Week 9 NLTK NLP

Tassos Karageorgiadis

Step 1:

> Import nltk library and download books from Gutenberg corpus.

Question A:

Step 2

> Get a list with all book-titles of Gutenberg and keep only the first 10.

Question B:

Step 3

> Iterate through books and read their raw text content.

Step 4

> Use word_tokenizer to split each book content to words/tokens and keep the unique set for each book into one list -- *vocabulary_per_book*--.

Step 5

> Because the books may have the same words in their vocabulary, we keep only the unique one of the previous step's list of words -- *vocabulary_final*--.

Question C:

Step 6

> For each book's sentences add a token at the start and at the end of the sentence.

Question D:

Step 7

> Use ngram() function to find unigrams,bigrams,and trigrams & Calcualte their Frequencies

Question E:

Step 8

> Generate random sentences of bigrams with legnth = 15, sentences are stored inside ***bigram_senteces.txt***

Step 9

> Couldn't Generate random sentences of trigrams with legnth = 15 sentences Empty list in cfdist in generate_model for trigrams

Resources

- Text processing with NLTK
- Ngrams
- Calculate Frequencies
- Gutenberg.words VS word_tokenize