IT550: Information Retrieval

Assignment 1: Preprocessing in IR

Instructor: Prasenjit Majumder

Learning Outcome: After completing this assignment you will be able to preprocess your documents before performing IR tasks.

1 Problem description

Data preprocessing is an essential step for building robust and reliable machine learning models. In IR, preprocessing consists of a pipeline which converts the corpus into a format that allows a model to efficiently solve a given task.

2 Implemetation

2.1 Dataset

- We will be using the dataset provided in the following tutorial for our implementation: https://drive.google.com/file/d/1CuxS5YVKaKzt_1kdHsspZPZ57tcV3uS_/view?usp=sharing
- The dataset contains text in <TEXT> tag.

2.2 Exercise

The primary objective of this assignment is to help you gain familiarity with loading and cleaning text data using python modules like Pandas, NLTK, RegEx etc.

1. Read and combine the text between the <TEXT> tags by using Beautiful Soup from all the documents.
2. Remove punctuation and numerical values from your text.
3. Perform tokenization on the text. There are two ways to perform tokenization here. One way is to split feature of string and another is using NLTK library. Display the difference between both approaches.
4. Convert the text to lowercase
5. Remove stop-words from the text
6. Perform Stemming (use Porter Stemmer) and Lemmatization (separately) and display the differences that you observe while using the two approaches. (For this you have to use NLTK library).

3 References

- NLTK documentation: https://www.nltk.org/
- NLTK tutorial: https://www.youtube.com/watch?v=X2vAabgKiuM
- RegEx tutorial: https://www.w3schools.com/python/python_regex.asp
- Pandas tutorials:
    1. https://data36.com/pandas-tutorial-1-basics-reading-data-files-dataframes-data-selection/
    2. https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html#min
    3. https://pandas.pydata.org/pandas-docs/stable/user_guide/cookbook.html#cookbook
- https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

4 Submission

- For this assignment you have to perform each step mentioned in the exercise and display a small extract of the text from your dataset.
- Create your assignment in Google colab and explain your approach and results clearly.
- You have to submit the assignment on Google Classroom
- **Assignment is due on 26th January 2020 at 11 PM**