Processamento de Texto

Luciano Barbosa

Centro de Informática
UFPE

Universidade Federal de Pernambuco

Corpus: Diversidade do Vocabulário

- N = número de tokens, V = vocabulário
- Heaps Law: $|V| = kN^b$
  - k,b: parâmetros que variam por corpus (valores típicos $30 \leq k \leq 100$ e $b \approx 0.5$)
  - O tamanho do vocabulário cresce na raiz quadrada do número de tokens

	Tokens = N	Types = |V|	N/|V|
Switchboard phone conversations	2,4 milhões	20K	120
Shakespeare	884K	31K	28
COCA	440 milhões	2 milhões	220
Google N-grams	1 trilhões	13+ milhões	77K

Fatores que Influenciam

- Quem escreveu (idade, gênero, raça)
- Quando foi escrito
- Língua: 7097 línguas no mundo
- Dialetos
- Finalidade: notícia, artigos científicos, livros de romance
- Múltiplas línguas: Hindi/Inglês

Tokenização

- Quebrar sequência de caracteres em palavras
- Quantas palavras existem nesta sentença falada?
  - "I do uh main- mainly business data processing"

- Exemplo de aborgadem simples
  - Qualquer sequência de caracteres alfanuméricos de tamanho mínimo 3
  - Terminado em espaço ou algum caracter especial
  - Letras maiúsculas convertidas para minúsculas

Tokenização

- Quebrar sequência de caracteres em palavras
- Exemplo de aborgadem simples
  - Qualquer sequência de caracteres alfanuméricos de tamanho mínimo 3
  - Terminado em espaço ou algum caracter especial
  - Letras maiúsculas convertidas para minúsculas

  - Ex 1: “Bigcorp’s 2007 bi-annual report showed profits rose 10%” -> “bigcorp 2007 annual report showed profits rose”

  - Ex 2: “Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing” - > “mr neill thinks that the boys stories about chile capital aren amusing”

Problemas em Tokenização

- Palavras pequenas podem ser importantes em algumas consultas
  - Ex: am, pm, el (paso), (world war) II
- Hífens
  - Algumas vezes são necessários
    - Ex: e-bay, wal-mart, cd-rom, t-shirts
  - Separam palavras
    - Ex: Dallas-Fort Worth, spanish-speaking

Problemas em Tokenização

- Caracteres especiais são importantes para URL, tags e código em documentos
- Palavras com letras maiúsculas podem ter significados diferentes
  - Bush, Apple
- Apóstrofo pode ser parte de uma palavra, parte de um possessivo (inglês), ou erro

rosie o'donnell, can't, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Problemas em Tokenização

- Números podem ser importantes
  - Ex: nokia 3250, united 93, quicktime 6.5 pro
- Pontos podem estar em números, abreviações, URLs, fim de sentenças etc
  - Ex: I.B.M., Ph.D.

Tokenização em Outras Línguas: Chinês

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Tokenização em Outras Línguas: Chinês

姚明进入总决赛 “Yao Ming reaches the finals”

3 palavras?
姚明 进入 总决赛
YaoMing reaches finals

5 palavras?
姚 明 进入 总 决赛
Yao Ming reaches overall finals

7 caracteres?
姚 明 进 入 总 决 赛
Yao Ming enter enter overall decision game

Tokenização: Ambiguidade

The two characters can be treated as one word meaning ‘monk’ or as a sequence of two words meaning ‘and’ and ‘still’.

Tokenização Sem Espaço em Branco: Japonês

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

Tokenização Sem Espaço em Branco: Alemão

Compounds in Dutch, German, Swedish

Computerlinguistik → Computer + Linguistik

Lebensversicherungsgesellschaftsangestellter

→ leben + versicherung + gesellschaft + angestellter

Life insurance company employee

Tokenização Sem Espaço em Branco: Turco

- Uygarlastiramadiklarimizdanmissinizcasina
- `(behaving) as if you are among those whom we could not civilize'
- Uygar `civilized' + las `become'
  + tir `cause' + ama `not able'
  + dik `past' + lar 'plural'
  + imiz 'p1pl' + dan 'abl'
  + mis 'past' + siniz '2pl' + casina 'as if'

Processo de Tokenização

- Simples
  - Palavra é qualquer sequência de caracters alfa-numéricos terminado por espaço ou caracter especial, tudo convertido para minúsculo
- Stanford tokenizer para inglês
  - http://nlp.stanford.edu/software/tokenizer.shtml

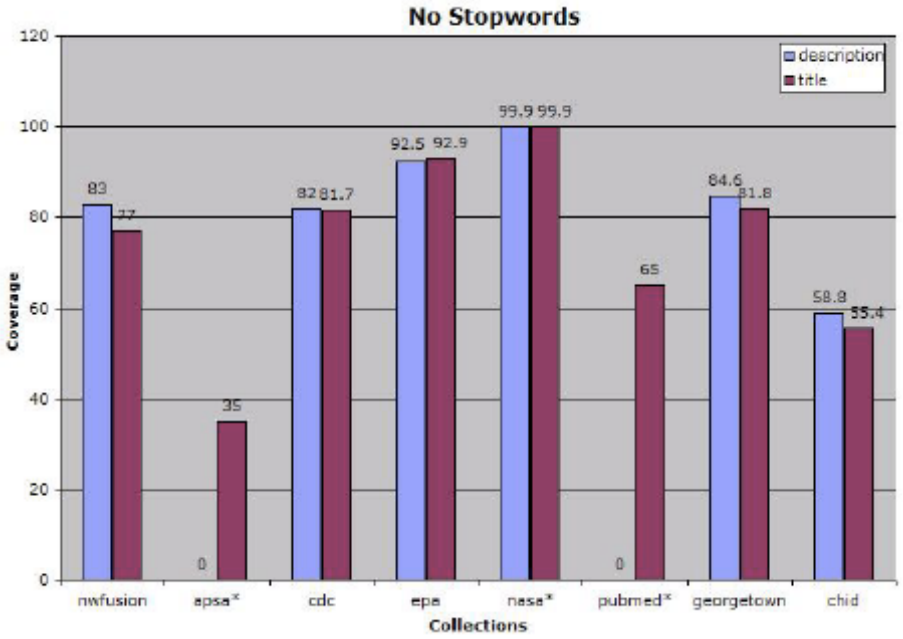
Stopwords

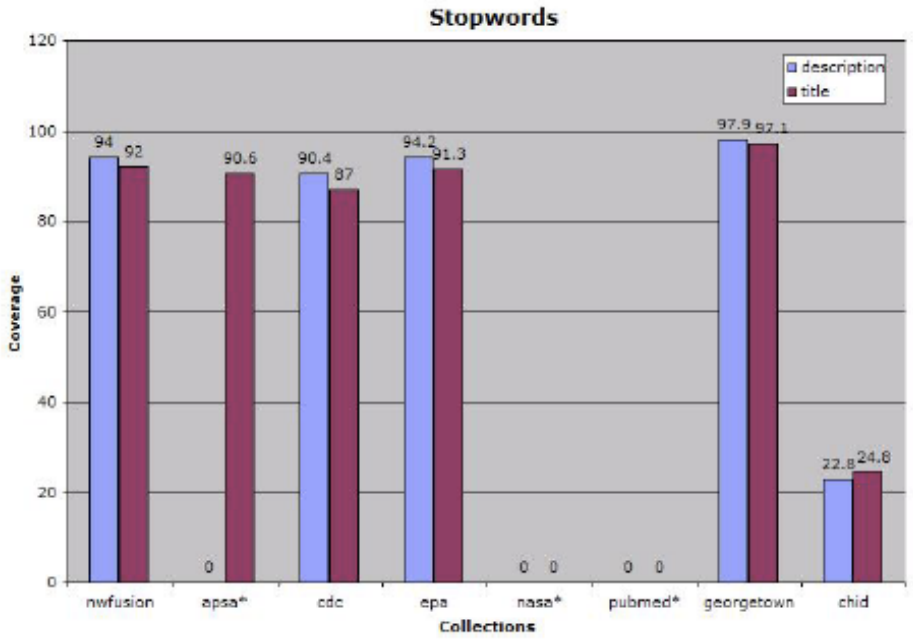
- Palavras que aparecem muito na coleção
- Não possuem muito significado
- Usualmente, não são boas para diferenciar
- Artigos, preposições, conjunções etc
- Criadas a partir de palavras com alta frequência ou de listas existentes
  - Ex: https://gist.github.com/alopes/5358189

Stopwords para Hidden-Web Crawling

- Siphon

Stopwords para Hidden-Web Crawling





Lematização

- Agrupar palavras com mesma raiz (lemma)
  - am, are, is → be
  - car, cars, car's, cars' → car
  - quero, queres → querer
  - He is reading detective stories → He be read detective story

Lematização: Morphological Parsing

- Quebra a palavra em morfemas
- Morfemas
  - Menor unidade com “significado” que compõe a palavra
  - Stem: unidade com o significado principal
  - Affixes: adiciona ao stem com uma função gramatical
  - Ex: unlikeliest
    - Stem: likely
    - Affixes: un-, -est
- Essencial para linguagens morfologicamente complexas como árabe

Stemming

- Reduz variações morfológicas das palavras para um stem em comum
- Remove prefixo ou sufixo -> stem
  - Ex: connect – connected, connecting, connection, connections
- Não há um consenso sobre benefícios (depende da língua)

Porter Stemmer

- Consiste de uma série de regras
- Comete erros e difícil de modificar

Original text:
Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:
document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

- https://github.com/stanfordnlp/CoreNLP

Remoção de Stopwords e Stemming na Reuters-RCV1

	number
unfiltered	484,494
no numbers	473,723
case folding	391,523
30 stop words	391,493
150 stop words	391,373
stemming	322,383

Segmentação de Sentenças

- Caracters ! e ? são precisos para separar
- Caracter “.” nem tanto
  - Abreviação: Dr.
  - Número: 7.5%
- Estratégia mais usada:
  - Toqueniza primeiro
  - Usa regras ou Machine Learning para classificar o ponto
  - Uso de dicionário de abreviação
- Exemplo de ferramenta: spacy (https://spacy.io/)