cebrap.lab - Captura de Dados com R

Thiago Meireles

Universidade de São Paulo — Departamento de Ciência Política
Cetic.br — Nic.br

27 de maio de 2022

O que é um corpus e como trabalhar com ele?

- É uma "biblioteca" do documento original em um formato mais simples acompanhado dos metadados
- É a base para qualquer manipulação do qual extrairemos os textos para criação de matrizes para realizar nossas análises
- Em outras palavras, é de onde retiramos as informações a atribuímos a novos objetos que serão analisados
  - Aqui fazemos a stemização, limpeza de *stopwords*, pontuação, etc.

Matriz Documento-Recurso

- Aqui temos a principal extração dos corpus utilizando o pacote *quanteda*, com a qual criamos matrizes documento-recursos
- Por que recursos e não termos?
  - Eles podem ser mais gerais, contendo termos brutos
  - Ou derivados, como classes gramaticais de termos, termos após remoção de palavras, etc.
  - Ainda podem ser mais específicos, como os n-gramas
- Tokenização: podemos extrair termos, caracteres ou frases inteiras, grupos de palavras
- Nossas análises são realizadas com essas matrizes

Modelagem de texto

- Aqui temos boa parte do conteúdo que foi apresentado no vídeo de ontem
- Veremos rapidamente alguns desses modelos no Tutorial 9, mas não entraremos nas especificações de cada modelo
- A ideia é apresentar códigos exemplo para desmistificar a produção de análises a partir da modelagem de texto
- Análise de equivalência, análise de sentimentos, análise de afinidade, semelhança entre os textos, clusterização em dendograma ("árvore"), posição de documentos (*Wordfish*), *topic modeling*.