Assignment 3- Tokenization using Python

Using "carroll-alice.txt"(Alice's Adventures in Wonderland by Lewis Carroll 1865) from Project Gutenberg to tokenization, compute the word frequencies and show the top 10 common words.

1. Tokenization & Remove stopword without Stemming/Lemmatization.

(a) Read text from file carroll-alice.txt"

(b) Normalize words: convert all upper-case letter to lower case (for example 'Word' and 'word' are considered as the same word).

(c)Tokenization: using nltk word_tokenize.

(d) Remove stopwords using stopwords from nltk. You can add more stopword if needed.

(e) Count the occurrence of words and display the top 10 common words. (the result will show the most common word and the counting value of this word in the document)

(*) Note that you need to remove "punctuation mark" token such as ","/ "." by using isalnum() function to remove punctuation mark after tokenization. Or you can remove the punctuation mark before tokenization by using re.sub() function.

2. Tokenization & Remove stopwords with Stemming/Lemmatization.

Do Tokenization and remove stopwords (similar to the steps in question 1), but add the step Stemming or Lemmatization before computing word occurrence. You need to show up 2 results (top 10 common words): one in case of using Stemming and one in case of using Lemmatization.

(*) Use **Porter** or Lancaster **stemming** and Lemmatization function from nltk package.

(*) Optional: you can use Lemmatization using the spacy package.

3. Compare the result of question 1 (without steaming) and question 2 (with stemming and with Lemmatization). Are the results different? Try to explain why it's different, or give your opinion about the result. In your opinion, which way is the best?

Submit your homework & Hint

- Students submit your file ".ipynb" to iLMS. If you use google colab, you can go to menu File/download .ipynb to download your python notebook file.
- The Submit file will contain your code and your explanation (in question 3). Using "add Text cell" to add text cell and write your answer directly to the ".ipynb" file.
- You can use my sample code in Appendix A1 (file week3_text to numerical vector.ipynb) or you can write your code by your self.
- You're welcome if you want to do this homework in a different way than I have shown in class.
- Any Problem/questions, you can write an email to me. (Daniel, cuongtv@iss.nthu.edu.tw).