**Corporación Universitaria Minuto de Dios**

**Asignatura:**

Minería de datos

**Taller minería de texto**

Trump Discursos

**Presenta:**

Adrian Yesith Hoyos Márquez

**NRC:**

13816

**Docente:**

Jonathan Freddy Narváez

Colombia Bogotá D.C.                                                                         12 de octubre 2020

Contenido

[Introducción 3](#_Toc53404258)

[Minería de texto 4](#_Toc53404259)

Introducción

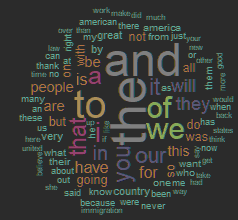
En este taller se dará análisis acerca de algunos los Discursos de Trump usando como herramienta el lenguaje de programación R para facilitar la minería de texto. En esta ocasión se dio uso de la herramienta anteriormente dicha iniciando por la lectura de los correspondientes archivos txt y terminando por el stemming donde se pintan las palabras más usadas de Trump con la finalidad de deducir y analizar lo que quiere decir.

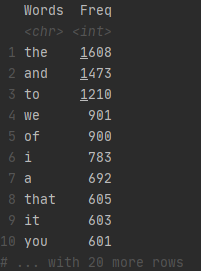
Minería de texto

Para la correspondiente minería de texto del presente taller, como bien se dijo introductoriamente, se empezó por leer los archivos txt correspondientes a algunos de los discursos del presidente de los estados unidos llamado Donal Trump, seguido de esto pasándolo por una fase de transformación para facilitar la manipulación de estas palabras específicamente en un vector.

Luego de esto, se hizo la correspondiente limpieza de los datos donde se eliminaron los números, los apostrofes se tuvieron en cuenta para eliminar dado que la expresión regular [[:punct:]] lo que hacia era eliminar los apostrofos pero no eliminar lo que seguía después de esta, entonces se optó por tomar una expresión regular manual teniendo en cuenta que hay palabras como por ejemplo don’t, weren’t entre otras que dan sentido negativo a las palabras en el idioma ingles y dejarlas de tal manera que dijeran do y were, y de esta manera dar cierto tipo de normalización a las palabras para posteriormente leerlas de manera adecuada. Luego de esto, se usó la expresión regular [[:punct:]] para eliminar todo tipo de signos de puntuación ya que esto no nos dice nada para su correspondiente análisis y además R lo toma como si fuese otra palabra cuando en realidad no lo es, seguido de esto se eliminaron los espacios, cadenas vacías y caracteres especiales.

Luego de esto, se tomaron las frecuencias de las palabras para poder encontrar stopwords que nos generan ruido, palabras que se repetían muchas veces pero no nos decían nada como se muestra a continuación:





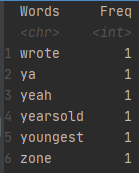
A lo cual se deduce que estas palabras deben ser eliminadas debido a que nos desvía de lo que realmente se quiere analizar, ya que estos stopwrods se repiten demasiadas veces pero no tiene significado alguno para el correspondiente análisis.

Posterior a esto, se procede a eliminar los stopwords donde investigando por internet se llego a que existe una función llamada stopwords() la cual contiene un vector de estas muy grande  a lo cual nos facilita en gran manera tener que descargar vectores externos o archivos para dar cumplimiento a la eliminación de dichos stopwords y la eliminación de lenguaje soez ingles, en donde pasamos de tener una nube de palabras con muchos stopwords pasamos a la siguiente:

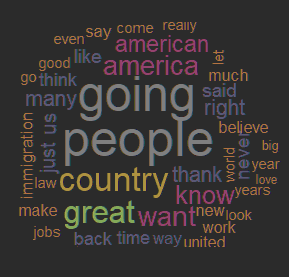


Donde se observa en gran manera que se reduce demasiado las palabras que Donald Trump menciona en los discursos analizados.

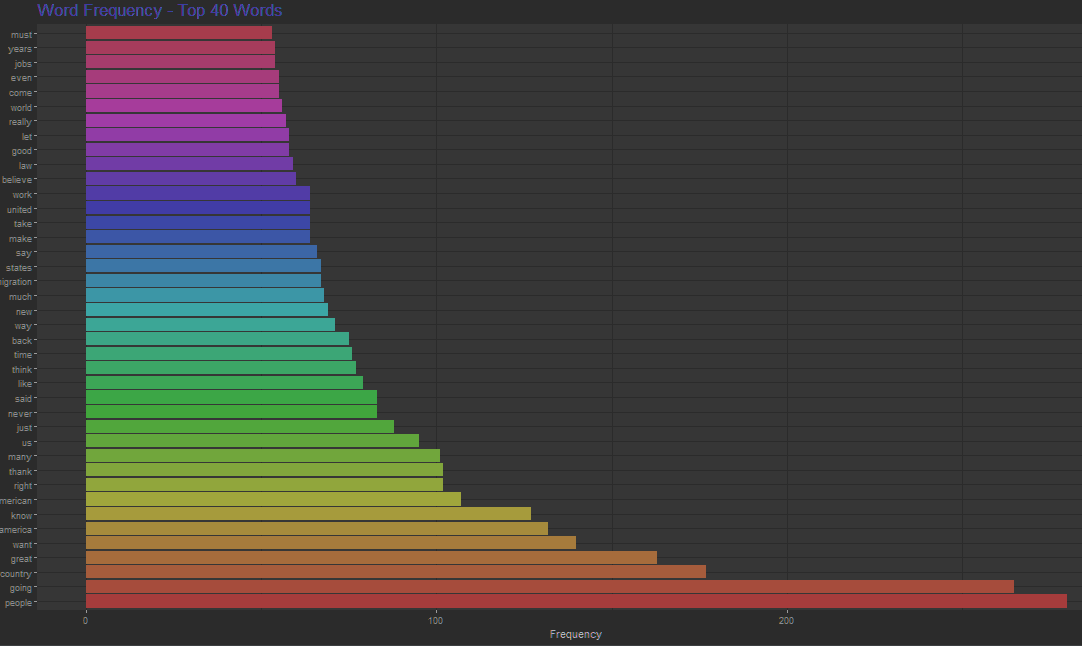
Se observa que de tener inicialmente 40380 palabras nos encontramos con una gran reducción de estas a tener 3555, se puede observar que las palabras que menos dice Donal Trump en los discursos analizados son las siguientes:



Posterior a esto y teniendo en cuenta la nube de palabras anterior con palabras que nos generaba mucho ruido pasamos a tener la siguiente:



Seguido de esto, pasamos a graficar las palabras que mas decía Trump en los discursos a analizar lo cuales son referentes a los 11 que se han propuesto en el taller surgiendo el siguiente gráfico:

Gráfico 1.

Se puede analizar del gráfico obtenido que Donald Trump en los 11 discursos que se están analizando, y teniendo en cuenta su correspondiente traducción de las siguientes 30 palabras más usadas:

"gente" "yendo" "país" "gran" "querer" "américa" "saber"

"americano" "correcto" "gracias" "muchos" "nosotros" "solo" "nunca"

"dijo" "como" "pensar" "tiempo" "atrás" "camino" "nuevo"

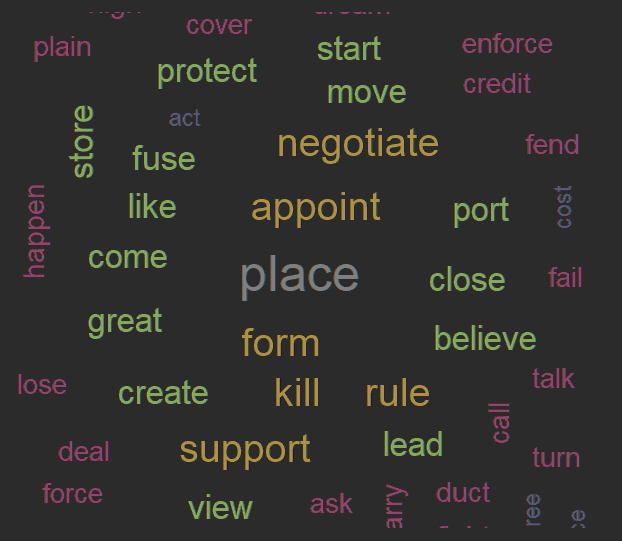
"mucha" "inmigración" "estados" "decir" "hacer" "tomar" "unidos"

"trabajar" "creer"

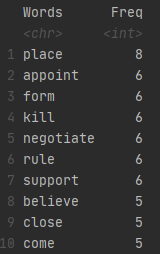
Hace más referencia hacia las personas que se están yendo de un gran país hacia América sabiendo que los americanos están de acuerdo o agradecidos con muchos de nosotros (es decir con los estadounidenses) haciendo referencia a que ellos nunca han dicho en un tiempo atrás o que habían pensado en un camino nuevo con mucha inmigración de estados creyendo en creer trabajar.

Finalmente, se hizo el stemming o la similaridad de las palabras con la función proporcionada por la librería SnowBallC a lo que a cada palabra, en cuanto a la relación de las palabras, les ha hecho a cada una el correspondiente stemming y de esta manera para finalizar se ha usado la función unique() ya que esta me permite poder traerme las palabras únicas, ya que la función stemming hace la correspondiente similaridad pero no las borra, en cuanto a unique() lo que hace es evitar la duplicidad y de tener las 3555 palabras finales al final se redujeron a 2606, lo que permite un análisis aun mas cómodo de las 40380 palabras iniciales.

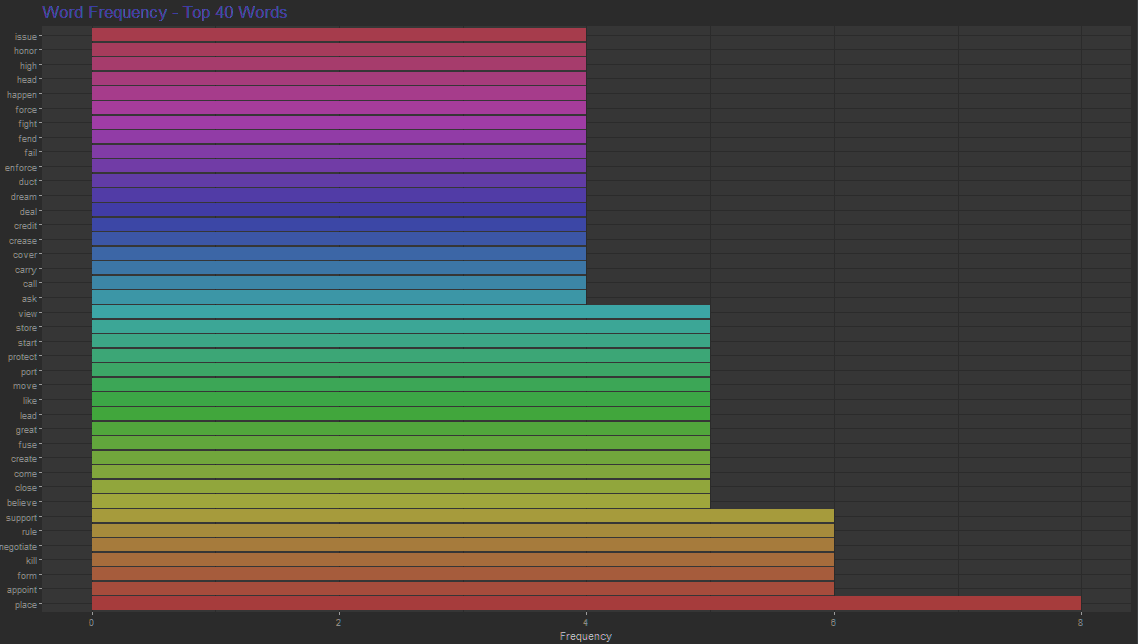
Posterior a esto, luego de realizado el stemming, se procedió a crear la nube de palabras correspondiente lo que dio como resultado el siguiente gráfico:



A lo que se observa que ahora las palabras con mas relevancia son las siguientes:



Finalmente se obtuvo el siguiente gráfico del stemming



Aquí se observa que ya la palabra “People” no lidera como el gráfico anterior, porque no estamos ya analizando el discurso como tal si no el Stemming, es decir la similaridad de todas las palabras que se han dado en el gráfico anterior, en este gráfico la palabra con mas similaridad la representa “place” con un valor de 5, la traducción de las 30 primeras palabras es:

"colocar" "nombrar" "formar" "matar" "negociar" "gobernar"

"apoyar" "creer" "cerrar" "venir" "crear" "fusionar"

"gran" "liderar" "como" "mover" "puerto" "proteger"

"empezar" "almacenar" "ver" "preguntar" "llamar" "llevar"

"tapa" "pliegue" "crédito" "trato" "sueño" "conducto"

Dando un análisis mas profundo de esto y teniendo en cuenta que es la similaridad la que se está analizando, puede deducirse que de las palabras tenidas en cuenta en elgráfico anterior representaban de una u otra forma similtud entre estas dando un total de 3555 palabras, a lo que posterior del stemming se redujo a 2606, es decir que ya se disminuye en gran manera y se hace mas fácil el análisis para reforzar lo anteriormente dicho.