UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
Departamento de Ciências de Computação

SCC0633/5908 PROCESSAMENTO DE LINGUAGEM NATURAL
LISTA DE EXERCÍCIOS 2

1. Escolha um corpus qualquer (pode ser o que foram utilizados na aula) e, usando as funções do NLTK, faça:
    a. *Tokenize* o corpus inteiro (palavras, números e pontuações)
    b. Verifique a quantidade de tokens do corpus
    c. *Tokenize* o corpus apenas por suas palavras
    d. Verifique a quantidade de palavras do corpus
    e. Verifique a frequência de palavras no corpus
    f. Verifique quais são as 5, 10 e 15 palavras mais frequentes do corpus
    g. Extraia as *stopwords* do NLTK (não do corpus ainda)
    h. Verifique a frequência dos tokens sem *stopwords* do corpus
    i. Extraia todos os bigramas do corpus
    j. Extraia todos os trigramas do corpus
    k. Extraia todos os 4-gramas do corpus
    l. Retorne as entidades nomeadas do corpus, usando os bigramas e trigramas
    m. Escolha 3 palavras do seu corpus e faça o *stemming* delas
    n. Separe uma sentença do seu corpus e retorne todas as classes gramaticais das palavras da sentença. Analise se o etiquetador acertou todas as classes gramaticais
    o. Retorne as classes gramaticais de todas as palavras do seu corpus
    p. Retorne as entidades nomeadas do seu corpus, usando a técnica de *chunking*

2. Utilizando um corpus de seu interesse (pode ser o que foi utilizado na aula) e utilizando o spaCy, faça:
    a. Tokenize todo o corpus
    b. Tokenize todo o corpus e retorne a lista com strings
    c. Tokenize todo o corpus e retorne apenas as palavras
    d. Retorne a quantidade de palavras do corpus
    e. Tokenize todo o corpus e retorne apenas os números
    f. Retorne a quantidade de números do corpus

g. Tokenize todo o corpus e retorne apenas as pontuações
h. Retorne a quantidade de pontuações do corpus
i. Retorne a frequência de pontuação do corpus (obs: pode usar o NLTK nesse exercício!)
j. Tokenize todo o corpus e retorne a quantidade de espaços presentes no corpus

3. Ainda utilizando o mesmo corpus e o spaCy, faça:
    a. Analise o 1º parágrafo do corpus, lematizando todos os tokens possíveis. Verifique se foi retornado os lemas corretos
    b. Ainda com o 1º parágrafo, retorne todas as classes gramaticais dos tokens.
    c. Retorne todas as classes gramaticais do corpus