

UNIVERSIDADE FEDERAL DE LAVRAS

DEPARTAMENTO DE CIÊNCIAS DA COMPUTAÇÃO

GCC 151 – Introdução ao Processamento de Línguas Naturais

Professores: Paula Christina Figueira Cardoso e Erick Galani Maziero

Prova 1

Data e horário de entrega: 28/Abr/2019 (domingo) até às 23h59

Linguagem de programação: Python 3.*

Modo e Local de entrega: códigos no repositório do Github, destinado às avaliações da disciplina GCC151. Será verificado o último *commit* antes da data e horário de entrega. Os corpora devem estar em um arquivo compactado, no Google Drive, compartilhado por link.

Nesta primeira avaliação, o aluno deve apresentar a documentação (na forma de comentários no código da biblioteca ou blocos *markdown* no Jupyter notebook) e código para realizar as seguintes tarefas principais:

1. (Jupyter notebook : 2.0 pts) Compilação de uns corpora (com no mínimo dois córpuses, de domínios textuais diferentes);
 - a. Cada córpus deve estar em um diretório específico e conter, pelo menos 500 textos cada
 - i. Cada texto deve estar no formato *.txt* codificado em *utf-8*
 - ii. Os corpora não devem ser *commitados* no Github, mas um link para download do mesmo (zipado) deve ser disponibilizado na documentação do código que o utilizar
 - b. As seguintes estatísticas devem ser apresentadas, por córpus:
 - i. 20 palavras mais frequentes
 - ii. 20 palavras menos frequentes
 - iii. Tamanho médio das palavras
 - iv. Tamanho médio das sentenças, em número de palavras
 - v. Outras duas estatísticas que achar interessante
2. (Arquivos *.py* : 3.0 pts) Criação de uma biblioteca de PLN, com rotinas de normalização textual do nível *lexical*:
 - a. *remoção de pontuação*
 - b. *remoção de acentos*
 - c. *remoção de stopwords*
 - d. *lowercase*
 - e. *stemming*
 - f. *tokenizar texto em sentenças*
 - g. *tokenizer texto em palavras*
3. (Jupyter notebook 2.5 pts) Normalização dos corpora, com vistas à criação de dois modelos, a saber, Word2Vec e Doc2Vec.
 - a. Para cada córpus, será criado um modelo Word2Vec
 - b. Para todos os corpora, apenas um modelo Doc2Vec será criado
4. (Jupyter notebook : 2.5 pts) Uso dos modelos gerados, conforme abaixo:
 - a. Uso do Word2Vec
 - i. Dada uma palavra w_1 de um córpus, quais as 10 palavras mais similares a w_1 ?

1. Exemplifique com três palavras e discuta como poderia melhorar os resultados. Pense no nível da morfologia ou outro do PLN.
 2. Tentar alguma abordagem para comparar dois documentos diferentes utilizando os vetores do Word2Vec
- b. Uso do Doc2Vec
- i. Dados os documentos (textos) de corporas diferentes, utilize os vetores para encontrar os documentos mais similares
 1. Exemplifique com três documentos e discuta os resultados. Ao ser ver, foram bons? Os documentos realmente são parecidos? O que poderia ser feito para melhorar os resultados?

Envie, para o e-mail erick.maziero@ufla.br o link do repositório git, junto com seu nome completo.