Primeira Lista de Exercícios - Mineração de Textos 2018.2

Raissa Camelo Salhab

[1]Departamento de Computação – Universidade Federal Rural Pernambucana (UFRPE)
Bacharelado em Ciência da Computação

srtacamelo@gmail.com

Resumo. *Este documento se trata de uma apresentação formal da(s) atividade(s) solicitada(s) em sala de aula. O conteúdo contido abaixo refere-se a resolução da lista de questões n[a] 1 da disciplina.*

1. Primeira Questão

No método in-line as anotações são inseridas em conjunto com os dados, o que possui a vantagem de reunir todos os dados necessários em um único arquivo, o que pode ser prático na hora de processa-los. Contudo este método não preserva o arquivo fonte original, logo se houver erros no processo de tagging ou necessidade de utilizar o arquivo dos dados limpo (sem anotações) não será possível. Outra desvantagem é que os arquivos ficam mais difíceis de compreender dado a grande quantidade de informações juntas aos dados.

O método stand-off preserva o arquivo dos dados originais e cria arquivos separados para cada camada de anotações, referenciando cada anotação a um dado especifico, onde existem chaves que ligam o dado e suas respectivas tags (metadados). Esse método organiza de forma consistente as anotações e permite que as anotações sejam trocadas de camadas sem a necessidade de alterar outras anotações. O arquivo fonte não possui anotações, sendo deixado inalterado, logo, se for necessário refazer as tags ou utilizar os dados brutos (sem tags), esse método torna isso mais fácil. Uma desvantagem é que dependendo da quantidade de anotações e camadas de anotações o processo de ligar cada uma das tags a seus respectivos dados pode ser longo e dispendioso.

2. Segunda Questão

O método de em camadas é mais organizado, separa os dados originais de suas tags e as divide em camadas consistentes com a tipologia de cada metadado. Com este método as anotações ficam mais praticas e mais fáceis de visualizar.

3. Terceira

A ferramenta Brat utiliza anotação estrutural (Structured annotation) que consiste em anotações com uma estrutura/ forma predefinida e organizada, que possa ser lida e interpretada por um computador, não um texto livre. Essa ferramenta pode ser utilizada em inúmeras situações diferentes. Dentre suas utilidades se encontram: Anotação de arquivos no formato brat pré- anotados por outras ferramentas, detecção de menção de identidade, extração de eventos, resolução de coreferencia, normalização e etc.

4. Quarta

POS tagging é um metodo de anotação no qual cada palavra (dado) recebe uma classificação, indicando geralmente a morfologia de cada palavra no texto original, analisando também a classe gramatical da palavra em relação a sentença completa. Parsing por sua vez é um método que analisa o texto levando em consideração sentenças completas ao invés de anisar cada palavra individualmente. No parsing são formadas árvores que ligam as palavras das frases, conferindo significado para a frase inteira e indicando tags para cada ligação.

5. Quinta

- Parsing Constintuinte No parsing constituinte são formadas as arvores sintáticas de maior probabilidade para a frase, geralmente são utilizadas gramaticas livres de contexto e as palavras são organizadas na árvore de forma hierárquica, formando sub-frases dentro da frase.
- Parsing de Dependências O parsing de dependências é baseado na teoria linguística de dependência gramatical, onde as estruturas sintáticas das frases correspondem a relacionamentos binários assimétricos entre as palavras que as compõem.

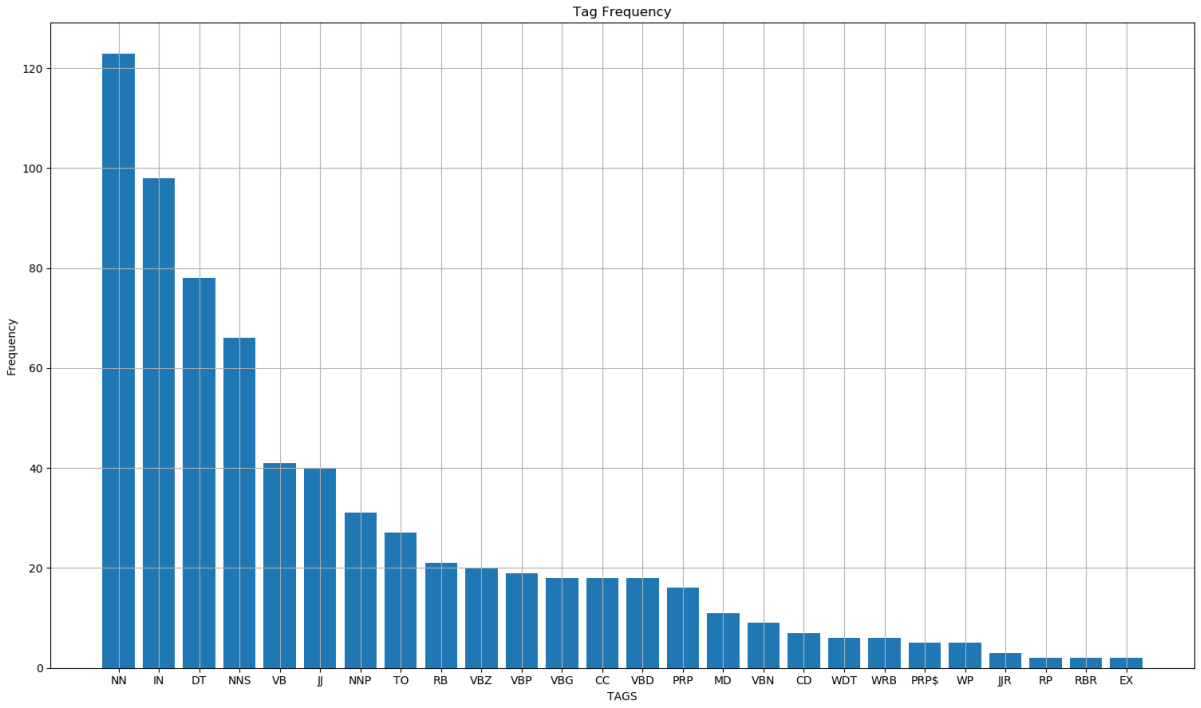
6. Oitava

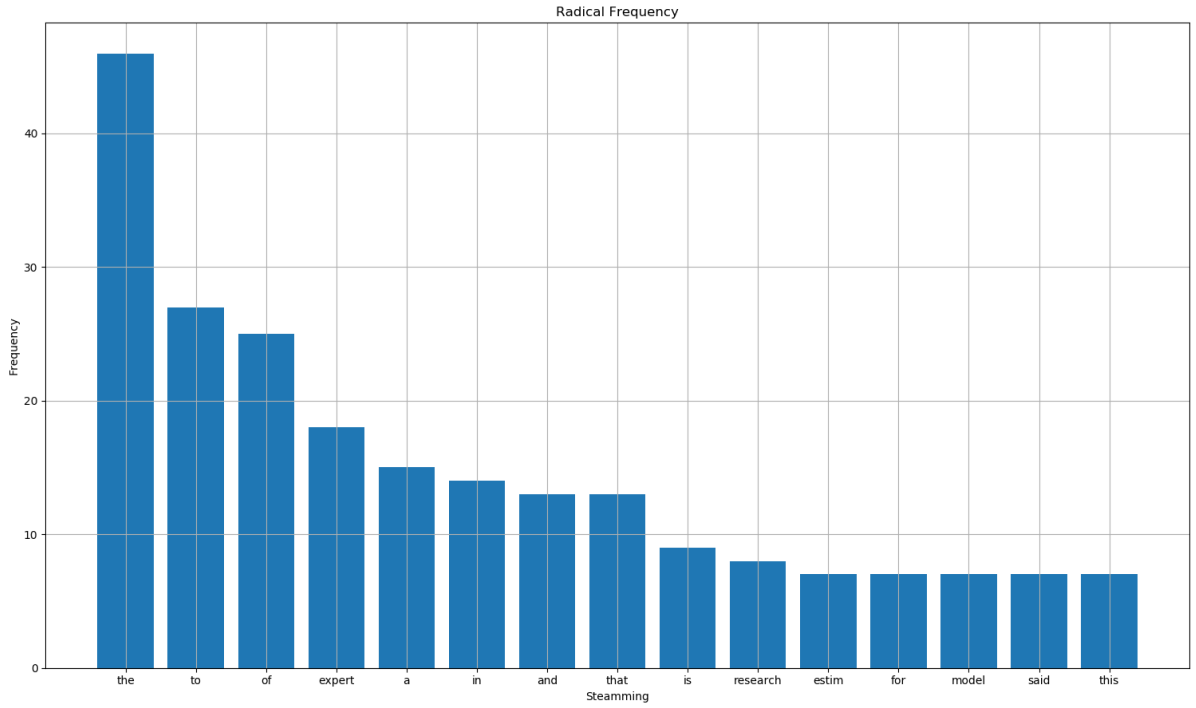
- A) Utilizando o tokenizer do NLTK a soma total de palavras encontradas foi 692
- B) A quantidade de radicais encontrados (sem repetições)foi 273, utilizando osteammer SnowBall do NLTK
- C) O número de sentenças detectados foi 24 A média de tokens por frases é 28 (28,8...)
- E) As Tags NN e IN, imagem no fim do arquivo
- F) Os dois grupos de radicais mais fraquentes são preposições e artigos. Imagem no final do artigo

7. Nona

Este stemmer é um pouco obsoleto, muitas palavras não sofrem grandes variações e outras não são nem radicalizadas. As ferramentas disponiveis nas bibliotecas de text mining são melhores.

Figura 1. Questão 8 Letra E




Radical Frequency
Frequency
0
10
20
30
40
the
to
of
expert
a
in
and
that
is
research
estim
for
model
said
this
Steamming