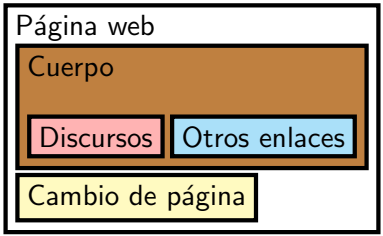
Scraping de discursos del Presidente Iván Duque Márquez.

Javier Andrés Vásquez Martínez

Junio, 2021

Introducción

La página web tiene la siguiente estructura:



Estructura de la página web

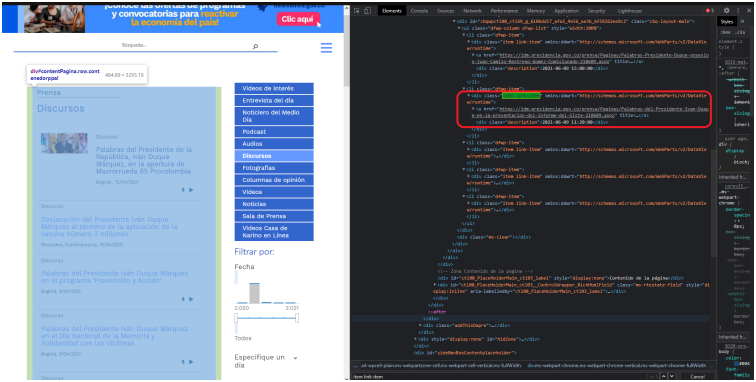


Figure: Obtención de las url de los discursos.

Estructura de la página web

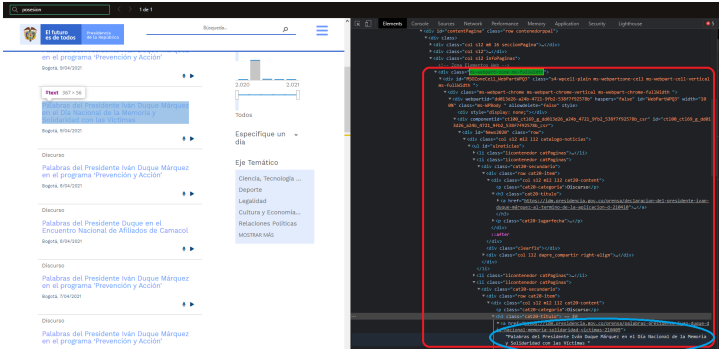


Figure: Obtención de los títulos de los discursos.

Estructura de la página web

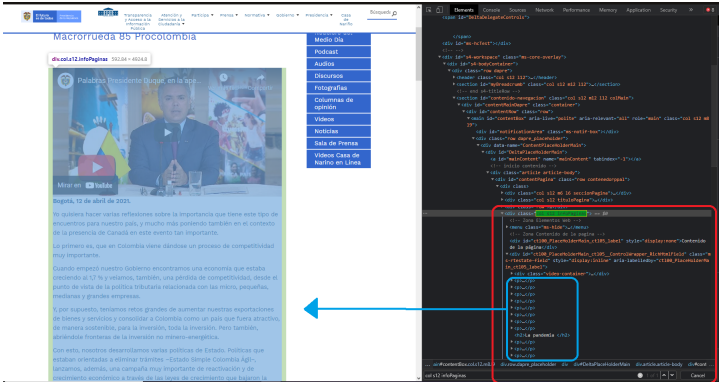


Figure: Obtención de los discursos.

TF-IDF

Para obtener el texto más relevante de cada discurso se utiliza la medida **frecuencia de término – frecuencia inversa de documento** (Term frequency – Inverse document frequency), en la que se calcula y se pondera la frecuencia de una palabra dentro de un texto y se compara con la frecuencia del término en una lista de textos.

$$\mathsf{tfidf}(t, doc, lista) = \mathsf{tf}(t, doc) \times \mathsf{idf}(t, lista).$$