Cmpe 493 Introduction to Information Retrieval, Spring 2020
Assignment 3 - Extractive Text Summarization for COVID-19
Due: 26/06/2020 (Wednesday), 23:59

In this assignment, you will implement a text summarization model similar to the LexRank algorithm. LexRank is an unsupervised model which computes the relative importance of the sentences based on the cosine similarity and the PageRank algorithm. The details of the algorithm are available at the following link.

`https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume22/erkan04a-html/erkan04a.html`

As the data set we will use articles from the COVID-19 Open Research Dataset (CORD-19) (`https://www.semanticscholar.org/cord19`) and relevance annotations provided by TREC-COVID Challenge (`https://ir.nist.gov/covidSubmit/`). We will use the Round 1 Data from `https://ir.nist.gov/covidSubmit/data.html`.

You should perform the following steps.

- Download the articles from (`https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases/cord-19_2020-04-10.tar.gz`). These articles are from the 10 April 2020 release of CORD-19 corpus. The size of the data is around 1.5GB.
- Download the list of topics from (`https://ir.nist.gov/covidSubmit/data/topics-rnd1.xml`). The file is in XML format and contains a list of topics and their descriptions.
- Download the relevance judgements from (`https://ir.nist.gov/covidSubmit/data/qrels-rnd1.txt`). Each line contains the relevance judgement for a document. The first column is the topic-id, the second field is iteration (you will NOT be using this field), the third column is the document-id (cord-id), and the last column is the relevance judgement, where 0 means not-relevant, 1 means partially relevant and 2 means fully relevant.
- We will use the documents of three topics, Topic 1: coronavirus origin, Topic 13: how does coronavirus spread, and a topic of your choice. Topic 1 tries to answer the question "what is the origin of COVID-19?", which can be described as "seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans". Topic 13 tries to answer the question "what are the transmission routes of coronavirus?", which can be described as "Looking for information on all possible ways to contract COVID-19 from people, animals and objects".
- You should compute the IDF scores of the terms by using all the documents in the downloaded corpus.

- Then, you should write a document summarization system that operates in two steps. Given a topic, in the first step, the system will identify the most salient 10 documents for that topic. In the second step, the 10 selected documents will be summarized by selecting the most important 20 sentences from among all the sentences in these 10 documents. So, for each topic, you will obtain a 20 sentence summary.
- To identify the most important documents in a given topic, you should create a document graph, where each node corresponds to a document and each edge represents the TF-IDF weighted cosine similarity between the ABSTRACTS of the corresponding two documents. Implement the PageRank algorithm and run it on this document graph to identify the 10 documents with the highest PageRank scores. For each topic, you will use the documents that have a relevance score of 2 in the document-relevance file (`https://ir.nist.gov/covidSubmit/data/qrels-rnd1.txt`).
- After identifying the most important 10 documents for a topic, you should create a sentence graph, where each node corresponds to a sentence (in the 10 documents) and each edge represents the TF-IDF weighted cosine similarity between the corresponding two sentences. Implement the PageRank algorithm and run it on this sentence graph to identify the 20 sentences with the highest PageRank scores. These 20 sentences will be the summary of the given topic.
- You should choose the cosine similarity threshold (t) as 0.10 to create arcs between two sentences/documents. In other words, if the cosine similarity between two sentences/documents is less than t, these sentences are not connected to each other (i.e., the corresponding entry in the adjacency matrix is 0), otherwise they are connected to each other in the graph (i.e., the corresponding entry in the adjacency matrix is 1) . Also, the teleportation rate should be set to 0.15 and the error tolerance (ϵ) in the power method should be set to 0.00001.

You can use numpy for matrix operations. You can use an equivalent library to numpy for other programming languages for the matrix operations. You can use NLTK or any other library for sentence boundary detection, tokenization, and XML processing. Other than these, you are not allowed to use any third party libraries.That is, everything else (including cosine similarity computation and the Power Method for PageRank) should be your own implementation. You may use any programming language of your choice. However, we should be able to run your program by following the instructions in your readme file.

Submission: You should submit a *".zip"* file named as YourNameSurname.zip containing the following files using the Moodle system:

1. Report:
   (i) Describe any assumptions or choices that you made while implementing your summarization system.
   (ii) For each of the selected three topics: Provide the IDs and the PageRank scores of the top 10 documents.
   (iii) For each of the selected three topics: Provide the top 20 sentences as well as their PageRank scores.

2. Source code and executable: Commented source code and executables of your summarization system.

3. Readme: Detailed readme describing how to run your program.

Late Submission: You are allowed 7 late days (one week) for this assignment with no late penalty. After 7 days, 10 points will be deducted for each late day (if you have difficulty in completing the assignment due to any reason, please contact me).