O objetivo deste trabalho é fazer um *web scrapping* de uma busca sobre uma assunto em páginas na internet. Utilizando o R aplicaremos um algoritmo que fará um pré-processamento do texto, uma limpeza dos dados, deixando-os mais homogêneos transformando as palavras em minúsculas, retirando pontuações e numerações, espaços desnecessários, etc. Dessa forma a base fica sem "lixo" e com palavras mais relacionadas ao tema selecionado.

A base de dados escolhida foi retirada da busca pelo tema "rachadinhas" na página do G1, https://g1.globo.com/busca/?q=rachadinhas. Por dentro da linguagem Html foi identificado os locais que as informações seriam retiradas, tanto do título das matérias quanto do corpo das mesmas (exemplo: para títulos – h1, corpo do texto – articles).

A partir deste momento copiamos este endereço no algoritmo do R e será copiados parte dos dados das 20 primeiras páginas (copiando os títulos e corpo do texto). Serão desconsiderados qualquer tipo de vídeo e páginas com erro. A próxima etapa é o pré-processamento do texto já mencionado acima. Nesta etapa também colocaremos as palavras que não queremos que faça parte da nuvem de palavras. No caso em particular, estou procurando assuntos relacionados ao tema rachadinhas, então estou removendo as palavras "rachadinhas" e "rachadinha" que não é de interesse que apareça, já que é de conhecimento e pode ofuscar outras palavras. Também tive que retirar outras palavras como preposições, artigos, verbos, etc. Um outro ponto é que também realizamos a contagem (Figura 2) de quantas vezes uma determinada palavra aparece, pois vai impactar na nuvem de palavras, em relação ao tamanho das palavras e na sua cor (Figura 1).

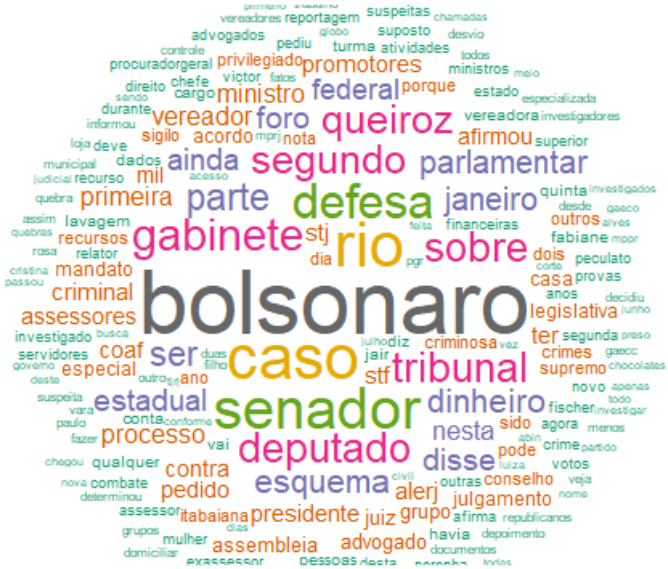

Figura 1 - Nuvem de palavra com a palavra "rachadinha" no G1

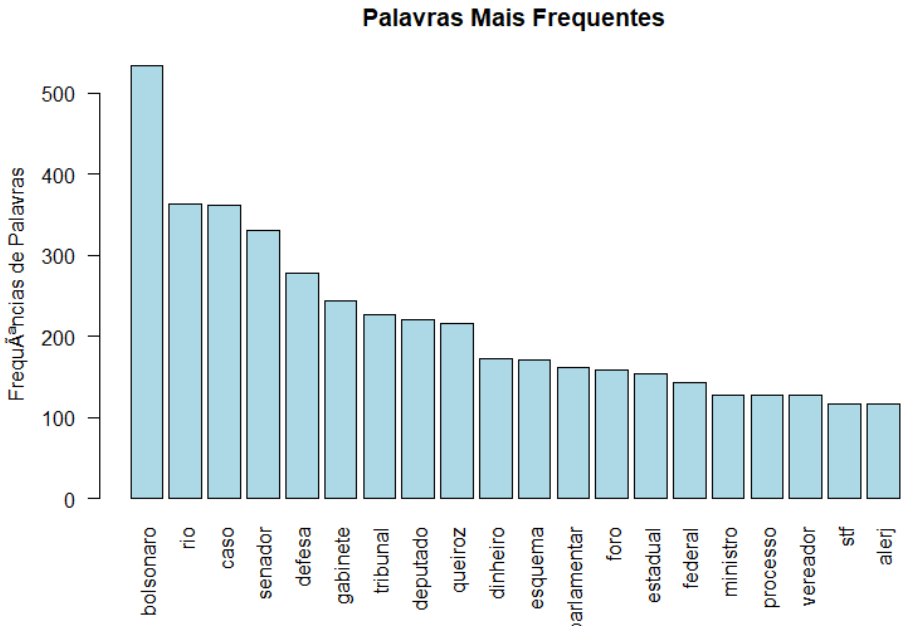

Figura 2 - Top20 palavras

Outro algoritmo é o spaCy (interligação com a linguagem python com o R) faz uma análise das palavras e cria entidades que irá analisar o grau de similaridade das palavras e suas dependências, gerando as arestas (entidades que acontecem juntas na mesma noticia). O algoritmo filtrará apenas pessoas e organizações como as entidades principais para as arestas. Por último o objeto do grafo será exportado para poder ser aberto no programa Gephi.

No Gephi é possível corrigir as palavras geradas na aba laboratório de dados, unir palavras que são iguais porém escritas de formas diferentes e remover palavras sem significados e/ou impactos. Trabalhando dentro dele é possível gerar um gráfico com as principais arestas, as que aparecem com maior frequência. Otamanho das arestas irá depender de sua frequência (Figura 3).

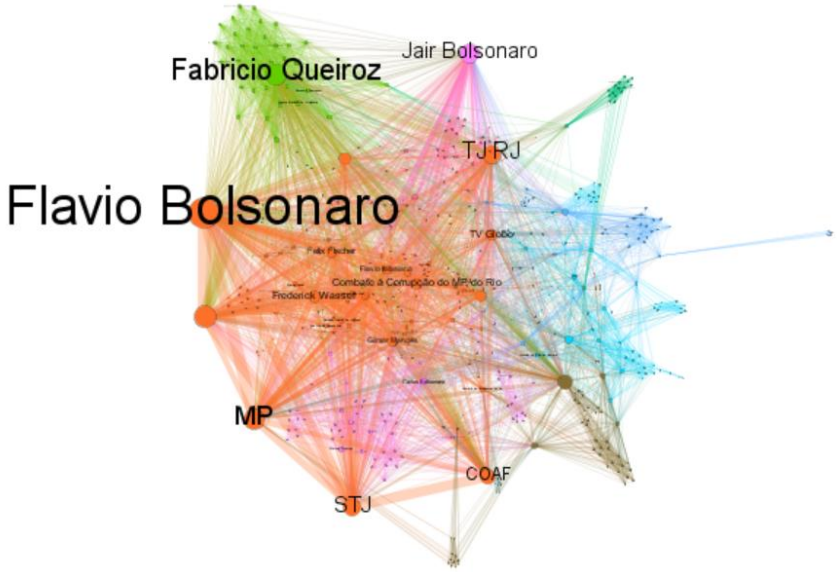

Figura 3 - Visualização gerada pelo Gephi

Os conhecimentos necessários para utilizar o algoritmo é entender o funcionamento do R e um pouco de programação. Foi necessário a instalação do Anaconda para criar a conectividade com o spaCy. E a instalação do Gephy para gerar o gráfico de arestas.

Resultados e Conclusões:

De acordo com a figura 3 gerada pelo Gephi, os dados batem com o que conhecemos sobre o que foi noticiado a respeito das “rachadinhas”:

A aresta principal é o Flavio Bolsonaro, seguida de seu assessor o Fabricio Queiroz. O tema estava sendo analisado pelo Tribunal de Justiça do RJ / Ministério Público, passou para o STJ se os dados do COAF poderiam ser analisados. Logo os relacionamentos (grau de similaridades) criados foram bem sucedidos.