CS 4960/6550: Introduction to Information Retrieval

Assignment 1

Instructor: Qingyao Ai *Teaching Assistant: Tao Yang*

Section 1: Evaluation Metrics

Suppose that we have a corpus with 6 documents: $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{E}$ and $\boldsymbol{F}$. Now consider a case where a user submits a query $\boldsymbol{q}$ and the goal is to retrieve 4 documents in response to the query. Here are five possible ranked lists:

- List 1: $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{E}$, $\boldsymbol{D}$
- List 2: $\boldsymbol{C}$, $\boldsymbol{B}$, $\boldsymbol{F}$, $\boldsymbol{D}$
- List 3: $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{C}$, $\boldsymbol{F}$
- List 4: $\boldsymbol{A}$, $\boldsymbol{D}$, $\boldsymbol{C}$, $\boldsymbol{E}$
- List 5: $\boldsymbol{F}$, $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{B}$

Question 1.1 (25 points)

Suppose that $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{E}$ are relevant to the query $\boldsymbol{q}$ while the others are not. Compute the Average Precision (AP) for each list. (5 points for each)

Question 1.2 (25 points)

Suppose that we have five-level relevance judgments (from 0 to 4) for each query-document pair, where 0 means *irrelevant* and 4 means *perfectly relevant*. The relevance level of $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{E}$ and $\boldsymbol{F}$ with respect to the query $\boldsymbol{q}$ are 4, 0, 3, 0, 1 and 0, respectively. Compute the Normalized Discounted Cumulative Gain (nDCG) for each list. (5 points for each, show DCG and IDCG respectively)

Section 2: Text Processing and Indexing

Question 2.1 (20 points)

Consider the following sentense:

> *According to Wikipedia, information technology is the use of computers to create, process, store, and exchange all kinds of electronic data and information..*

Prior to indexing, we want to conduct tokenization, normalization, stopping, and Krovetz stemming. Please explain how each technique changes the sentence (4 points for each), and write down the final sentence after applying all these techniques (4 points). For stopping, please explain which words have been treated as stopwords and why.

Question 2.2 (10 points)

As we know, inverted indexing is one of the key techniques in modern search engines. To understand why it is important, please answer the following questions:

- What is advantage of searching with inverted index comparing to searching by sequentially reading each document? (5 points)
- Does inverted indexing improve the efficiency of a search system in all cases? If so, explain why; if not, give an example (5 points).

Question 2.3 (20 points)

Bitwise compression is frequently used in index compression. Two of the common bitwise compression methods are γ-code and δ-code.

- Encode the number 2021 with both γ-code and δ-code (10 points).
- Consider two codes 000010100 and 001010101. They are either encoded in γ-code or δ-code. Identify their coding schemes and decode their value after decoding (10 points).