CS 6550: Introduction to Information Retrieval

Assignment 1

UID: student_id *Student Name: student_name*

Question 1:

Question 1.1

$Average\ Precision = \frac{1}{|R|}\Sigma_{i=1}^{n} prec(i) rel(i)$

List 1: *A, C, E, D*

$ap = \frac{1}{3}\left(1 + 0 + \frac{2}{3} + 0\right) = \frac{1}{3}\left(\frac{5}{3}\right) = \frac{5}{9}$

List 2: *C, B, F, D*

$ap = \frac{1}{3}\left(0 + \frac{1}{2} + 0 + 0\right) = \frac{1}{6}$

List 3: *E, D, C, F*

$ap = \frac{1}{3}\left(1 + 0 + 0 + 0\right) = \frac{1}{3}$

List 4: *A, D, C, E*

$ap = \frac{1}{3}\left(1 + 0 + 0 + \frac{2}{4}\right) = \frac{1}{3}\left(\frac{6}{4}\right) = \frac{6}{12} = \frac{1}{2}$

List 4: *F, A, C, B*

$ap = \frac{1}{3}\left(0 + \frac{1}{2} + 0 + \frac{2}{4}\right) = \frac{1}{3}\left(1\right) = \frac{1}{3}$

Question 1.2

Normalized Discounted Cumulative Gain

$$nDCG = \frac{DCG}{IDCG}$$

where DCG is the Discounted Cumulative Gain

$$DCG = \Sigma_{i=1}^{n} \frac{rel(i)}{\log_2(i+1)}$$

and IDCG is the Ideal Discounted Cumulative Gain.

The ideal list for this query would be ***A, B, E, C***

$$IDCG = 4 + \frac{2}{\log_2(3)} + \frac{1}{\log_2(4)} = 5.7618$$

List 1: *A, C, E, D*

$nDCG = \frac{4+0+\frac{1}{\log_2(4)}+0}{5.7618} = \frac{4.5}{5.7618} = 0.781$

List 2: *C, B, F, D*

$nDCG = \frac{0+\frac{2}{log_2(3)}+0+0}{5.7618} = \frac{1.262}{5.7618} = 0.219$

List 3: *E, D, C, F*

$nDCG = \frac{1+0+0+0}{5.7618} = \frac{1}{5.7618} = 0.174$

List 4: *A, D, C, E*

$nDCG = \frac{4+0+0+\frac{1}{\log_2(5)}}{5.7618} = \frac{4.4307}{5.7618} = 0.769$

List 4: *F, A, C, B*

$nDCG = \frac{0+\frac{4}{\log_2(3)}+0+\frac{2}{\log_2(5)}}{5.7618} = \frac{3.3851}{5.7618} = 0.588$

Question 2:

Question 2.1

Tokenization

Tokenization splits the sentence into individual tokens. Tokens can be words, name entities (i.e. people's names, city names, etc), or Email Addresses, URLs, etc. Punctuation's are also removed. Below each token is in its own cell.

According	to	Wikipedia	Information	Retrieval	is	the
activity	of	obtaining	information	resources	relevant	to
an	information	need	from	a	collection	of
information	resource					

Normalization

Normalization transforms text into a single canonical form such a lower-casing or removing whitespace.

according	to	wikipedia	information	retrieval	is	the
activity	of	obtaining	information	resources	relevant	to
an	information	need	from	a	collection	of
information	resource					

Stopping

Stopping removes stopwords. Stopwords appear frequently in text, but are usually uninformative. The words 'to', 'is', 'the', 'of', 'an', 'from', and 'a' were chosen as stopwords because they appear frequently in text and are non-informative.

according		wikipedia	information	retrieval		
activity		obtaining	information	resources	relevant	
	information	need			collection	
information	resource					

Krovetz Stemming

Stemming reduces inflected or derived words to their word stem, base or, root, i.e. plural to singular and normalizes verb tense.

accord		wikipedia	information	retrieveal		
activity		obtain	information	resource	relevant	
	information	need			collection	
information	resource					

Question 2.2

What is advantage of searching with inverted index compared to searching by sequentially reading each document?

The time, memory, and processing is much lower; and in the case of the web or large databases of documents the inverted index makes the search feasible.

Does an inverted index improve the efficiency of a search system in all cases? If so, explain why; if not, give an example.

No, the updates time (adding a document) for an inverted index is much slower.

Question 2.3

Encode the number 646 with both γ-code and δ-code

γ-code

$$x_d = \lfloor \log_2(646) \rfloor = 9$$
$$x_r = 646 - 2^{\lfloor \log_2(646) \rfloor} = 134$$

Representing x_d as unaray and x_r as binary we get: 0000000001 010000110.

δ-code

$$x_d = \lfloor \log_2(646) \rfloor = 9$$
$$x_{dd} = \lfloor \log_2(9+1) \rfloor = 3$$
$$x_{dr} = (9+1) - 2^{\lfloor \log_2(9+1) \rfloor} = 10 - 2^3 = 2$$
$$x_r = 646 - 2^{\lfloor \log_2(646) \rfloor} = 646 - 2^9 = 134$$

Representing x_{dd} as unaray and x_{dr} and x_r as binary we get: 0001 010 10000110.

Determine encoding and decode 0001010 and 001010101

0001010

This is γ-code.

$$x_d = 0001 = 3$$
$$x_r = 010 = 2$$
$$2 = x - 2^3$$
$$x = 10$$

001010101

This is δ-code

$$x_{dd} = 001 = 2$$

$$x_{dr} = 01 = 1$$

$$x_r = 0101 = 5$$

Solving for x_d

$$2 = \log_2(x_d + 1)$$

$$2^2 = x_d + 1$$

$$4 - 1 = x_d$$

$$x_d = 3$$

Solving for x

$$5 = x - 2^3$$

$$5 = x - 8$$

$$5 + 8 = x$$

$$x = 13$$

Experimental Question

Task 1: PageRank

Document ID	PageRank Score
1	1.03773405
2	0.62264043
3	0.0
4	0.41509362
5	0.83018724
6	0.0
7	0.20754681