CS 4390/5390: Information Retrieval and Information Visualization
Fall 2019

Assignment# 1

Due: September 22, 2019 in Blackboard

We have used inverted index to answer Boolean queries. Let's practice more. For a set of given input files, write a program to create an inverted index and respond to user queries.

You will be given:

1. Documents to be indexed in a *data* directory. Each document in that directory will be a represented using a txt file.
2. User queries in a *query.txt* file. Each query will be composed of two query terms joined by either an **AND** or **OR**.

Your tasks are:

1. Write a program that generates an inverted index from data files and writes the index in a file named *index_lastname.txt* (*lastname* should be replaced by your last name).
    a. The file should have the following format

    ```
    t1      d1, d2, d3
    t2      d2, d4
    t3      d1, d5, d6
    …       ……
    tn      d1, d8, di
    ```

    where t1 is term1 and d1 is document1.

    b. It is up to you to decide how you want to implement the dictionary and postings. Efficiency is not the prime concern for this assignment.

3. Using the queries found in the *query.txt* file, the program should return the document IDs for the successful queries. Write the results in a file named *answer.txt.* Use and update the intersection algorithm introduced in the class. The query file should have the following format

    ```
    q1      d1, d3
    q2      d1
    q3      -1
    ```

    where q1 is the first query.

    a. Return -1 for unsuccessful queries.

2. You may use the porter stemmer from the NLTK package.
3. Do not use any stopword list. That is, do not remove the stop words.

Deliverables: 3 files

1. The “very well-documented” source code of the program. The name of the file should be lastname_firstname.xyz (replace xyz with proper extension).
2. The inverted index, named as *index_lastname.txt*.
3. The query results in the *answer.txt* file.

Resources:

1. **Porter Stemmer:** https://tartarus.org/martin/PorterStemmer/
2. **NLTK package:** https://www.nltk.org