cebrap.lab - Captura de Dados com R

Thiago Meireles

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

25 de maio de 2022

Pacote *stringr*

- Permite entender as características dos textos
  - Tamanho, conteúdo, posições de termos, etc.
- Criação de subconjuntos a partir de termos específicos
- Mas a principal função é a manipulação do conteúdo
  - Desde substituições simples até o uso de expressões regulares (*regex*)

Pacote *tm*

- Criação e manipulação de *corpus*
- Fácil remoção de pontuação, stopwords e stemização
- Tokenização
  - Bigramas ou n-gramas

Pacote *tidytext*

- Trabalha o texto como dado em formato dataframe comum
- Permite também a tokenização
- A utilização de uma linguagem tidy traz algumas vantagens
  - Integração com pipes, ggplot, etc.