CS378 Information Retrieval and Web Search: Midterm Exam Solution

Oct. 13, 2011

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

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

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) Show the 3-gram (inverted) index constructed for a spelling correction system for the small dictionary containing only the words "gram", "spam", "cram", and "scram". List the 3-grams alphabetically in a table assuming the word-boundary character ($) is alphabetized after "z" and show the posting lists for each.

Answer:

```
am$ -> {gram,spam,cram,scram}
cra -> {cram, scram}
gra -> {gram}
pam -> {spam}
ram -> {gram,cram,scram}
scr -> {scram}
spa -> {spam}
$cr -> {cram}
$gr -> {gram}
$sc -> {scram}
$sp -> {spam}
```

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).

Answer:

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:

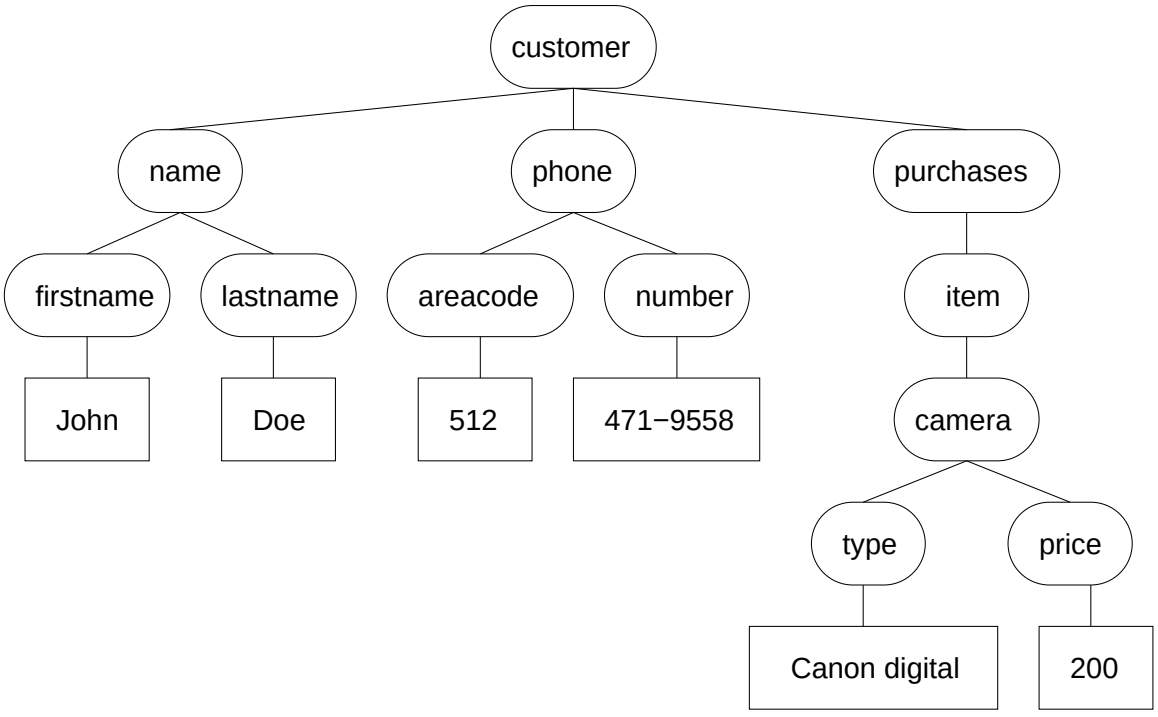
$$\begin{aligned} p_1 = \frac{0.1}{1} &= 0.1 \\ p_2 = \frac{0.1}{2} &= 0.05 \\ p_3 = \frac{0.1}{3} &= 0.033 \end{aligned}$$

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

6. (12 points) Draw the DOM tree for the following XML document:

```
<db>
  <customer>
    <name>
      <firstname>John</firstname> <lastname>Doe</lastname>
    </name>
    <phone>
      <areacode>512</areacode> <number>471-9558</number>
    </phone>
    <purchases>
      <item>
        <camera>
          <type>Canon digital</type> <price>200</price>
        </camera>
      </item>
    </purchases>
  </customer>
</db>
```

Answer:



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

- What is the difference between database management and information retrieval?

  Answer: Database management is focused on *structured* data stored in relational tables, while information retrieval is focused on retrieval and indexing of *unstructured, free-form textual documents*. Database management deals with efficient processing of well-defined queries in a formal language (e.g. SQL), while information retrieval is concerned with retrieving documents relevant to free-text queries.

- Why is Euclidian distance not a good metric for judging the (dis)similarity of documents in vector-space retrieval?

  Answer: Since most queries are quite short, Euclidian distance would tend to prefer arbitrary short documents that may contain none of the words in the query but, like the query, are close to the origin in Euclidian distance, compared to long documents that contain many copies of the words in the query, but whose vectors are far from the origin in Euclidian distance. A metric is needed that normalizes for document length and prefers documents with the same *relative* frequencies of words, rather than the same absolute frequencies.

- 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.

- What additional step must one be careful to perform when experimentally evaluating human-provided relevance feedback?

  Answer: Remove the documents that the user rated from the testing corpus so as not to bias the results by testing on the training data.

- Define "pseudo relevance feedback".

  Answer: Pseudo relevance feedback assumes that top m retrieved documents are relevant, and uses them to reformulate the query using a relevance feedback algorithm like Ide or Rocchio, but without any user interaction. This allows for query expansion that includes terms that are correlated with the query terms.

- 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)?

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

- (Extra credit) What was the first complete web search engine and where was it developed?

  Answer: Lycos at Carnegie Mellon University

- (Extra credit) What was the first complete web browser and where was it developed?

  Answer: Mosaic at the University of Illinois at Urbana Champaign