APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE TEXTO

Andrea Amaya, Laura Agudelo, Camila Ospina, Luisa Acosta, Estefanía Echeverry, Esteban Bermúdez

Escuela de Estadística, Universidad Nacional de Colombia.

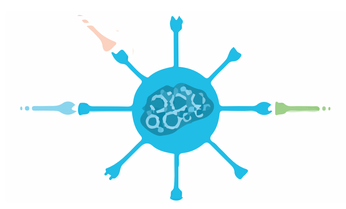
Introducción

El objetivo de este trabajo es el desarrollo de una aplicación que facilite a los usuarios la lectura y el análisis de textos por medio de resúmenes extractivos y gráficos descriptivos, obtenidos mediante análisis de minería de texto, procesamiento de lenguaje natural y métodos de clusterización

Metodología

1. Recolección de la información:

Web Scraping



API

Primero se obtiene el texto, ya sea desde una url, utilizando la técnica de web scraping sobre formatos HTML o mediante API

2. Limpieza de los datos:

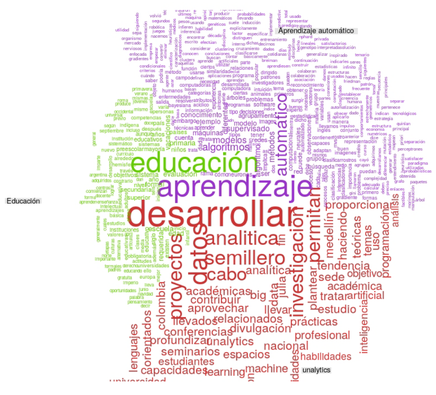
Una vez obtenida esta información en formato de cadenas de texto procedemos con la limpieza de las mismas, comenzamos removiendo palabras de parada (stopwords) y caracteres especiales que no aportan información relevante, como bien lo son los números, signos de puntuación, pronombres, artículos, preposiciones. Por ejemplo: ,

```
[1] "lusecitagomes: RT @JuMaJaRa: Daniel Quintero dice que hay que presionar al gobierno para hacer el metro subterráneo en Medellin, que es la misma propuesta…"
```

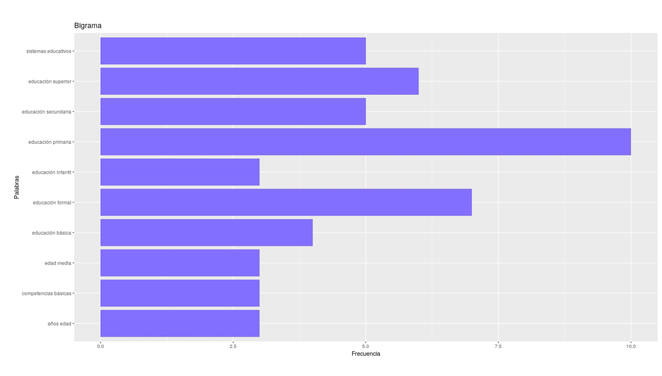
```
[1] "lusecitagomes rt jumajara daniel quintero dice presionar gobierno hacer metro subterráneo medellin misma propuesta "
```

3. Gráficos descriptivos:

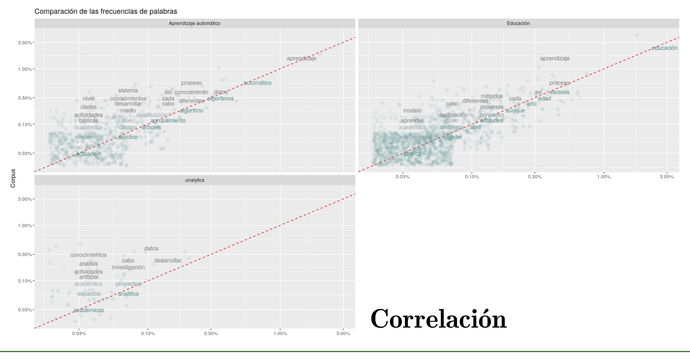
Preparadas las cadenas de texto y puestas en un marco de datos, se divide el texto por palabras individuales (token) para contar la frecuencia de las palabras, permitiendo así obtener distintos gráficos descriptivos.



Nube de palabras



Bigrama



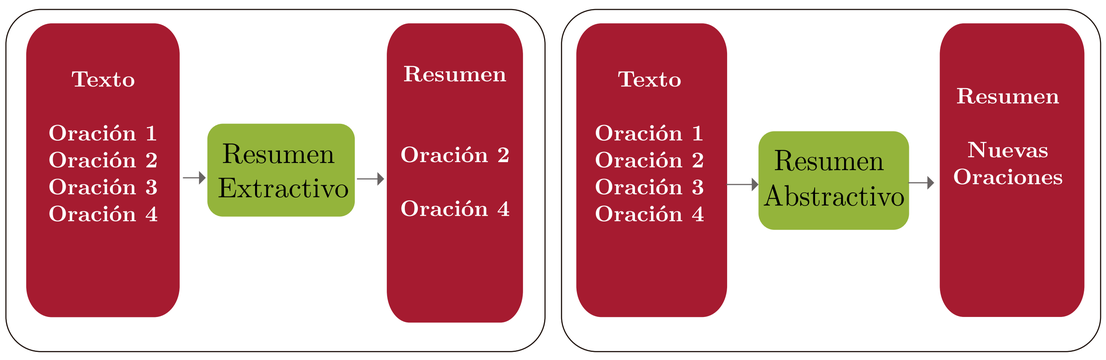
Correlación

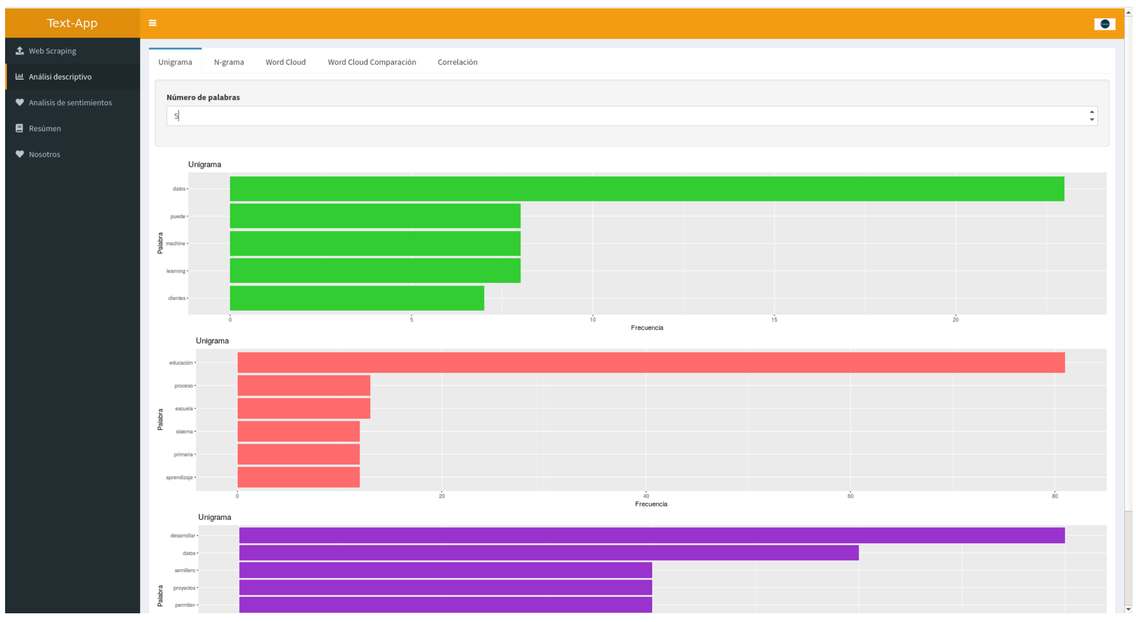
4. Aplicación

Los pasos clave dentro de la aplicación son una interfaz de usuario y un servidor que permitan controlar el funcionamiento de esta. La aplicación recibe como parámetro la URL de sitios web para extraer su texto, visualizar distintos tipos gráficos como son: nubes de palabras, bigramas, digramas de correlación y gráficos para el análisis de sentimientos, también permite crear un resumen extractivo.

3. Resumen extractivo y abstractivo:

Hay principalmente dos tipos de resumen: El primero es el extractivo, que toma las frases más relevantes del documento de entrada por medio de una ponderación y luego, sin alterarlas, las concatena. El segundo es el abstractivo que interpreta y examina el texto para generar uno más corto implementando el modelado.





Conclusiones

La minería de texto reduce los recursos y el tiempo empleado para analizar textos extensos, extrayendo patrones e ideas generales por medio de métodos automáticos, facilitando el acceso a conocimiento. Es por esto que esta metodología representa una de las mejores opciones para extraer y analizar este tipo de dato no estructurado.

Referencias

[1] RStudio Team (2019). RStudio: Integrated Development for R. RStudio,Inc., Boston, MA URL http://www.rstudio.com/.

[2] Text Mining with R https://www.tidytextmining.com/

[3] Dashboards en R: https://github.com/departmenttransport/janus-dashboard