INFO 300 – Assignment 1 (8 points)

Weimao Ke, Drexel University September 17, 2020

1 Objectives

This assignment is for students to get started with the concept of inverted indexing and initial software setup for information retrieval. Please use the provided **A1_wk77_template.md** template and follow the steps below to prepare your answers in the **Markdown** format.

2 Data

Consider a small collection of 4 text documents (titles only):

1. breakthrough drug for schizophrenia
2. new schizophrenia drug
3. new approach for treatment of schizophrenia
4. new hopes for schizophrenia patients

Note the numbers $1, 2, 3, 4$ are document IDs and not part of the text contents.

3 Tasks

3.1 Term-Document Matrix Representation

Please draw the term-document incidence matrix (with binary representation) for this document collection.

3.2 Inverted Index Representation

Please draw the inverted index representation for the document collection (similar to Figure 1.3. on page 6 of the text book).

3.3 Indexing with ElasticSearch

With the *Kibana Dev Tools Console* or *Curl*, put the above Schizophrenia documents into an index called **YourUserName_info300_schizophrenia** on the ElasticSearch cluster (via https://tux-es1.cci.drexel.edu:5601).

Note:

- Use a field such as **content** for the text information and do NOT include doc 1, doc 2, etc. in the text.
- Instead, use each number as the **document ID** in the index.

Once indexing requests are successfully processed, copy and paste **your request/command** (in JSON) to your Markdown file.

3.4 Retrieval with ElasticSearch

Run a search command to retrieve all documents and make sure all four documents have been posted to the above Schizophrenia index. Copy and paste your: 1) **search request** and 2) **response/results**.

3.5 Boolean Query

What are the returned documents for following boolean queries?
"schizophrenia" **AND** "drug"

3.5.1 Manual Analysis

Answer the question first with your own analysis and write down your result.

3.5.2 ElasticSearch Query

Write a request to query the above index on ElasticSearch to verify your answer. Copy and paste the complete search **request and response**.

3.6 Compound Query

What are the returned documents for following boolean queries?
"for" **AND** ("drug" **OR** "approach")

3.6.1 Manual Analysis

Again, answer the question first with your own analysis and write down your result.

3.6.2 ElasticSearch Query

Write a request to query the above index on ElasticSearch to verify your answer. Copy and paste the complete search **request and response**.

4 Assignment Submission

Please submit all your answers in a Markdown file. Make sure:

1. File name has to be **A1_YourDrexelID.md**. If you pair up for the assignment, include both students' IDs in the name: **A1_id1_id2.md**.
2. All data, code, and answers should be **properly numbered and formatted** according to the template.