UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

3a lista de exercícios **29 de agosto de 2023**

Instruções:

A lista deve ser respondida por grupos de até 2 pessoas (graduação) e individualmente (mestrado).
Resoluções idênticas de grupos distintos serão desconsideradas.
O código e demais dados devem ser anexados a cada questão.
Data limite para entrega: 12/09/2023.

Usando sua base de textos após os pré-processamento realizados na lista 1, realize as seguintes tarefas:

1. O objetivo dessa questão é desenvolver buscadores de palavras e documentos.

a) escolha e aplique um modelo do tipo word2vec a seus textos, compatível com o idioma de seus textos (inglês ou português).
b) escolha 5 palavras de consulta que não estejam em nenhum dos textos. Para cada palavra de consulta, encontre as 3 palavras **de seu conjunto de textos** mais parecidas com cada uma das palavras de consulta e exiba os documentos onde estas palavras aparecem.
c) Seja d um documento da base e w uma palavra de consulta. Implemente o seguinte algoritmo para buscar documentos:
    1. Encontre $d_{10}(w)$: a lista com as 10 palavras mais parecidas com w em um certo documento d.
    2. Para cada documento d, calcule a distância média $DM_{10}(w)$ entre w e as palavras de $d_{10}(w)$.
    3. Recupere os 3 documentos da base cuja $DM_{10}(w)$ é menor.
d) aplique o algoritmo da letra c) para buscar documentos em 5 palavras distintas, e exiba os 3 documentos mais próximos de cada um.

2. Resolva novamente a primeira questão da 2a lista e compare com os resultados obtidos anteriormente:

a) Aplicando a representação vetorial Doc2Vec combinado com os classificadores usados anteriormente.
b) Usando pelo menos duas arquiteturas de redes neurais que utilizem camadas Embedding, convolucionais e LSTM.

3. Usando sua base de textos, e se baseando no exemplo disponível em `https://keras.io/examples/generative/lstm_character_level_text_generation/`:

a) Treine uma rede LSTM para gerar texto a nível de caractere, que receba uma quantidade fixa de caracteres Q como entrada. O treinamento deve ser realizado considerando um conjunto supervisionado que prevê o próximo caractere de uma sequência de tamanho $Q = 10$, usando sequências de sua base.
b) Após o treinamento, exiba pelo menos 5 exemplos de textos dados de entrada, e do texto gerado em seguida pela rede treinada. Para cada exemplo, gere pelo menos 50 caracteres consecutivamente.

4. Usando sua base de textos e a biblioteca spaCy, realize as seguintes tarefas:

a) Extraia as etiquetas gramaticais (POS) de cada token do seu textos.
b) Calcule e plote um gráfico com as frequências de cada tipo gramatical.
c) Extraia entidades do tipo pessoa e lugar dos seus textos.
d) Identifique e liste as pessoas mais frequentes nos seus textos. Você só deve contar cada entidade 1 vez por documento.