ATIVIDADES

Setembro a Novembro

REMOÇÃO DE PLURAL

```
# -ões, -ãos, -ães no plural para -ão no singular
txt <- gsub('\\b(.*)ões\\>\\b', '\\1ão', txt)
txt <- gsub('\\b(.*)ãos\\>\\b', '\\1ão', txt)

# -ais, -éis, -óis, -uis no plural para -al, -el, -ol, -ul no singular
txt <- gsub('pais', 'pai', txt) # exceção
txt <- gsub('\\b(.*)ais\\>\\b', '\\1al', txt)
txt <- gsub('\\b(.*)éis\\>\\b', '\\1el', txt)
```

STOPWORDS

- stopwords_bigramas.txt
- stopwords_discurso.txt
- stopwords_nomes_compostos.txt
- stopwords_nomes_simples.txt
- stopwords_partidos.txt
- stopwords_portugues.txt

INVERSE DOCUMENT FREQUENCY

- permite determinar as palavras mais importantes em cada discurso
- cada documento pode compartilhar palavras que não apenas as *stopwords*
- essas palavras devem ter peso baixo quanto à importância no discurso
- palavras frequentes no discurso específico devem ter alto peso

$$w_{i,j} = tf_{i,j} * log\left(\frac{N}{df_i}\right)$$

$$w_{i,j} = peso\ da\ palavra\ i\ no\ discurso\ j$$
$$tf_{i,j} = ocorrências\ da\ palavra\ i\ no\ discurso\ j$$
$$df_{i,j} = quantidade\ de\ discursos\ que\ contêm\ a\ palavra\ i$$
$$N = número\ total\ de\ documentos$$

MATRIZ TERMO DISCURSO

- 15.000 termos *vs* 6.000 discursos ano => matriz com 90 milhões de células
- MTD são esparsas, i.e., contém majoritariamente **zeros**
- estipular frequência de corte de zeros (fcz) para cada termo na MTD, i.e., a percentagem de zeros admissível para cada termo
- quanto mais próxima de 1, mais termos são mantidos na MTD

CLUSTERIZAÇÃO

- A análise de cluster hierárquica usa o conjunto de diferenças para os objetos em cluster.
- Inicialmente, cada objeto é atribuído ao seu próprio cluster e, em seguida, o algoritmo prossegue iterativamente, em cada etapa, juntando os dois clusters mais semelhantes, continuando até existir apenas um único cluster.
- Em cada estágio, as distâncias entre clusters são recalculadas pela fórmula de atualização de dissimilaridade de Lance-Williams, de acordo com o método *complete linkage* que encontra clusters similares.

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$

$d(x,y)$ é a distância entre $x \in X$ e $y \in Y$

X e Y são dois clusters

https://en.wikipedia.org/wiki/Complete-linkage_clustering

LATENT DIRICHLET ALLOCATOIN

No processamento de linguagem natural, a alocação latente de Dirichlet (LDA) é um modelo estatístico generativo que permite que os conjuntos de observações sejam explicados por grupos identificam partes de dados são semelhantes.

Por exemplo, se as observações são palavras coletadas em documentos, ele postula que cada documento é uma mistura de um pequeno número de tópicos e que a criação de cada palavra é atribuível a um dos tópicos do documento.

Os modelos de tópicos agrupam tanto documentos que usam palavras semelhantes, quanto palavras que ocorrem em um conjunto de documentos semelhantes.

https://en.wikipedia.org/wiki/Topic_model

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

PRÓXIMOS PASSOS

- identificação ideológica
- ajuste dinâmico da fcz de esparsidade
- supervised learning: classificação do discurso por partido
- avaliação do modelo de predição
- elaboração do relatório pesquisa