ITCS 4111/5111 Introduction to Natural Language Processing

Assignment 4 Due: November 17th, 2019 11:59 pm

Total points: 100

Your submissions should be in a single python notebook. Make sure code is documented well. All submissions should be uploaded to Canvas. Email submissions will not be considered for grading.

The problems marked with ** are optional for students enrolled in the ITCS 4111 section of this course, all other problems are mandatory. ALL problems are mandatory for the students enrolled in ITCS 5111.

The learning goals of this assignment are:

- how to build inverted indexes
- how to implement Boolean information retrieval
- adding weighting e.g. TF-IDF to IR algorithm

Information Retrieval

Problem 1 (40 points): We have given you a small set of data in the form of tweets. Each line in the file begins with a document ID, followed by the text of the tweet. Implement a function to create an inverted index of these documents. You may use the approach described in the Manning book (link below), or you may come up with your own algorithm.
https://nlp.stanford.edu/IR-book/html/htmledition/a-first-take-at-building-an-inverted-index-1.html

Problem 2 (40 points): Write a function to implement the merge algorithm as discussed in class. Your code should allow intersecting the postings of two terms, as well as process simple Boolean queries. When there are multiple query terms, make sure that your algorithm uses the optimization described in Manning book of performing the most restrictive intersection first.
https://nlp.stanford.edu/IR-book/html/htmledition/processing-boolean-queries-1.html

Problem 3* (20 points):** Extend the system from Problem 2 to perform simple TF-IDF scoring of the retrieved results. There is no need to worry about any weight normalizations.