UFRPE – Departamento de Computação
Disciplina: *Mineração de Texto – Prof. Rinaldo Lima*
Aula 02: *Introdução ao Processamento de Linguagem Natural*
Lista de Exercícios (em dupla)
Data de Entrega: **26/09/2018**
Enviar email anexando arquivo zipado com o título [**LE_01_Text_Mining**]: para
rinaldo.ufrpe@gmail.com

UFRPE
Universidade Federal Rural de Pernambuco

1ª. LISTA DE EXERCÍCIOS

1. Quais as vantagens e desvantagens dos tipos de anotação de corpora *inline vs. stand-off?*

2. Ao seu ver, porque o formato híbrido, isto é, a anotação "em camadas" é o mais usado hoje em dia?

3. Quais tipos de anotação são realizadas pela ferramenta BRAT (http://brat.nlplab.org)? Para quais problemas de mineração de textos ela pode ser usada?

4. O que se entende por POS tagging? E qual a diferença entre POS taging e parsing?

5. Discuta sobre as propriedades do parsing constituinte e o parsing de dependências

6. Escreva, em sua linguagem de programação favorita, um algoritmo de *tokenização* para a língua portuguesa. Use o arquivo "texto_pt.txt" em seus testes.
Dica: baseie-se no algoritmo dos slides da Aula 02.

7. Escreva, em sua linguagem de programação favorita, um algoritmo para *segmentação* de *sentenças(orações)* de textos em português. Use o arquivo "texto_pt.txt" em seus testes.
Dica: baseie-se no algoritmo do slides da Aula 02.

8. Use o NLTK para criar um *pipeline* que realize as seguintes tarefas, nesta ordem:

- *Tokenization, Sentence Splitting, Lemmatization, Stemming e POS tagging*

Em seguida gere as seguintes informações estatísticas e gráficos de barras em relação ao texto em inglês "*texto_en.txt*":

a. Quantas palavras temos em todo o texto?
b. Quantos radicais (stemming) diferentes existem?
c. Qual o número de sentenças e a média de tokens por sentença?
d.

e. Gere um gráfico de barra do conjunto de POS tags de todas as palavras do texto. Ordene os resultados e responda: quais classes gramaticais correspondem a mais de 70 ou 80% do total?
f. Gera um outro gráfico de baras de todos os radicais presentes no texto. Ordene-os por sua contagem no texto e depois responda:
    - Existe alguma característica que se sobressai em relação as categorias gramaticais dos primeiros colocados na lista ordenada?

9. Use o stemmer em português disponível no link http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html para gerar os stemming (radicais) de todas as palavras do texto em anexo "texto_pt.txt".

Pergunta:

- este stemmer funciona bem ao seu ver? Isto é, existem palavras que ele reduziu demais seu radical e em outras ele não fez nada? Discuta sobre isso.

MATERIAL DE APOIO PARA OS EXERCÍCIOS

NLTK (Instalação e uso), subtarefas e geração de gráficos (histogramas)

https://www.howtoforge.com/tutorial/install-and-use-nltk-for-human-language-processing/

http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html

NLTK tutorial

https://dzone.com/articles/natural-language-toolkit-nltk

https://www.cs.toronto.edu/~frank/csc2501/Tutorials/cs485_nltk_krish_tutorial1.pdf

http://www.nltk.org/howto/portuguese_en.html

Stemming

http://www.nltk.org/api/nltk.stem.html

http://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization

Gerando estatísticas a partir de um texto

http://www.nltk.org/book/ch01.html

Spacy – NLP Toolkit

https://spacy.io/