

Processamento de Linguagem Natural com Deep Learning

Lista de Exercícios 1

Considerando os slides *2.2 Tokenization DLForNLPLastversion* e o arquivo *A-Semana-Machado de-Assis* que acompanha esta atividade, escrevam comandos em UNIX que respondam corretamente as seguintes questões: ~

1. Faça a contagem de cada palavra presente no texto.
 - Saída esperada: Um arquivo que contem cada palavra acompanhada do número de ocorrências das mesmas no texto. ^
2. Aplique *downcasing* (deixar todos os caracteres em minúsculo) no texto inteiro. ^
 - Saída esperada: Um arquivo que contem o texto com todos os caracteres em minúsculo. ^
3. Faça a contagem da ocorrência de cada sequência de vogais distintas presentes no texto (ex.: ^ As sequências “e”, “ei” e “o” da palavra “Prefeito”). ^
 - Saída esperada: Um arquivo que contem cada sequência acompanhada do número de ocorrências das mesmas no texto. ^
4. Encontre as 50 palavras mais comuns do texto.
 - Saída esperada: Um arquivo que contem cada uma dessas palavras acompanhada do número de ocorrências das mesmas no texto. ^
5. Encontre as palavras no texto que terminam em “ar” (ex. “amar”).
 - Saída esperada: Um arquivo que contem cada uma dessas palavras acompanhada do número de ocorrências das mesmas no texto. ^
6. Conte quantas palavras de 4 caracteres existem no texto.
 - Saída esperada: Um arquivo que contem o resultado dessa contagem.
7. Conte quantas palavras distintas sem vogais existem no texto.

- Saída esperada: Um arquivo que contem o resultado dessa contagem.

8. Indique quais são os tipos (elementos do vocabulário) que as palavras sem vogais pertencem. ´ • Saída esperada: Um arquivo que contem todos os tipos retornados.

9. Conte quantas palavras que contem exatamente uma sequência de vogais. ^ • Saída esperada: Um arquivo que contem o resultado dessa contagem.

Entrega:

Um arquivo de texto simples com todos os comandos utilizados para responder as questões acima, bem como as saídas geradas por cada comando. O arquivo com os comandos deve ser enviado pelo Classroom, mas as saídas dos comandos podem estar armazenadas em um link do Google Drive, desde que o mesmo esteja fornecido no arquivo com os comandos.

E importante que o arquivo de entrada indicado anteriormente não seja modificado. 2