CS371R: Sample Solution to Midterm Exam

October 18, 2012

NAME: _____

1. (a) (13 points) Corpus C consists of the following three documents:

"new york times"
"new york post"
"los angeles times"

Assuming that term frequencies are normalized by the maximum frequency in a given document, calculate the TF-IDF weighted term vectors for all documents in C. Assume that words in the vectors are ordered alphabetically.

Answer:

Term frequencies:

	angeles	los	new	post	times	york
"new york times"	0	0	1	0	1	1
"new york post"	0	0	1	1	0	1
"los angeles times"	1	1	0	0	1	0

Inverse document frequencies:

angeles	los	new	post	times	york
$\log_2 \frac{3}{1}$	$\log_2 \frac{3}{1}$	$\log_2 \frac{3}{2}$	$\log_2 \frac{3}{1}$	$\log_2 \frac{3}{2}$	$\log_2 \frac{3}{2}$

Since $\log_2 3 = 1.5850$ and $\log_2 \frac{3}{2} = 0.5850$, we have the following TF-IDF weighted term vectors:

	angeles	los	new	post	times	york
"new york times"	0	0	0.5850	0	0.5850	0.5850
"new york post"	0	0	0.5850	1.5850	0	0.5850
"los angeles times"	1.5850	1.5850	0	0	0.5850	0

(b) (15 points) Given the following query:

"new new times"

calculate the TF-IDF weighted query vector, and compute the score of each document in C using the cosine similarity measure. Assume that term frequencies are normalized by the maximum frequency in a given query.

Answer:

The TF-IDF weighted query vector is as follows:

angeles	los	new	post	times	york
0	0	$\frac{2}{2} \times \log_2 \frac{3}{2} = 0.5850$	0	$\frac{1}{2} \times \log_2 \frac{3}{2} = 0.2925$	0

The vector lengths for the query and the documents are:

"new new times"	$\sqrt{0.5850^2 + 0.2925^2} = 0.6540$
"new york times"	$\sqrt{0.5850^2 + 0.5850^2 + 0.5850^2} = 1.0132$
"new york post"	$\sqrt{0.5850^2 + 1.5850^2 + 0.5850^2} = 1.7879$
"los angeles times"	$\sqrt{1.5850^2 + 1.5850^2 + 0.5850^2} = 2.3165$

Hence, the scores of the documents in terms of cosine similarity are:

"new york times"	$(0.5850 \times 0.5850 + 0.2925 \times 0.5850)/(0.6540 \times 1.0132)$ $= 0.7746$
"new york post"	$(0.5850 \times 0.5850)/(0.6540 \times 1.7879) = 0.2926$
"los angeles times"	$(0.2925 \times 0.5850)/(0.6540 \times 2.3165) = 0.1129$

(c) (13 points) Assume in response to the results of the query "new new times," that the user rates the following documents as *irrelevant*:

"new york times"
"new york post"

Reformulate the query to account for relevance feedback using the Ide "Dec Hi" method, assuming $\alpha = \beta = \gamma = 1$.

Answer:

The highest ranked among the irrelevant documents is "new york times," according to Part (b). Therefore, the query is reformulated by subtracting the document vector for "new york times" from the query vector:

$$\begin{aligned}
\overrightarrow{q_m} &= \alpha \overrightarrow{q} - \gamma \max_{non-relevant} (\overrightarrow{d_j}) \\
&= \begin{bmatrix} 0 \\ 0 \\ 0.5850 \\ 0 \\ 0.2925 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0.5850 \\ 0 \\ 0.5850 \\ 0.5850 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -0.2925 \\ -0.5850 \end{bmatrix}
\end{aligned}$$

2. (14 points) Given a corpus that consists of the following two documents:

   "new orleans"
   "new hampshire"

   Compute a normalized association matrix that quantifies term correlations in terms of how frequently they co-occur. Order terms in the matrix alphabetically.

Answer:

The *unnormalized* association matrix C is as follows:

	hampshire	new	orleans
hampshire	1	1	0
new		2	1
orleans			1

By applying $s_{ij} = c_{ij}/(c_{ii} + c_{jj} - c_{ij})$, we have the following normalized association matrix:

	hampshire	new	orleans
hampshire	$\frac{1}{1+1-1} = 1$	$\frac{1}{1+2-1} = 0.5$	0
new		$\frac{2}{2+2-2} = 1$	$\frac{1}{1+2-1} = 0.5$
orleans			$\frac{1}{1+1-1} = 1$

3. (12 points) What is the Levenstein distance between the following pairs of strings?

"thorough" and "throughout"

Answer:

The Levenstein distance is 4. Edit operations : delete 'o', insert 'o', insert 'u', insert 't'

"filosofy" and "philosophy"

Answer:

The Levenstein distance is 4. Several edit operation sequences of length 4 are possible, for example: replace 'f' with 'p', insert 'h', insert 'p', replace 'f' with 'h'.

4. (12 points) Given the following document type definition (DTD):

```
<!DOCTYPE db [
  <!ELEMENT db (person*)>
  <!ELEMENT person (name,age,(parent|guardian)?)>
  <!ELEMENT name (#PCDATA)>
  <!ELEMENT age (#PCDATA)>
  <!ELEMENT parent (person)>
  <!ELEMENT guardian (person)>
]>
```

Write a valid document for this DTD that contains the following information:

John Doe is 30 years old. His parent is Robert Doe, who is 55 years old.

The header for the document has been done for you.

Answer:

```
<?xml version="1.0"?>
<!DOCTYPE db [
  <!ELEMENT db (person*)>
  <!ELEMENT person (name,age,(parent|guardian)?)>
  <!ELEMENT name (#PCDATA)>
  <!ELEMENT age (#PCDATA)>
  <!ELEMENT parent (person)>
  <!ELEMENT guardian (person)>
]>
<db>
  <person>
    <name>John Doe</name>
    <age>30</age>
    <parent>
      <person>
        <name>Robert Doe</name>
        <age>55</age>
      </person>
    </parent>
  </person>
</db>
```

5. (18 points) Provide short answers (1-3 sentences) for each of the following questions:

- List **two** disadvantages of the Boolean retrieval model.

  Answer:
  - The Boolean model predicts that each document is either relevant or irrelevant. There is no notion of a *partial* match to the query.
  - Exact matching may lead to retrieval of too few or too many documents.
  - It is difficult to rank the output, since *all* matched documents logically satisfy the query.
  - It is difficult to perform relevance feedback.

- Briefly describe the steps involved in exact phrasal retrieval using a standard word-based inverted index. What piece of information is missing from the `ir.vsr.InvertedIndex` class, which is required for phrasal retrieval?

  Answer: Phrasal retrieval requires an inverted index that stores the positions of each term appearing in a document, which are missing in the `ir.vsr.InvertedIndex` class. The steps involved in the retrieval of phrase P are as follows:
  (a) Retrieve the documents and positions for each term in P.
  (b) Obtain the documents containing *all* terms in P by intersecting the sets of retrieved documents.
  (c) Check for ordered contiguity of the term positions.

- Compare query expansion using global analysis and local analysis.

  Answer: Global analysis requires intensive term correlation computation across all documents only once at system development time, while local analysis requires intensive computation across only the top ranked documents for every query at run time. Local analysis gives better results by avoiding term ambiguity but slows query response time considerably.

- List **two** ways in which Zipf's Law manifests itself on the Web (other than the original manifestation in terms of word frequency)?

  Answer:
  - The number of in-links to and out-links from a page has a Zipfian distribution.
  - The length of web pages has a Zipfian distribution.
  - The number of hits to a web page has a Zipfian distribution.

- What is the fundamental difference in the purposes of HTML and XML?

  Answer: HTML is a markup language for formating documents so they can be diplayed properly in a browser for human readability. XML is a a markup language for formating semi-structured documents containing data so that the structured data can be easily extracted and processed by a computer.

- List **two** ways in which web search is more difficult than traditional IR.

  Answer:

  - Documents are spread over millions of different web servers.
  - Many documents on the web change or disappear rapidly.
  - There are billions of separate documents on the web.
  - There is no uniform structure for web documents. Many of the documents have HTML errors. Up to 30% of them are near duplicates.
  - There is no editorial control. False information and poor quality writing abounds.
  - There are media types other than text, such as images and videos. Documents are in different languages, encoded in different character sets.

- What is the current best explanation for why word frequencies, wealth, city population, and many other properties of human society obey Zipf's law?

  Answer:

  The fact that all of these phenomenon have a "rich get richer" growth process, in which the increase in the size or popularity of an item in a given time period is proportional to its current value.

- (Extra credit) In what decade and in what general geographic area were both hypertext and the graphical user interface originally developed?

  Answer:

  In the 1960's in Silicon Valley (San Francisco Bay Area)

- (Extra credit) What IR pioneer originally developed the hashtable?

  Answer:

  H. P. Luhn