IR-assignment 1

Yulan Wang, uID: u1264235

February 10, 2020

Question 1.1:

- AP(list 1) = $\frac{1}{3} \times (1 \times 1 + \frac{2}{3} \times 1) = \frac{5}{9}$
- AP(list 2) = $\frac{1}{3} \times (\frac{1}{2} \times 1) = \frac{1}{6}$
- AP(list 3) = $\frac{1}{3} \times 1 = \frac{1}{3}$
- AP(list 4) = $\frac{1}{3} \times (1 \times 1 + \frac{2}{4} \times 1) = \frac{1}{2}$
- AP(list 5) = $\frac{1}{3} \times (\frac{1}{2} \times 1 + \frac{2}{4} \times 1) = \frac{1}{3}$

Question 1.2:

- nDCG (list 1) = $(4 + \frac{1}{\log_2 4})/(4 + \frac{1}{\log_2 3}) = 0.97$
- nDCG (list 2) = $(\frac{2}{log_2 3})/(2) = 0.63$
- nDCG (list 3) = $1/1 = 1.0$
- nDCG (list 4) = $(4 + \frac{1}{\log_2 5})/(4 + \frac{1}{\log_2 3}) = 0.96$
- nDCG (list 5) = $(\frac{4}{log_2 3} + \frac{2}{log_2 5})/(4 + \frac{2}{\log_2 3}) = 0.64$

Question 2.1:

- *Tokenization:* Split the text into individual tokens, like words, name entities, and email address, etc.
  According | to | Wikipedia | Information | Retrieval | is | the | activity | of | obtaining | information | resources | relevant | to | an | information | need | from | a | collection | of | information | resources
- *Normalization:* Transforming text into a single canonical form that it might not have had before, like lower-casing, whitespaces, and numbers.
  according | to | wikipedia | information | retrieval | is | the | activity | of | obtaining | information | resources | relevant | to | an | information | need | from | a | collection | of | information | resources
- *Stopping:* Removing stopwords, including "to", "is", "the", "of", "an", "a", and "from". These words are very frequent, and we couldn't get any information for them.
  according | - | wikipedia | information | retrieval | - |- | activity | - | obtaining | information | resources | relevant | - | - | information | need | - |-| collection | - | information | resources
- *Krovetz stemming:* reducing inflected (or sometimes derived) words to their word stem based on dictionary.
  The final sentence after all these techniques:
  accord | - | wikipedia | inform | retrieve | - |- | act | - | obtain | inform | resource | relevant | - | - | inform | need | - |-| collect | - | inform | resource

Question 2.2a:

- First, efficiency. Inverted indexing could retrieve any desired subset of records, and doesn't have to retrieve all main record files.
- Second, more information could be recorded. Inverted indexing could contain more information, like appear frequency of the words, positions of the words. All these information could encoded in inverted indexing.

Question 2.2b:

- No. If we want to search some queries in some specific document not all documents, then inverted indexing couldn't improve the efficiency.

Question 2.3a:

- γ-code: $x_d = 9$ and $x_r = 134$, code: $0000000001, 010000110$.
- δ-code: $x_d = 9$, $x_{dd} = 3$, $x_{dr} = 2$, and $x_r = 134$ code: $0001, 010, 010000110$.

Question 2.3b:

- 0001010 is coded by γ-code. From left to right, there are three zeros, so the unary coding number is 3. Then reading three bins after 1, there are only three number left which is binary code, and the number is 2. Therefore, the code number is: $2^3 + 2 = 10$.
- 001010101 is coded by δ-code. There are two zeros in the beginning, so the unary coding number is 2. Then reading two bins after 1, the binary coding number is 1. Then reading all remaining bins 101, the binary coding number is 5. So, the δ-coding number is: 21.

Task 1:

Node	PageRank Score
1	0.26391561
2	0.16940098
3	0.02857143
4	0.14896637
5	0.24955901
6	0.05142857
7	0.08815803

Task 2:

- There are 10429 documents in the corpus.
- The average length of documents is 1509.52.
- The number of unique words in the corpus is 92287.
- The document id B009RXU59C of the longest document and the document length is 61311.
- The number of documents that contain the word information" is 727.