ASSIGNMENT 1 SOLUTION (INFORMATION RETRIEVAL)

HASAN POURMAHMOODAGHABABA
UID: U1255635
H.POURMAHMOODAGHABABA@UTAH.EDU

1. Evaluation Metrics

Question 1.1.

- List 1: $\frac{1}{3}(1+1+1+0) = \frac{3}{3} = 1$
- List 2: $\frac{1}{3}(1+0+0+0) = \frac{1}{3} = 0.3333$
- List 3: $\frac{1}{3}(1+0+\frac{2}{3}+0) = \frac{5}{9} = 0.5556$
- List 4: $\frac{1}{3}(1+0+\frac{2}{3}+\frac{3}{4}) = \frac{29}{36} = 0.8056$
- List 5: $\frac{1}{3}(0+\frac{1}{2}+\frac{2}{3}+0) = \frac{7}{18} = 0.3889$

Question 1.2.

$iDCG = 4 + \frac{3}{\log_2 3} + \frac{1}{\log_2 4} = 6.3928$

- List 1: $DCG = 4 + \frac{3}{\log_2 3} + \frac{1}{\log_2 4} + \frac{0}{\log_2 5} = 6.3928 \Longrightarrow nDCG = \frac{DCG}{iDCG} = \frac{6.3928}{6.3928} = 1$
- List 2: $DCG = 3 + \frac{0}{\log_2 3} + \frac{0}{\log_2 4} + \frac{0}{\log_2 5} = 3 \Longrightarrow nDCG = \frac{DCG}{iDCG} = \frac{3}{6.3928} = 0.4693$
- List 3: $DCG = 1 + \frac{0}{\log_2 3} + \frac{3}{\log_2 4} + \frac{0}{\log_2 5} = 2.5 \Longrightarrow nDCG = \frac{DCG}{iDCG} = \frac{2.5}{6.3928} = 0.3911$
- List 4: $DCG = 4 + \frac{0}{\log_2 3} + \frac{3}{\log_2 4} + \frac{1}{\log_2 5} = 5.9307 \Longrightarrow nDCG = \frac{DCG}{iDCG} = \frac{5.9307}{6.3928} = 0.9277$
- List 5: $DCG = 0 + \frac{4}{\log_2 3} + \frac{3}{\log_2 4} + \frac{0}{\log_2 5} = 4.0237 \Longrightarrow nDCG = \frac{DCG}{iDCG} = \frac{4.0237}{6.3928} = 0.6294$

2. Text Processing and Indexing

Question 2.1.

- tokenization: We split the sentence by space and remove symbols like comma, dot, semicolon.

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

- normalization: We change all words to the same written shape like changing them to lowercases.

according	to	wikipedia	information	technology	is	the	use
of	computers	to	create	process	store	and	exchange
all	kinds	of	electronic	data	and	information	

- stopping: We remove stop words like "a, an, the, and, of, to, is, ...".

according	–	wikipedia	information	technology	–	–	use
–	computers	–	create	process	store	–	exchange
all	kinds	–	electronic	data	–	information	

- Krovetz stemming: We change all words to their "algorithmic + dictionary-based" versions, like plural to singular and normalizing verb tense.

accord	–	wikipedia	inform	technology	–	–	use
–	computer	–	create	process	store	–	exchange
all	kind	–	electronic	data	–	inform	

I treated the words {the, and, of, to, is} as stopping as they don't cary a useful information and are frequent.

Question 2.2.

- Inverted index is efficient for big cuprous' queries where reading each document sequentially can be inefficient.
- It depends on data. In fact if data we are exploring is not a big data, I assume there would be no considerable improvement; however, it would have a significant improvement in time efficiency for big data sets. For example, for a query to search within all webpages (like search in Google) inverted index has a dramatic improvement in the search system.

Question 2.3.

- - $x_d = \lfloor \log_2 2021 \rfloor = 10$ and so its unary code is 00000000001,
  - $x_r = x - 2^{\lfloor \log_2 2021 \rfloor} = 2021 - 2^{10} = 997$ which has the binary code 1111100101,
  - So the γ**-code** of 2021 is $00000000001, 1111100101$.
- - $x_d = \lfloor \log_2 2021 \rfloor = 10$,
  - $x_{dd} = \lfloor \log_2(10+1) \rfloor = 3$, and so its unary code is 0001,
  - $x_{dr} = (x_d + 1) - 2^{x_{dd}} = (10+1) - 2^3 = 3$ which has the binary code 11,
  - $x_r = x - 2^{x_d} = 2021 - 2^{10} = 997$ with binary code 1111100101,
  - So the δ**-code** of 2021 is $0001, 11, 1111100101$.
- - 000010100 is encoded in γ-code and we can decode it to get $2^4 + 2^2 = 20$.
  - 001010101 is encoded in δ-code and we can decode it to get $2^4 + (2^2 + 1) = 21$.

    Calculations are as follows:

    $x_{dd} = (001)_{\text{unary}} = 2 \Rightarrow 1 = (01)_2 = x_{dr} = (x_d + 1) - 2^{x_{dd}} = x_d + 1 - 2^2 \Rightarrow x_d = 4 \Rightarrow x = x_r + 2^{x_d} = (0101)_2 + 2^4 = 5 + 2^4 = 21$.

School of Computing, University of Utah, Utah, USA
Email address: h.pourmahmoodaghababa@utah.edu