

# Procesamiento del Lenguaje Natural: Análisis de Sentimientos

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

## 1. Anexo

En este anexo, se presentarán, algunos resultados obtenidos con la aplicación creada para nuestro proyecto, así como su explicación de cada una de ellas, observaciones y conclusiones relevantes.

### 1.1 Resultados.

Se tomó una base de datos, con más de 11,000 palabras en inglés cada una de ellas con su polaridad, negativa o positiva, como parte de su enteramiento.

Después se tomó un fragmento de una base de datos sobre opiniones de algunos consumidores sobre productos electrónicos. Teniendo como sintaxis la siguiente:

<Pros>Picture Quality, Manual Controls, Battery Life</Pros>  
<Pros>Photo color & clarity, flip-out LCD screen, remote included</Pros>  
<Pros>Lots of features, Great battery, great outdoor pictures.</Pros>  
<Pros>Good feature list, RAW mode, flip-out LCD etc</Pros>  
<Pros>Easy point and shoot, GREAT color and resolution. Very good battery life.</Pros>  
<Pros>Great quality pics, even in low light</Pros>  
<Pros>Image quality, battery life, hot-shoe for ext. flash</Pros>  
<Pros>Multiplicity of modes and features, flexibility, DURABILITY!, MOVEABLE LCD-SCREEN! GREAT PICTURE QUALITY</Pros>  
<Pros>Great images, easy to use, many features.</Pros>  
<Pros>great camera</Pros>

En base a esta gran base de datos, se le realizaron los procesos de textos, adecuados (tokenizar, transformación de mayúsculas a minúsculas, etc.)

Después se procedió a aplicar el algoritmo de agrupación en este caso, clustering jerárquico, más se realizaron pruebas con distintas distancias, estos resultados se presentan a continuación:

- Distancia Euclidea:

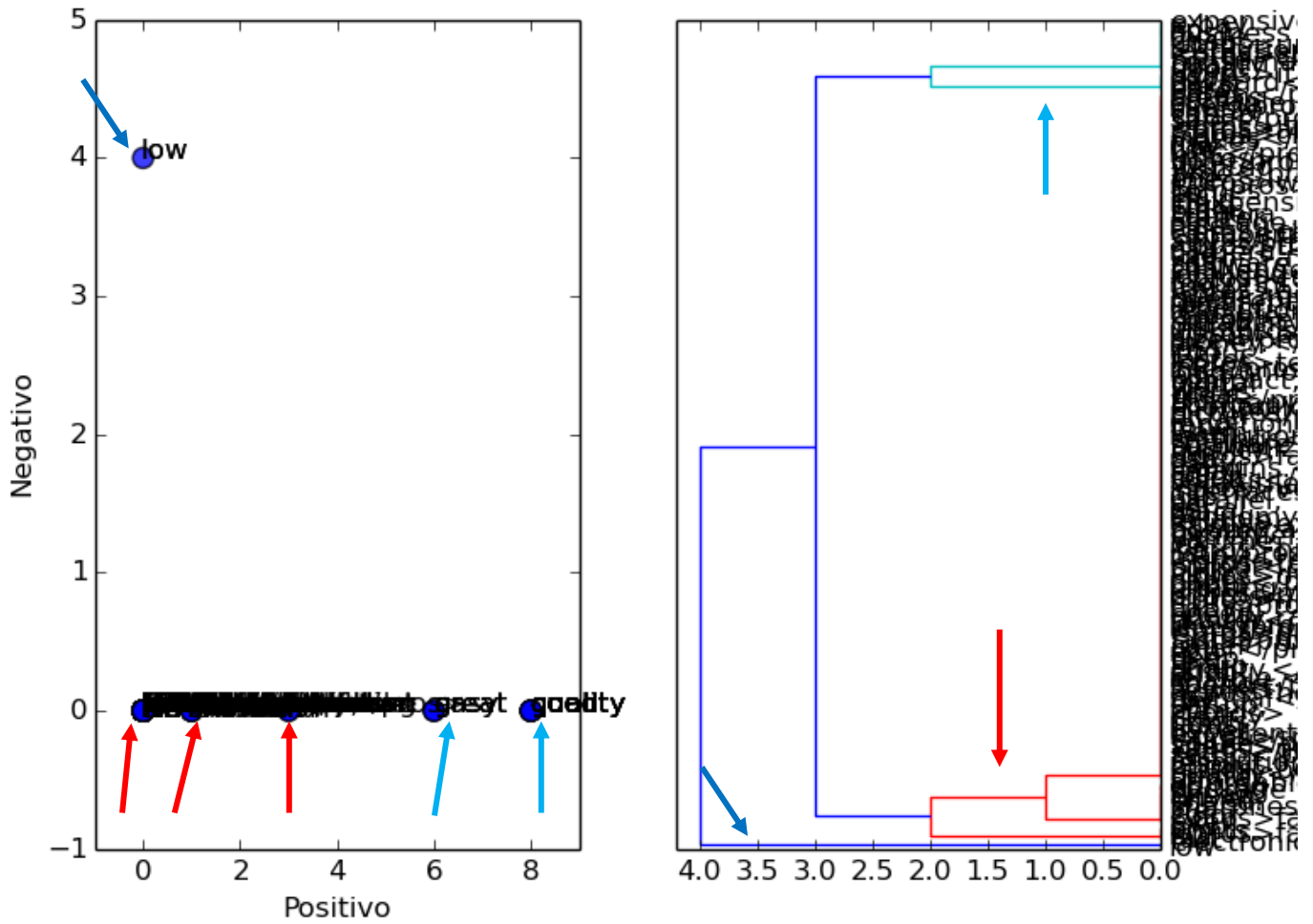




Figura1.Clustering Jerárquico, con distancia Euclidea

-  Palabras Negativas
-  Palabras Neutras
-  Palabras Positivas

En esta primera figura, de lado izquierdo se puede ver la graficacion de los tres clústers, cada uno señalado por flechas de colores, la flecha de color azul oscuro representa el clúster con la(s) palabras negativas, en la parte de abajo, representado por las dos flechas de color azul claro, tenemos los clusters que representan las palabras positivas, mas por otro lado con las tres flechas rojas están representadas las palabras que caen en el caso de “neutras” es decir aquellas palabras que no están en ninguna base de datos, como por ejemplo: photo , color, screen , etc.

Del lado derecho tenemos la graficacion de los clusters, por un dendograma, representado las palabras neutras las líneas de color rojo , las palabras positivas las líneas de azul claro, y la palabra(s) negativas de color azul oscuro pegado a las líneas rojas. Recordemos que un dendograma se interpreta, como: “la similitud de dos objetos es la ”altura” entre los nodos “.

Así que como vemos la altura entre los nodos de las palabras positivas (azul claro) y las palabras neutras(rojo) es similar, más en la altura con la línea azul oscuro que son las palabras negativas, es realmente más alta y totalmente alejada de los demás clusters.

Tomando estas pequeñas observaciones, podemos **sugerir** que el texto que se está analizando tiene una polaridad positiva, ya que solo identifico una palabra negativa, en contra de las grandes cantidades de palabras positivas que no se muestran del todo.

- Distancia Manhattan:

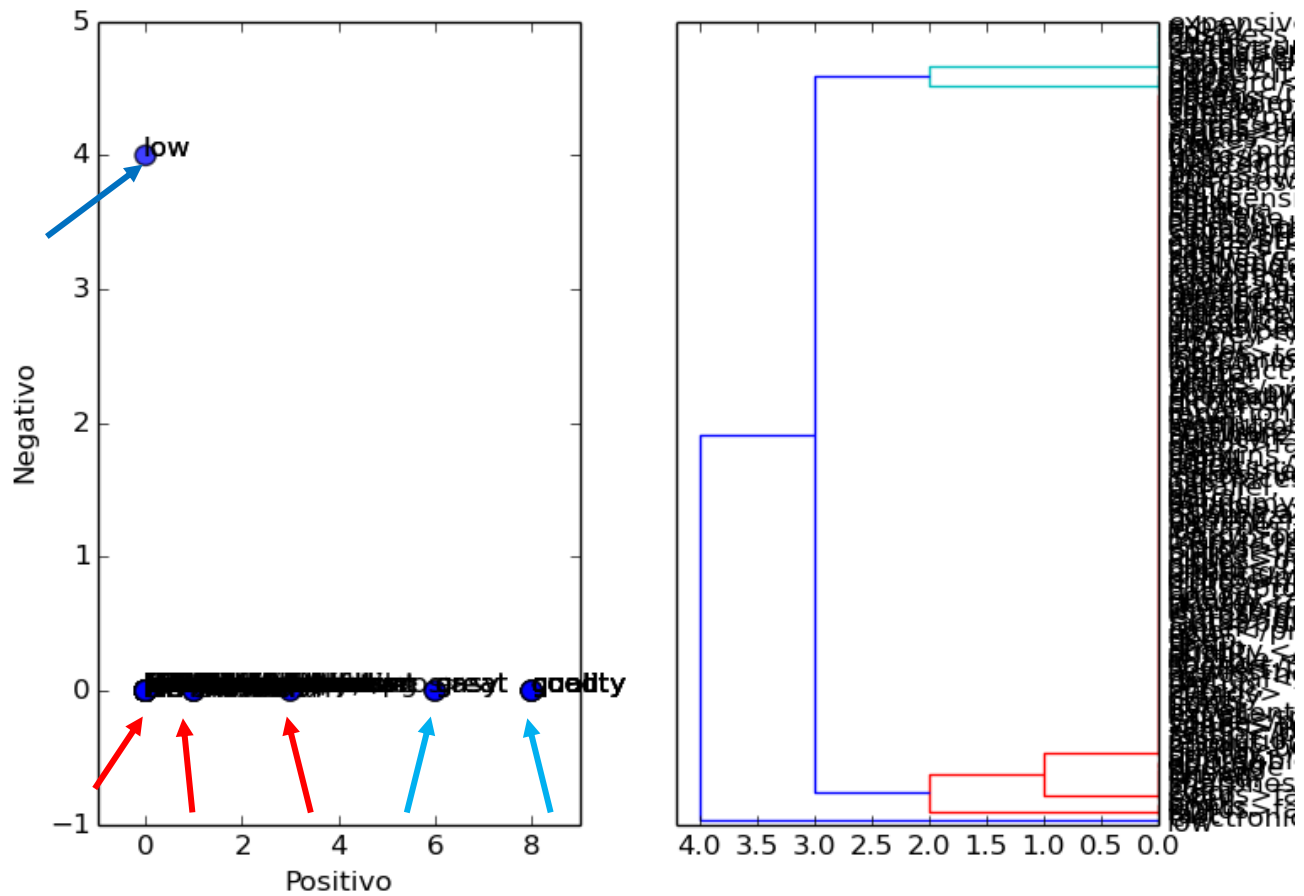


Figura2.Clustering Jerárquico, con distancia Manhattan

-  Palabras Negativas
-  Palabras Neutras
-  Palabras Positivas

○ Distancia Chebyshev:

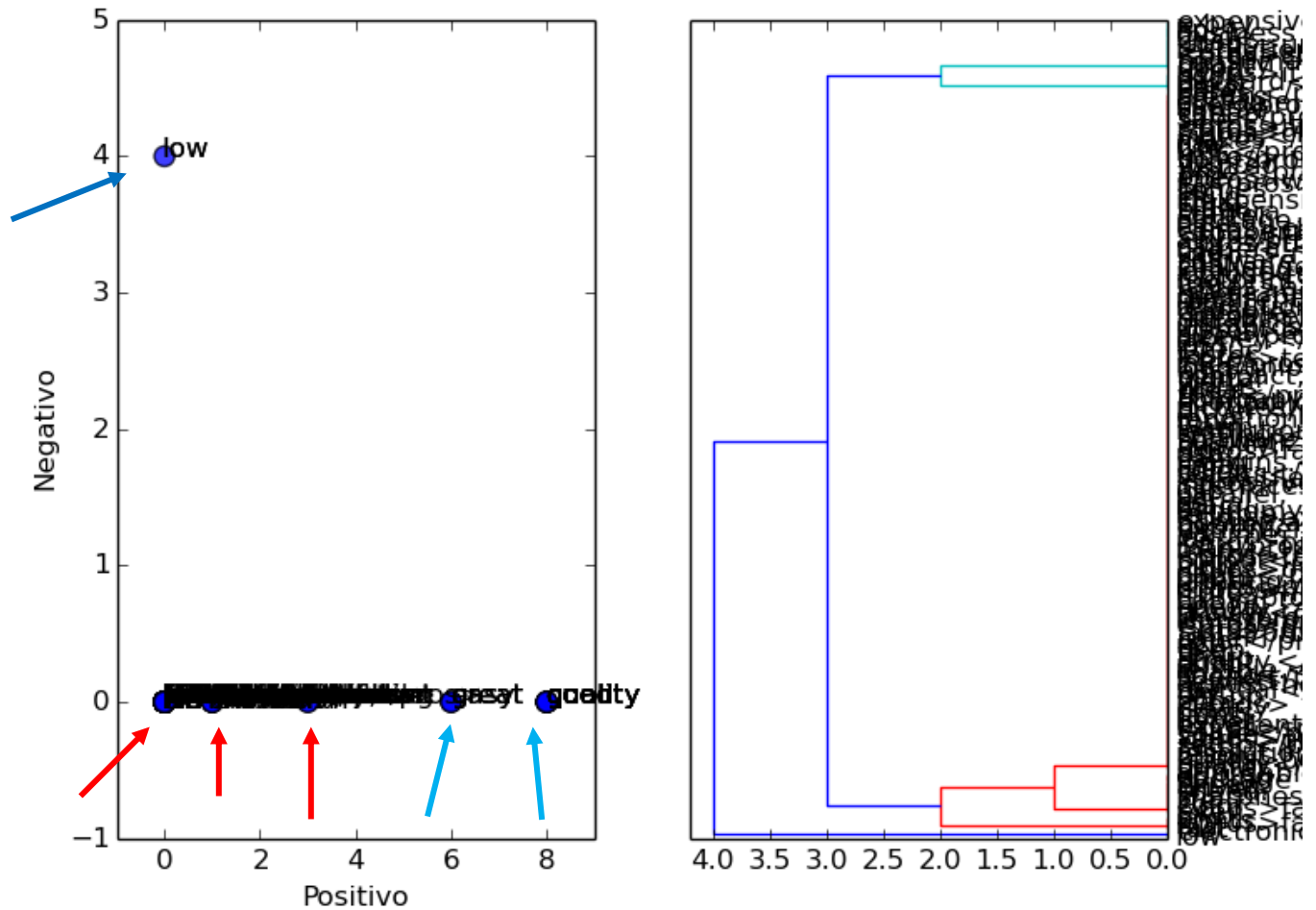


Figura3.Clutering Jerárquico, con distancia Chebyshev

Como se puede observar, en las figuras 2 y 3 no existe una gran diferencia de los clústers y su graficación en el dendrograma, con la distancia Euclidea, esto nos deja ver la gran imprecisión que existe a la hora de calcular la similitud entre objetos en este caso palabras que son demasiado imprecisos

Realizamos diferentes pruebas aquí presentamos una mas que hace ver algo importante.

Ahora procesamos un pequeño cuento.

Titulado: "The arrogant Swans", aquí un fragmento:

"In a far away kingdom, there was a river. This river was home to many golden swans. The swans spent most of their time on the banks of the river. Every six months, the swans would leave a golden feather as a fee for using the lake. The soldiers of the kingdom would collect the feathers and deposit them in the royal treasury."

Los Resultados fueron:

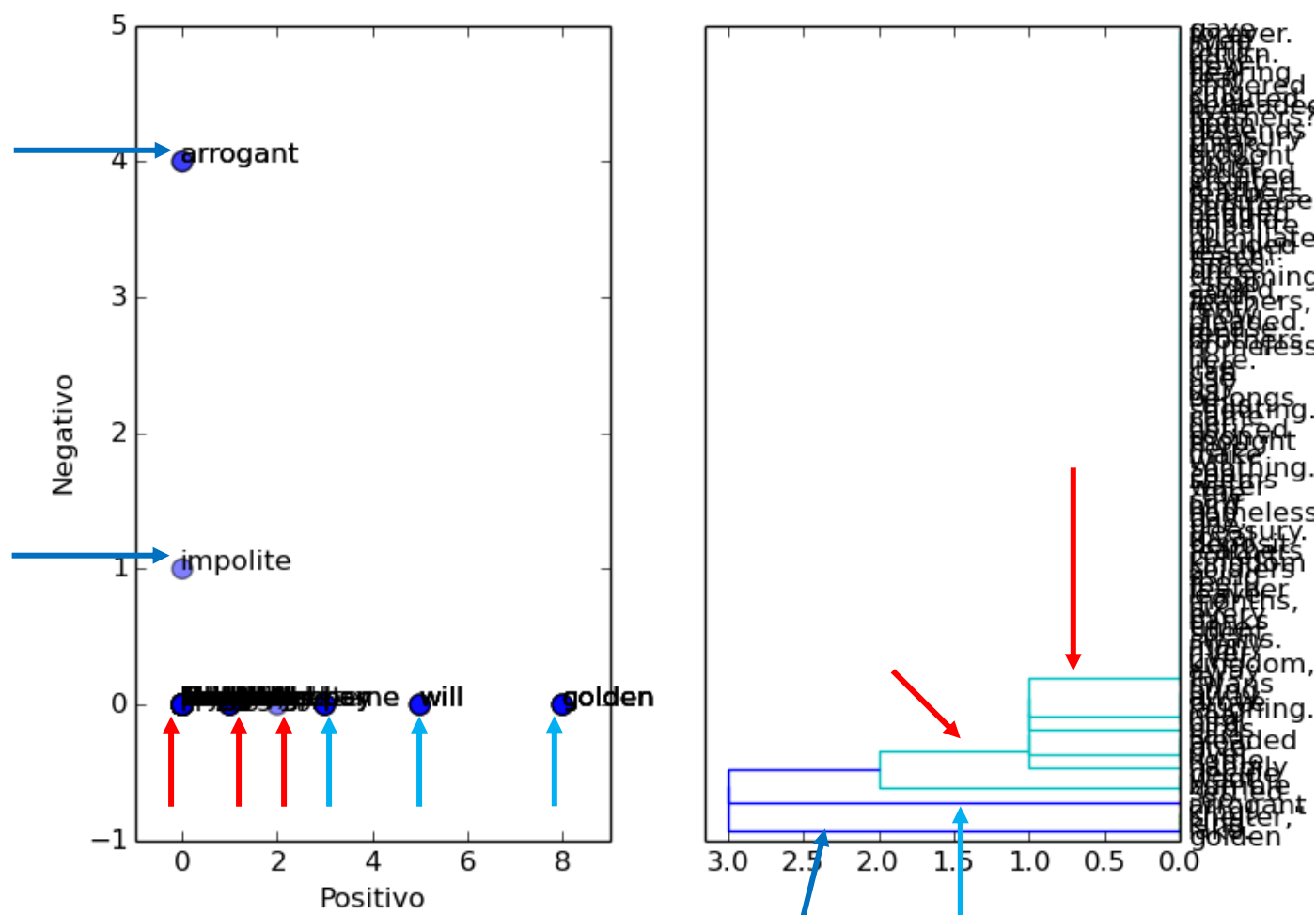
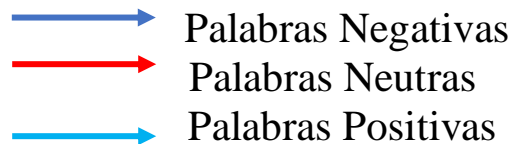


Figura4.Clustering Jerárquico, con distancia Euclidea



Como podemos observar, en la parte izquierda de los cluters, se grafica claramente dos palabras (impolite y arrogant) representadas por las flechas de color azul oscuro, que tienen la connotación negativa. También vemos tres clusters señalados con las flechas de color azul claro entre las palabras que se pueden ver están: “golden”, ” will “(en este caso pusimos a will como positivo en nuestra base)

Y por último tenemos las palabras neutras señaladas con flechas rojas.

Mas si nos fijamos ahora en la parte derecha del dendograma solo podemos ver que nos grafica dos, de diferente color , alguna de las hipótesis del porque esto sucede , es por el cálculo de la similitudes, como podemos ver la palabra “arrogant” e “impolite” están cerca, uno del otro así que su similitud es pequeña, y las grafica como la línea azul oscuro más cercana al eje de los números, por otro lado tenemos las palabras “golden” y “will” que son palabras positivas en el dendograma las grafica como la línea oscura de en medio donde tienen una gran similitud, y por ultimo las palabras “neutras” las líneas de color azul cielo donde dentro de estas podemos ver que algunos de los clúster son más cercanos entre ellos, que otros, y sus alturas entre ellos también teniendo pequeños sub-grupos de clusters.

Aquí sería difícil decir que polaridad tiene el texto por la falta de precisión de las gráficas. Mas podemos sugerir que solo al tener dos cluters o palabras negativas, tiene una polaridad positiva.

También calculamos este ejemplo con diferentes distancias mas no obtuvimos alguna diferencia relevante como en las figuras 2 y 3 del caso anterior.



## 2. Conclusiones

Viendo los resultados obtenidos, podemos deducir algunas conclusiones:

- Existe una gran falta de precisión a la hora de calcular las distancias, entre los pesos de las palabras, así que se debe buscar algún método más eficiente para poder trabajar con las diferentes características y complejidades de las palabras, el cual tome no solo “superficialmente” un objeto para deducir si son similares. Si no pasara lo que presentamos en las figuras 2 y 3 ,donde no existe alguna diferencia entre las distancias usadas
- Con respecto al uso del método de Clustering Jerárquico, no es el adecuado, para este tipo de análisis, ya que no se adapta a la gran base de datos y su complejidad perdiendo su calidad y dando resultados muy inexactos.
- El procesamiento del texto es muy irregular también, este debe de pasar por un gran número de pasos para ser “limpiado” y ser usado por el algoritmo dando algunos inconvenientes, como podría ser caracteres que no reconoce como emoticonos, o separar algunas palabras como: “don’t give up” donde la frase tiene una connotación positiva, mas al procesarla y separar cada una de sus palabras nos arrojaría que don’t es negativa, give positiva y up neutra, lo cual ya no representa su polaridad.

También algunas palabras su significado puede variar según el contexto por ejemplo will, esta palabra por si sola nos da referencia a decisiones que podemos hacer en el futuro, en nuestro analisis seria llevada a las palabras positiva , mas si se tiene : “After

six months in hospital she began to lose the will to live”, aquí will precedido por el lose tendría una conotación negativa ya que significa que “a perdido la voluntad de

vivir”, mas al ser procesado el texto y ser tokenizado will iría a palabras positivas y el resultado del análisis no sería preciso.

## 2.2. Notas

Algunas cosas que nos sucedió a la hora de procesar una gran base de datos es que el programa simplemente paro. Realizamos varias pruebas, mas aquí solo presentamos algunas que nos parecieron más relevantes

## 3. Bibliografía:

Aquí las paginas donde fueron descargadas las bases de datos y el cuento

- <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>
- <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>
- <http://www.english-for-students.com/The-Arrogant-Swans.html>

Se anexa también el link de drive donde se puede descargar el programa, así como las diferentes bases de datos usadas, y otras encontradas si se quiere probar

- <https://drive.google.com/drive/folders/0B4k63jnwksNNR1VKcXNmUI9WQ2c>

