Análisis de Presentaciones Web

Manuela Jaramillo Rendón
Proyecto final Datos y Visualización
UdeM

1 Acerca del Proyecto

Las presentaciones web muestran colecciones de material textual del cual se puede inferir mucha información,como temas y conceptos clave,relaciones ocultas y tendencias existentes sin necesidad de conocer las palabras o términos exactos que han utilizado los autores para expresar dichos conceptos.

En este proyecto se unen las herramientas necesarias para realizar un análisis estadístico básico de la información contenida en estas presentaciones y a partir de esto generar representaciones gráficas,utilizando para este fin python y otros paquetes libres.

En este caso se realizará un análisis a una presentación de slideshare títulada "La vida media del ADN impedirá la colnación de dinosaurios".

2 Funte de Datos

Como se hizo mención anteriormente la fuente de datos proviende de una presentación de Slideshare; con 26 diapositivas,para realizar esta presentación se tomo como base un artículo científico.

La presentación se puede encontrar en: *https://es.slideshare.net/manunicol10111998/la-vida-media-del-adn-impedir-la-clonacin-de-dinosaurios/1*

Artículo base: *https://blogthinkbig.com/la-vida-media-del-adn-impedira-la-clonacion-de-dinosaurios*

Aunque en este caso se realizó un ejemplo específico el código que surge del proyecto esta diseñado para el análisis de cualquier diapositiva en slideshare.

3 Análisis estadístico y productos gráficos

A partir del análisis del texto de estas diapositivas se pueden obtener frecuencias; para asi poder identificar posibles patrones en algunos contextos,además se puede porcentualizar los sentimientos manifestados en la temática para determinar la polaridad textual y a partir de esto generar gráficos agradables.

3.1 Frecuencias

La frecuencia de las palabras es un buen indicador del tema o del sentimiento expresado en los textos. En ocaciones las palabras con aparición más frecuente son los "Stop-words" o denominadas palabras vacias; ellas no nos brindan indicaciones de los patrones que buscamos,así que se eliminan en su mayoria.

Al tomar la diapositiva "La vida media del ADN impedirá la colnación de dinosaurios",obtener el número de repeticiones de las 10 palabras más frecuentes y dividirlas por el número total de palabras se obtuvo la frecuencia,esta posteriormente se paso a porcentaje y se represento graficamente obteniendo :

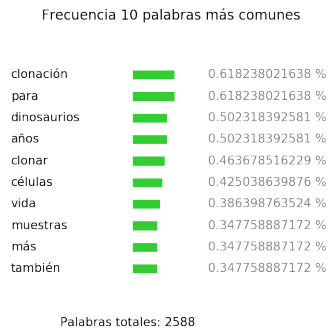

Figure 1:

Como se puede obvservar en la Figura 1 algunas de palabras más comunes son vacias,asi que no se toman en cuenta para inferir temáticas,otras palabras que se pueden tomar como claves son más dicientes,asi pues observamos que las palabras proporcionan información acerca del contenido de las diapositivas; aunque no leamos siquiera la diapositiva podemos saber los temas primordiales.

También resulta interesante observar la frecuencia de cada elemento de vocabulario en el texto,de una manera más general.NLTK nos brinda la distribución del número total de tokens de palabras en el texto através de los elementos de vocabulario mediante FreqDist.

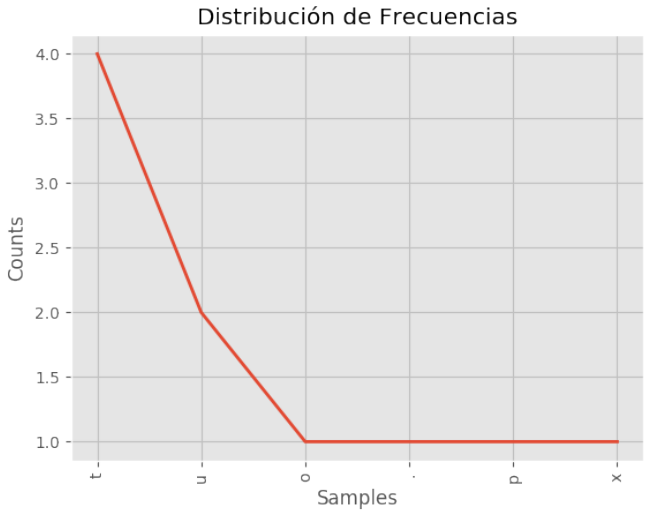

Figure 2:

Estas frecuencias son no acumuladas,FreqDist devuelve la frecuencia de casi cualquier elemento en el texto que presente repeticiones,incluyendo las stop-words.

3.2 Análisis de Sentimientos

Realizando este tipo de estudio se puede determinar la polaridad general de un documento, una frase o un aspecto del mismo, tratando de detectar la actitud del creador en base a las posibles emociones, juicios o evaluaciones contenidas en el documento. Las etiquetas más extendidas para clasificar la

polaridad son: positiva, negativa o neutra.

El resultado analisis de sentimientos para diapositiva "La vida media del ADN impedirá la colnación de dinosaurios" fue:

Etiqueta: Neutra
Positivo: 0.5
Negativo: 0.5

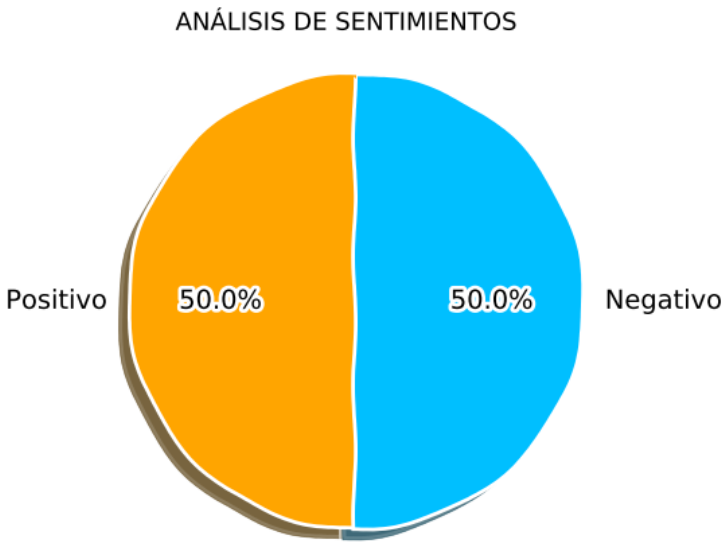

Figure 3:

El texto analizado no presento polaridad,se da un valence entre lo negativo y lo positivo.El tipo de texto analizado fue un artículo científico neutral.

Para estudiar un poco más a fondo la polaridad,se pueden tomar párrafos en particular,por ejemplo el pírrafo 8:

El párrafo 8 por su parte se etiqueto como negativo.

Etiqueta: Negativo

Positivo: 0.114
Negativo: 0.886

Figure 4:

4 Word Cloud

Generalmente usan nubes de palabras para producir fácilmente un resumen de documentos.

5 Más acerca del proyecto

Para conocer más acerca de proyecto diríjase a la documuntación,tanto a la documentación del proyecto como la del código.


forma
investigadores
especie
condiciones
ADN
hace
ciencia
vlda
vida
organismo
hueso
clonaci
equipo
mama
ADN
clonauun
posible
millones
enla
lula
muestras
vlda mama
clonar dinosaurios
gen
dinosaurio
clonar
dela
Dmusaunus
an
ahora
siles
conservaci
vida media
restos
aunque
secuencia

Figure 5: