CS 582 Information Retrieval

Pranay Kumar Rasulury
prasul2@uic.edu

Program details:

- For each file in the "citeseer" directory, retrieving the files as string and applying lower and split methods. Then stored them in a list
- For each word in list, removed punctuations with regular expressions
    - **remove_specical_characters(string)**: this method takes a string and removes punctuation and returns it as a string
- Loading stop words list from stopwords.txt and filtering the stop words from the list
    - **load_stop_words():** from the file "stopwords.txt" extracting all the stop words and returning as a list of strings
- Porter stemmer algorithm from nltk library is used to stem words from the list and are stored in the dictionary with frequencies
- All the specified questions are then calculated
    - **find_stats(dictionary):** This method takes a dictionary containing words as keys and frequency as value, then calculates all the questions mentioned in the homework

Instructions to execute:

- Run the command "python hw1.py" in the terminal

Answers:

(2a) **Total no. of Words:** 476203

(2b) **Total vocabulary size:** 19889

(2c) **Top 20 words**

1. the	8. is	15. with
2. of	9. we	16. as
3. and	10. that	17. by
4. a	11. this	18. data
5. to	12. are	19. be
6. in	13. on	20. information
7. for	14. an	

(2d) **Stop words from the top 20 words:**

the, of, and, a, to, in, for, is, we, that, this, are, on, an, with, as, by, be

(2e) 4 words account for 15% or more of the total number of words in the collection

Stats after removing stop words and applying Porter stemmer algorithm

(2a) **Total no. of Words:** 265781

(2b) **Total vocabulary size:** 13559

(2c) **Top 20 words**

1. system	8. user	15. present
2. data	9. learn	16. base
3. agent	10. algorithm	17. web
4. inform	11. 1	18. databas
5. model	12. approach	19. comput
6. paper	13. problem	20. method
7. queri	14. applic	

(2d) No Stop words from the top 20 words

(2e) 22 words account for 15% or more of the total number of words in the collection