CS378 Information Retrieval and Web Search: Midterm Exam Solution

Oct. 18, 2018

NAME: ____

Be sure to show your work on all problems in order to allow for partial credit.

1. (14 points) Assume that simple term frequency weights are used (no IDF factor), and the only stopwords are: "is", "am" and "are". Compute the cosine similarity of the following two simple documents:

   (a) "precision is very very high"

   (b) "high precision is very very very important"

   Answer:

   The word "is" is ignored because it is a stopword:

	high	precision	very	important
Doc 1	1	1	2	0
Doc 2	1	1	3	1

$$\begin{aligned}\Rightarrow \text{Cosine similarity} &= \frac{1\cdot 1+1\cdot 1+2\cdot 3+0\cdot 1}{\sqrt{(1^2+1^2+2^2+0^2)}\times\sqrt{(1^2+1^2+3^2+1^2)}} \\ &= \frac{8}{\sqrt{6}\times\sqrt{12}} = 0.9428\end{aligned}$$

2. (14 points) Assume that an IR system returns a ranked list of 10 total documents for a given query. Assume that according to a gold-standard labelling there are 5 relevant documents for this query, and that the only relevant documents in the ranked list are in the 2nd, 3rd, 4th, and 8th positions in the ranked results. Calculate and clearly show the interpolated precision value for each of the following standard recall levels: {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} for this individual query.

Answer:

Document Number	Recall	Precision
2	$1/5 = 0.2$	$1/2 = 0.5$
3	$2/5 = 0.4$	$2/3 = 0.67$
4	$3/5 = 0.6$	$3/4 = 0.75$
8	$4/5 = 0.8$	$4/8 = 0.5$

Table 1: Precision-Recall values corresponding to relevant documents positions

Recall	Precision
0.0	0.75
0.1	0.75
0.2	0.75
0.3	0.75
0.4	0.75
0.5	0.75
0.6	0.75
0.7	0.5
0.8	0.5
0.9	0
1.0	0

Table 2: Interpolated Precision-Recall values

3. (13 points) The table below shows the final ranked list of results for an IR search together with their continuous human-rated relevance values. Assume the table contains all documents with non-zero relevance. Compute the values of the DCG and NDCG evaluation metrics for each value of n and add them to the table. Complete the second table to show the idealized DCG (IDCG) values.

n	doc	relevance (gain)	——DCG——	——NDCG——
1	D23	0.6	0.6000	0.6000
2	D78	1.0	1.6000	0.8421
3	D90	0.0	1.6000	0.7022
4	D17	0.5	1.8500	0.7316
5	D78	0.9	2.2376	0.8849

n	doc	relevance (gain)	——IDCG——
1	D78	1.0	1.0000
2	D78	0.9	1.9000
3	D23	0.6	2.2786
4	D17	0.5	2.5286
5	D90	0.0	2.5286

4. (13 points) Write a Perl regular expression (regex) for matching the final line in a US Postal address in Texas or California. Assume that it consists of a city name of one or two alphanumeric words followed by a comma and then any amount of optional whitespace, followed by one of the two-letter state abbreviations (TX or CA) followed by some whitespace and then a 5 digit zip-code with an optional "plus four" digits introduced by a hyphen.

   Answer:

   ```
   \b\w+(\b\w+)?,\s*(TX|CA)\s+\d{5}(-\d{4})?\b
   ```

5. (13 points) Assuming Zipf's law with a corpus independent constant $A = 0.1$, what is the fewest number of most common words that together account for more than 18% of word occurrences (i.e. the minimum value of m such that at least 18% of word occurrences are one of the m most common words).

   <u>**Answer:**</u>

   Zipf's law: $p_r = A/r$

   Since the probability of seeing any of the first m words must be 18% or greater, we would like to find the minimal m such that:

$$\frac{A}{1} + \frac{A}{2} + .. + \frac{A}{m} \geq 0.18$$

   Because no closed-form formula exists for the harmonic series $\sum_{i=1..m} 1/i$, we add up the terms for most common words manually to find minimal m:

$$\begin{aligned} p_1 = \frac{0.1}{1} &= 0.1 \\ p_2 = \frac{0.1}{2} &= 0.05 \\ p_3 = \frac{0.1}{3} &= 0.033 \end{aligned}$$

   $p_1 + p_2 = 0.15$, while $p_1 + p_2 + p_3 = 0.183 > 0.18$, therefore $m = 3$.

6. (12 points) Consider the following web pages and the set of web pages that they link to:

   Page A points to pages B and D.
   Page B points to pages C, F, and G.
   Page C points to page D.
   Page D points to page H.
   Page G points to pages E and H.
   Page H points to page C.

   Show the order in which the pages are indexed when starting at page A and using a breadth-first spider with duplicate page detection. Assume links on a page are examined in the orders given above.

   A (start) Indexed
   B (from A) Indexed
   D (from A) Indexed
   C (from B) Indexed
   F (From B) Indexed
   G (From B) Indexed
   H (from D) Indexed
   D (from C) Already visited
   E (from G) Indexed
   H (from G) Already visited
   C (from H) Already visited

   Indexing order in BFS is: A B D C F G H E

7. (21 points) Provide short answers (1-3 sentences) for each of the following questions:

- What are two aspects of the web that make make web search fundamentally different from earlier, traditional IR?

  Answer: 1) Distributed Data: Documents spread over millions of different web servers. 2) Volatile Data: Many documents change or disappear rapidly (e.g. dead links). 3) Large Volume: Billions of separate documents. 4) Unstructured and Redundant Data: No uniform structure, HTML errors, duplicate documents. 5) Quality of Data: No editorial control, false information, poor quality writing, typos, etc. 6) Heterogeneous Data: Multiple media types (images, video), languages, character sets, etc.

- In the vector-space model, why is it **not** necessary to normalize term frequencies when the resulting document vectors are only used for computing cosine similarity?

  Answer: Since cosine similarity only depends on the angle between vectors, and term frequency normalization only affects the length of the vectors and not their angles.

- But why **is** it necessary to normalize term frequencies when the resulting document vectors are used for standard vector-space relevance-feedback methods?

  Answer: Because relevance feedback involves adding and subtracting vectors which does change the angle of the resulting vector. Without normalization, longer documents would affect the result more than shorter documents, which is undesirable.

- What is the functional role (i.e. purpose) of the IDF factor in standard term weighting?

  Answer: The role of IDF is to decrease the weight of common terms that occur in many documents and therefore do not make useful, discriminative search terms and to upweight uncommon terms (ones that appear in few documents) since they are typically more specific and therefore better index terms.

- How does stemming typically affect recall? Why?

  Answer: Stemming typically increases recall because morphological variations of words are collapsed onto a single token, enabling retrieval of relevant documents that contain slight variations of the query tokens in addition to those that contain the query tokens themselves.

- Why does thesaurus-based query expansion typically not work very well?

  Answer: Thesaurus-based query expansion may significantly decrease precision. Many words have multiple meanings in the thesaurus, and adding synonyms for these multiple meanings results in irrelevant terms that cause retrieval of documents that are not relevant to the original query.

- On what type of plot does a power law result in a straight line? What is the slope of the line (in terms of the parameters of the power law, $y = kx^c$)?

  Answer: A power law results in straight line on a log-log plot. The slope of the line is c, the constant in the exponent of the power law $y = kx^c$.

- (Extra credit) Before moving to the US, both the founding father of IR, Gerald Salton, and the inventor of the hash table, Hans P. Luhn, were born in what country?

  Answer: Germany

- (Extra credit) Google settled a lawsuit with what early web company after that company acquired the company that had patented the pay-per-click model?

  Answer: Yahoo