencia o palabra y el comienzo e las que sigue.

- Steamming:

Existen algoritmos llamados Steamming (o Stemmers) que permiten obtener la raíz o stem de una palabra eliminando terminaciones, con el objetivo de unificar aquellos términos que aportan la misma información al clasificador. Por ejemplo, los términos "recomendable, recomendamos, recomendación" son remplazadas por su stem "recomendar".

- Stopwords:

Stopwords o también llamadas palabras vacías es el nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc.

- Conversión de mayúsculas a minúsculas

2.2. Creación de la matriz de Procesos

Para poder analizar la información de los textos debemos estructurar la información contenida. Para ello, convertiremos la columna de

texto en una matriz en la que cada palabra es una columna cuyo valor es el número de veces que dicha palabra aparece.

Por ejemplo:

```
¡Rosa no presentida que quitara
a la rosa la rosa que le diera
a la rosa le diera la rosa!
```

Convertida en matriz:

- **Contadores**: El número de veces que la palabra aparece en la frase.

diera	la	le	no	presentida	que	quitara	rosa
0	0	0	1	1	1	1	1
1	2	1	0	0	1	0	2
1	2	1	0	0	0	0	2

Así podremos trabajar con estos vectores en vez de con texto plano.

2.3 Calculo de la Similitud

2.3.1 Calculo de la Distancia

El siguiente paso sería el cálculo de la distancia similitud, existen varios métodos para obtenerla, nosotros elegimos la distancia Euclidea a continuación se explica su funcionamiento:

Distancia Euclidea:

"Nos dice que el punto la distancia más corta es la línea recta que une dos puntos."

- Formula:

$$d_{ij} = \sqrt{\sum_{p=i}^{k} (x_{ip} - x_{jp})^2}$$

-)Ejemplo de su uso:

A. Teniendo la matriz de datos

Objetos/individuos	X_1	X_2
A	1	1
B	2	1
C	4	5
D	7	7
E	5	7

Matriz de pesos/datos

B. Aplicamos la formula

Caso particular para dos variables:

$$d(P_1,P_2)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

Tomando el objeto A y B de la matriz tenemos:

$$d(A,B)=\sqrt{(2-1)^2+(1-1)^2}=1$$

Calculando con cada uno de los objetos, obtenemos una matriz de distancias o similitudes:

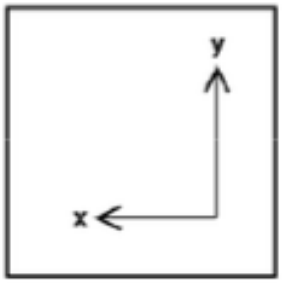
	A	B	C	D	E
A	0				
B	1	0			
C	5	4.5	0		
D	8.5	7.8	3.6	0	
E	7.2	6.7	2.2	2	0

Matriz de Similitudes

Si analizamos un poco la tabla nos daremos cuenta que los objetos A y B son más similares entre si ya que solo están a una distancia o unidad del uno del otro así que, se estos se fusionarán y construirán otro punto, así se deberá aplicar la formula sucesivamente hasta ya no poder fusionar más puntos.

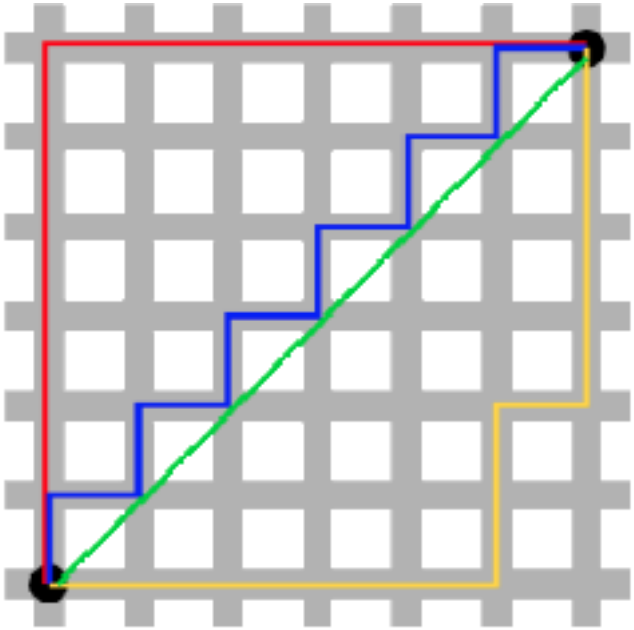
Existen varios métodos para poder calcular las distancias, aquí si da una pequeña muestra de estas:

Distancia Minkowski (Manhattan):



Manhattan

“Nos dice que la distancia se mide por el mínimo número de calles que se recorren “



Manhattan: Líneas azul, rojo y amarilla.
Miden=12

Euclidea: Línea verde
Mide=8.5

Problema del Taxicab

Formula:

$$d_1(x, y) = \sum_{j=1}^{J} |x_j - y_j|$$

Distancia Chebyshev:

También conocida como distancia de tablero de ajedrez , esta mide el número de movimientos que el rey ha de hacer para llegar de una casilla a otra en un tablero.

Formula:

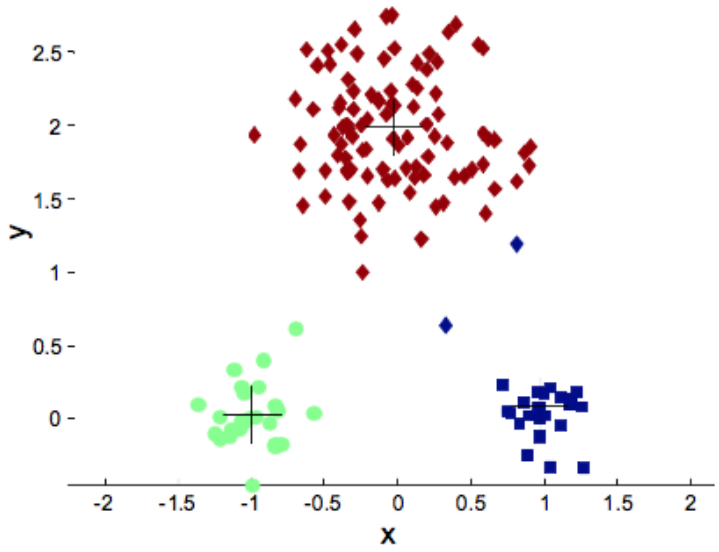
$$d_\infty(x, y) = \max_{j=1..J} |x_j - y_j|$$

Al final se dan algunas observaciones más sobre la distancia Euclidea utilizada en este proyecto
(sección 3.1.1)

2.4 Aplicación del Método de Agrupamiento

Cluster

El objetivo de un clúster es encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre si y diferentes de los objetos de otros grupos .



Clúster

Existen diversos métodos de agrupamientos como:

I. K-Means: Es un algoritmo de agrupamiento por particiones, cada clúster tiene asociado un centroide. Los puntos se asignan al clúster cuyo centroide este más cerca. Iterativamente, se van actualizando los centroides.

II. DBSCAN: Es un método baso en densidad, donde un clúster en una región densa de puntos, separada por regiones poco

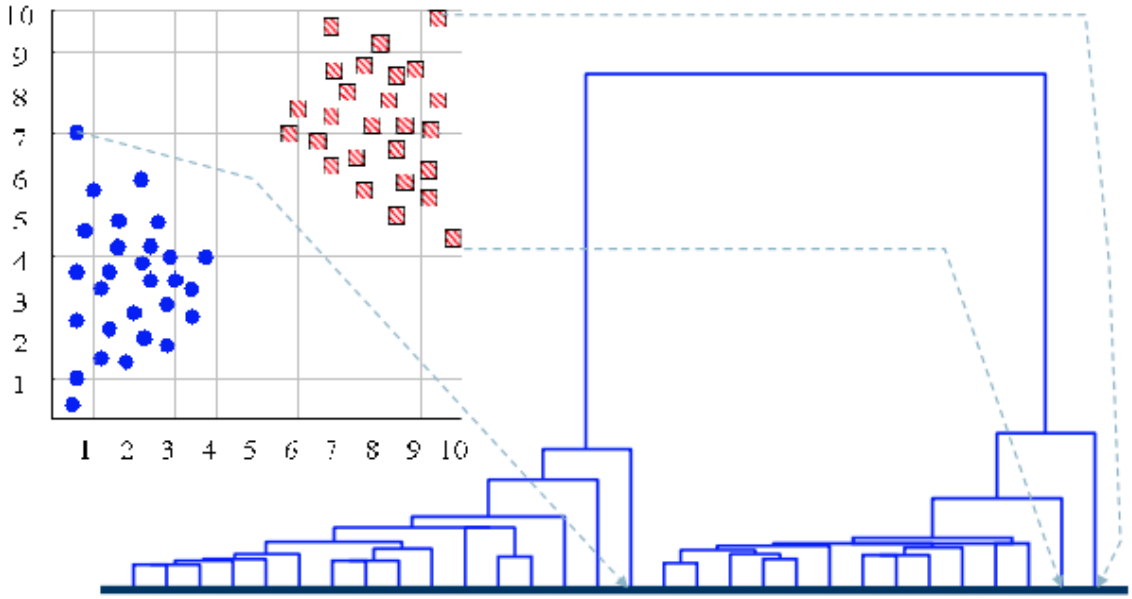
densas. Son muy útiles cuando los clúster tienen formas irregulares.

Clustering por DBSCAN

Nosotros aplicamos el método de agrupamiento por clustering jerárquico.

III. Clustering Jerárquico:

El clustering jerárquico se caracteriza por representar cada clúster en un diagrama llamado dendograma, este nos puede ayudar a determinar el número adecuado de agrupamientos (aunque normalmente no será tan fácil).



La similitud entre dos objetos viene dada por la “altura” del nodo común mas cercano.

Al final se detalla mas acerca de este clustering y su uso en el proyecto (sección 3.1.1)

3.0 Aplicación

Se realizó una pequeña aplicación para simular el análisis de sentimientos.
Detalles:

- ❖ Esta fue programada en Python
- ❖ Se utilizaron algunas librerías como:
  - Pickle – interfaz grafica
  - Nltk - procesamiento del texto
  - Matplotlib – graficacion
  - Numpy - manejo de matrices
  - Scipy.cluster.hierarchy – graficacion de dendogramas

3.1 Observaciones y Resultados

Los resultados obtenidos a la hora de ejecución, fueron para nuestro propósito de análisis de sentimientos, inconcluso ya que tuvimos una falta de entrenamiento de nuestro aplicación, que tenia que ser obligatoria para su funcionamiento, esta era dar una vasta base datos o corpus, por el mismo, los datos arrogados no fueron nada demostrativos.
Con lo que se refiere a los algoritmos hubo un buen trabajo en la programación del algoritmo de clustering jerárquico y por ende un buen resultado en su graficacion.

3.1.1 Observaciones y Conclusiones de los diferentes métodos

- Distancias:
  Como se mencionó con anterioridad se utilizó la distancia Euclidea,, por la facilidad de explicarla, y mostrar algunos ejemplos en clase que no se llevaran mucho tiempo, mas cabe destacar que esta , se nos hizo ver que esta no es la más adecuada a la hora de calcular similitudes en objetos tan complejos como en nuestro caso lo hacemos en las palabras, o texto ya que esta ese queda muy en la “superficie”, es decir si

tenemos una naranja y un plátano y a estos le aplicamos la distancia Euclidea, nos dirá que efectivamente que son similares, por qué ambos son frutas, mas realmente no está tomado otros datos como: la textura, el tamaño, el sabor, el color, la forma, de que temporada son, de que país provienen ,etc.

Tomando esto en cuenta, al final se dará un anexo, mostrando una comparación de los resultados obtenidos de la aplicación, pero utilizando otras distancias.

- Clustering Jerárquico:

  Se eligió el clustering jerárquico, ya que la mayoría de los algoritmos ya se habían elegidos, mas también se quería , conocer su funcionamiento y si podría ser utilizado para nuestro proyecto de análisis de textos.
  Al final nos dimos cuenta que no es el mas indicado para este tipo de análisis, ya que nos dimos cuenta que este tiene una baja escalabilidad , es decir no tiene una habilidad para reaccionar y adaptarse a la hora de que los datos se hacen mas grandes y mas complejos ya que pierde bastante su calidad.

4.0 Trabajando

A la falta de entrenamiento como se menciono, se está trabajando en mejorar esa parte de la aplicación añadiendo un corpus o base de datos.

Pronto se subirá el proyecto mejorado, así como los anexos mencionados .

5.0 Bibliografía

1. http://pdln.blogspot.mx/2014/01/analisis-de-sentimientos-un-algoritmo.html
2. http://materias.fi.uba.ar/7500/Dubiau.pdf
3. http://isa.umh.es/asignaturas/crss/turorialWEKA.pdf
4. http://materias.fi.uba.ar/7500/Dubiau.pdf
5. https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf?sequence=1&isAllowed=y
6. https://www.python.org
7. https://dlegorreta.wordpress.com/category/de-lo-concreto-a-lo-abstracto/nlp-con-python-y-r-project/
8. https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/#Perform-the-Hierarchical-Clustering
9. https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.pdist.htm
10. https://es.wikipedia.org/wiki/Distancia_de_Chebyshov