Information Retrieval

Assignment 1

Due: Next section.

Note: for task 2, you should present a running program for the lab instructor.

---

1. (5 Points) Assume you have a set of documents with the following contents:

File name	Contents
Big Data	Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
SAS	Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured.
IBM Big Data	Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.

a. **Create a Boolean matrix table for the previous set of documents.**
b. **Use the Boolean retrieval technique to find all the documents contains the word "data", and "Processing".**
c. **Construct an Inverted Index for the same table.**
d. **Which technique (Boolean matrix, Inverted Index) is better, Why?**

Note: you can omit common words such as (a, is, the, …, etc.)

2. (10 Points) Write a java program named (Tokenizer), that take a text file as input. and produce another text file contains all the "distinct terms" in the input file such that each term in a separated line with its frequency. For example:

Input file: (Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. To extract meaningful value from big data.**)**

Output file:

Big 2

Data 2

Is 1

….