Mineração de Texto

Críticas de adorocinema.com

Brendha Lima

Universidade Federal do Paraná

13 de junho de 2019

Sumário

- Motivação e objetivo
- Web Scraping
- Mineração de texto
- Conclusão

Motivação e objetivo

A Mineração de texto é um processo que utiliza algoritmos capazes de analisar coleções de documentos texto com o objetivo de extrair dados, informações e/ou resumos.

Neste trabalho, abordaremos críticas de 5 séries com maiores pontuações de usuários do site ADOROCINEMA.COM, através de web scraping.

Sobre o site:

- 6,1 milhões de visitantes únicos por mês (fonte : ComScore, janeiro 2014);
- 52 milhões de páginas vistas (fonte : Google Analytics, maio 2014);
- 4 milhões de vídeos vistos por mês (fonte : Google Analytics, maio 2014);
- Mais de 14 mil fichas de filmes;
- Mais de 32 mil fichas de personalidades;
- Mais de 13 mil trailers;
- Mais de 320 mil fotografias;
- Mais de 600 mil usuários cadastrados no site;
- Mais de 1 milhão notas de filmes dadas pelos usuários;
- Mais de 80 mil críticas de filmes escritas pelos usuários, etc.

Web Scraping

As 5 séries mais votadas (nº de páginas com críticas/nº de críticas)

- Game of Thrones (41/612)
- Breaking Bad (25/375)
- Sherlock (6/82)
- Suits (6/79)
- Supernatural (32/480)

Processo de criação de funções e extração das críticas

```
url2 <- "www.adorocinema.com/series-tv/melhores"
url <- "www.adorocinema.com"
fim <- "criticas"
urls <- paste0(url, sufist, fim)
```

** "sufist" significa o código de cada série, extraído através de expressões XPaths da url2.

** Criadas as urls principais, é feita outra função para ler todas as páginas, através da expressão XPath que indica as numerações.

Extraindo os comentários de cada série

Por fim, é criada a função onde é retirado apenas os textos de cada página, de forma similar utilizando também expressões XPaths. Essas expressões foram obtidas através de inspeções das urls dentro do navegador. Algumas delas:

-"//*[@id='col_content']/div/div[1]/div/div[2]/div/h2/a"

-“//nav[@class ='pagination cf']”

-“.//div[@class = 'content-txt review-card-content']”

Mineração de texto

Extraídas as críticas, as etapas para cada série se deram da seguinte forma:

- Transformação para caixa baixa, remoção acentos, pontuações, espaços, números e stopwords;
- Criação de cada Corpus;
- Checagem de dimensões das Matrizes de Documentos e Termos;
- Cálculo de esparsidade;
- Termos frequentes, nuvem de palavras;
- Correlação;
- Polarização.

O contexto de cada análise de texto acarreta em termos frequentes particulares. Como as críticas abordadas são de séries, as palavras "episodio", "serie"e "temporada"foram retiradas na etapa de remoção de stopwords.

Termos frequentes e nuvem de palavras

Game of Thrones

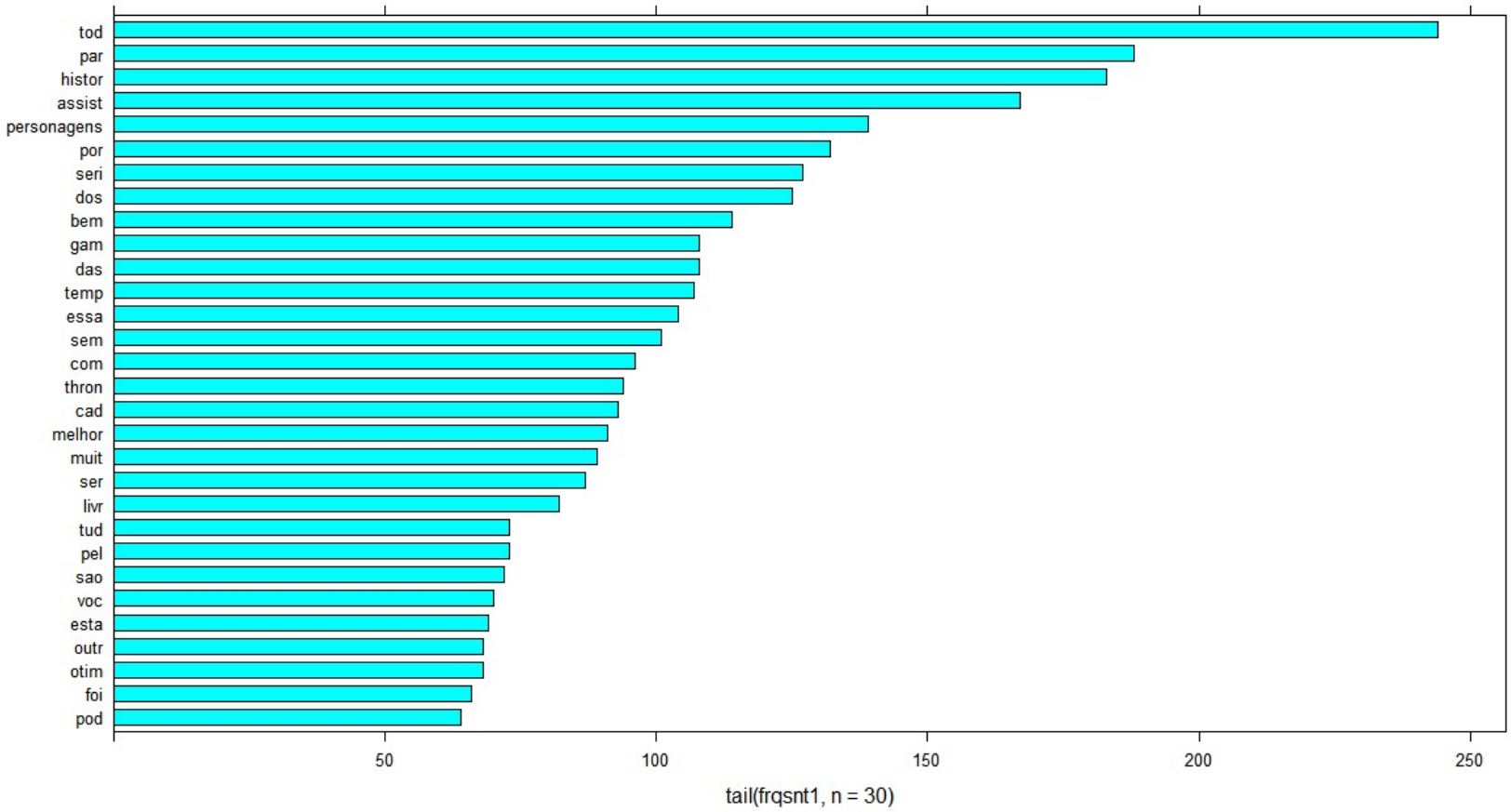

Figura 1: Termos frequentes; Game of Thrones

Figura 2: Nuvem de palavras; Game of Thrones

Breaking Bad

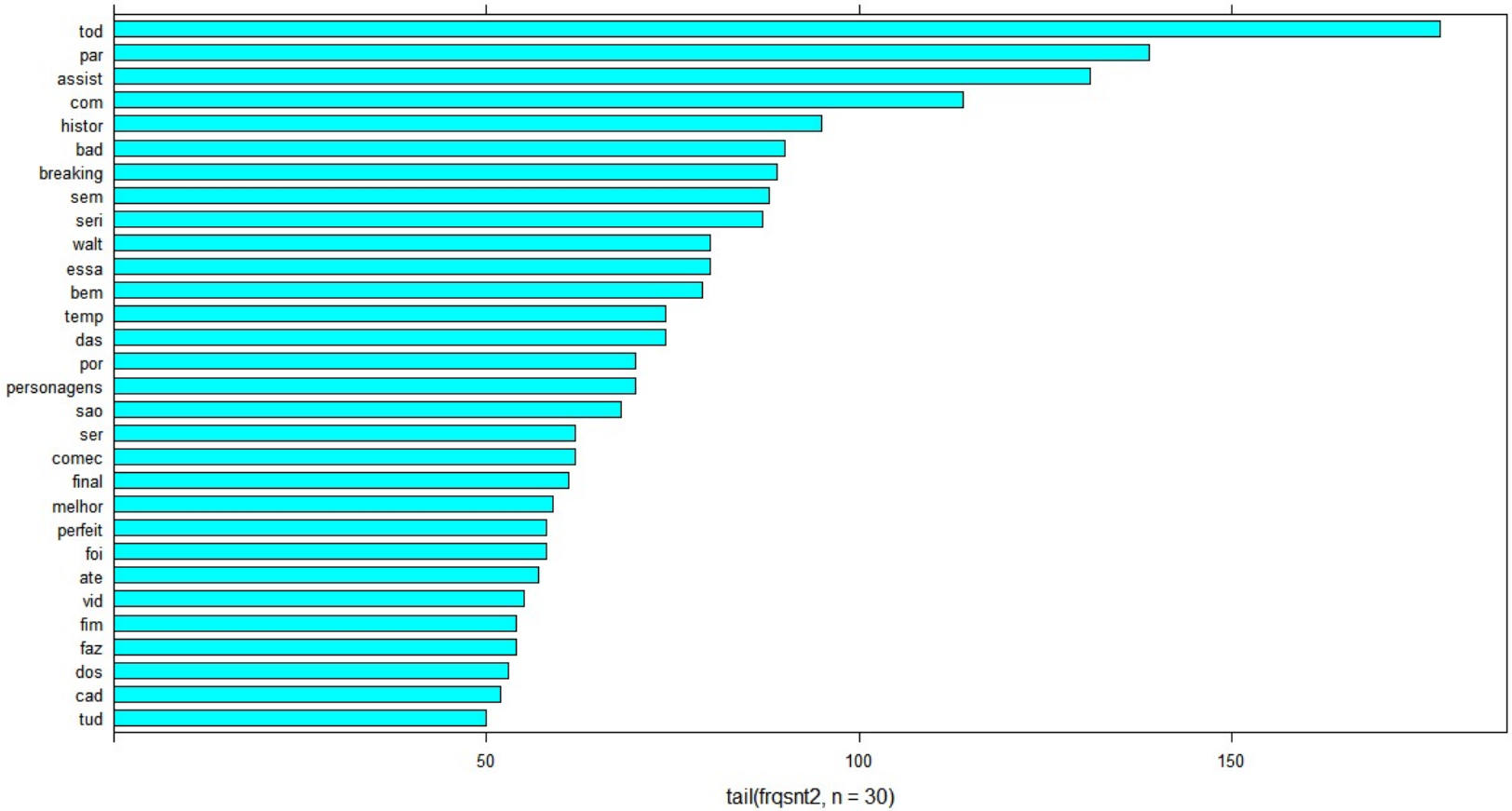

Figura 3: Termos frequentes; Breaking Bad

Figura 4: Nuvem de palavras; Breaking Bad

Sherlock

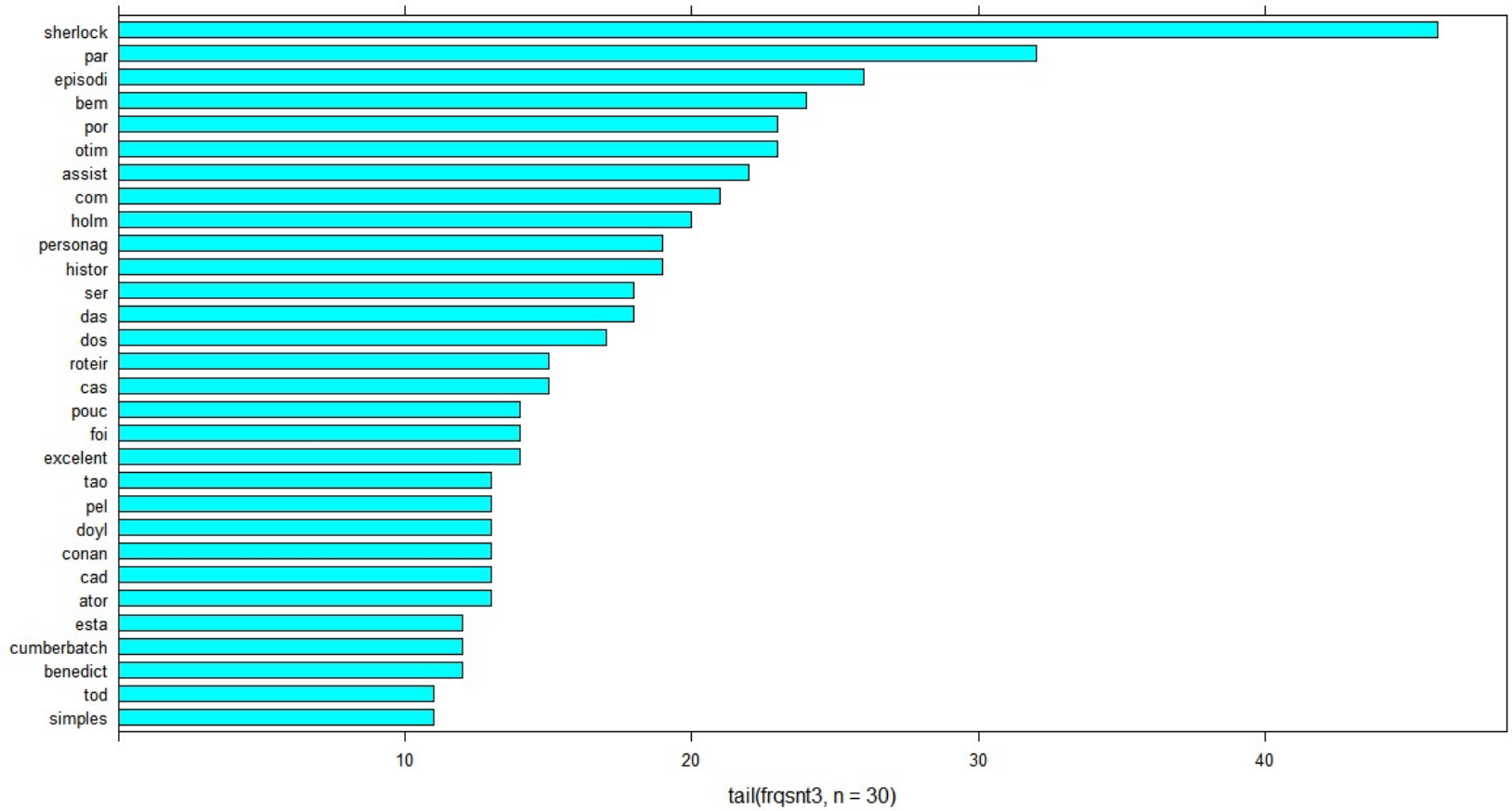

Figura 5: Termos frequentes; Sherlock

Figura 6: Nuvem de palavras; Sherlock

Suits

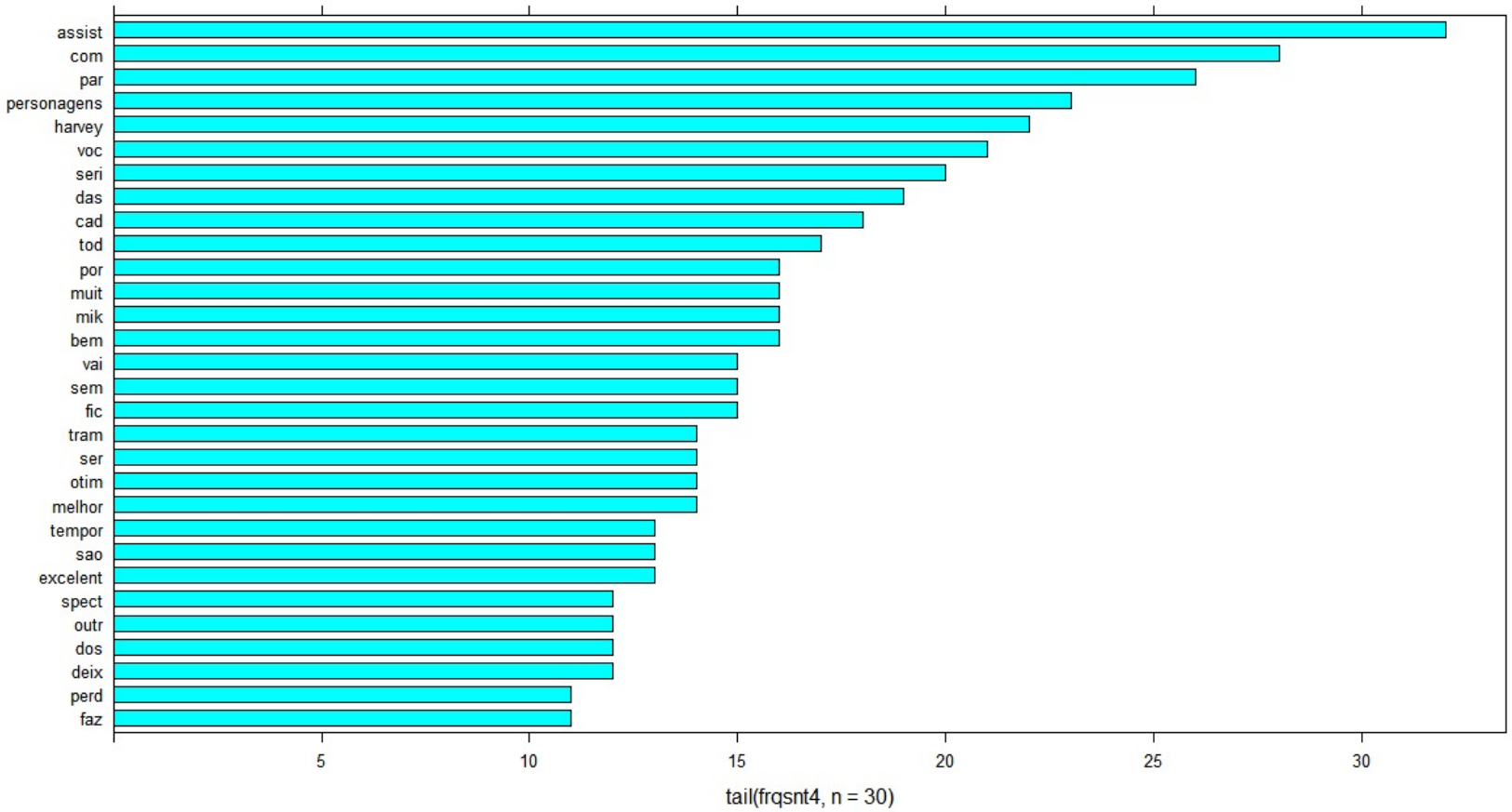

Figura 7: Termos frequentes; Suits

Figura 8: Nuvem de palavras; Suits

Supernatural

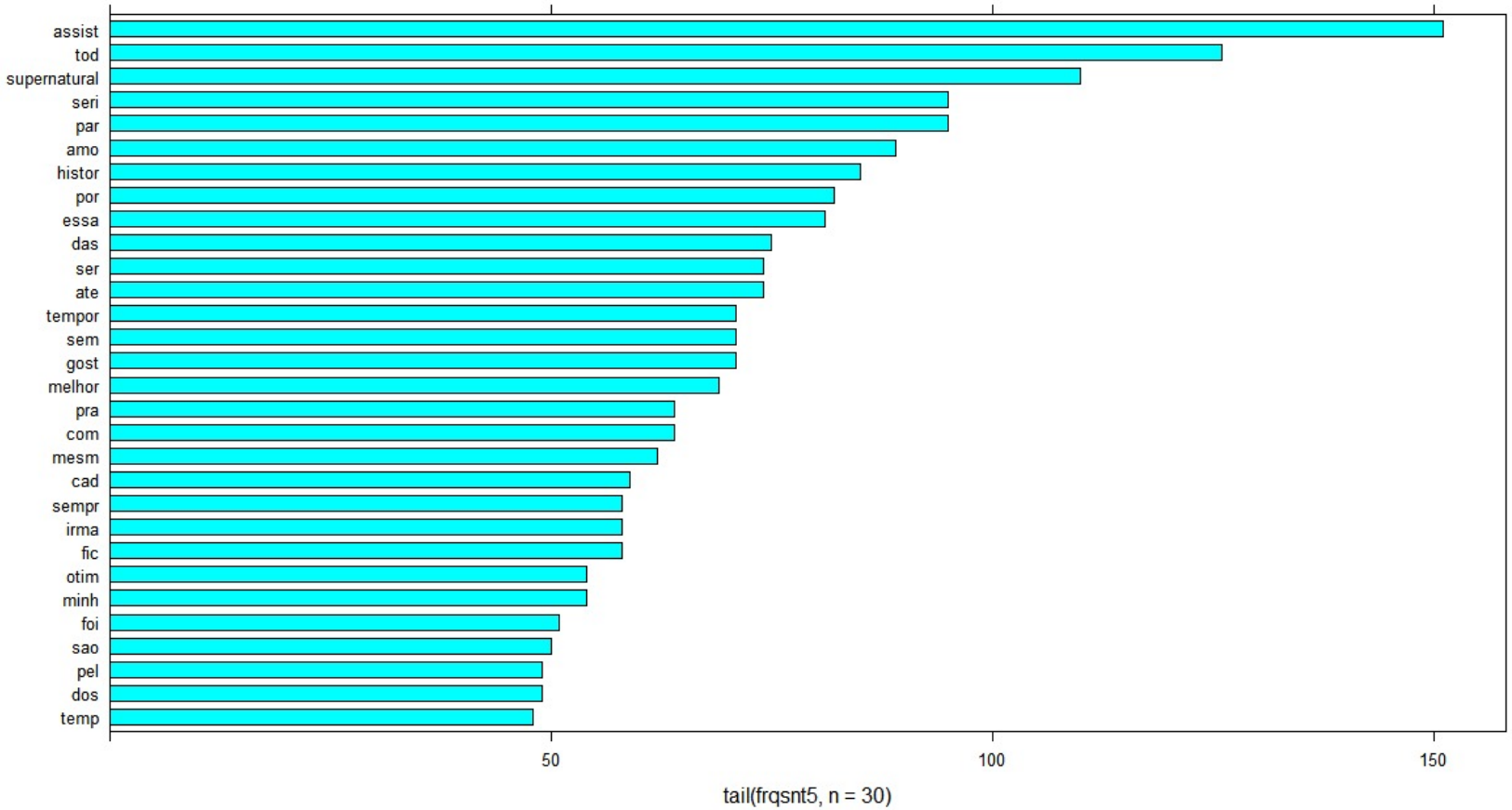

Figura 9: Termos frequentes; Supernatural

Figura 10: Nuvem de palavras; Supernatural

Conclusão

Podemos observar os principais resultados das críticas de cada série e ter uma noção do que elas relatam. Ainda, existem outras metodologias que podem ser aplicadas dependendo do interesse a ser estudado.

Por exemplo, inserir novamente as palavras **“serie”**, **“episodio”** e **“temporada”** para ver quais termos estão associados a elas.

Obrigada!