

Student Name: _____

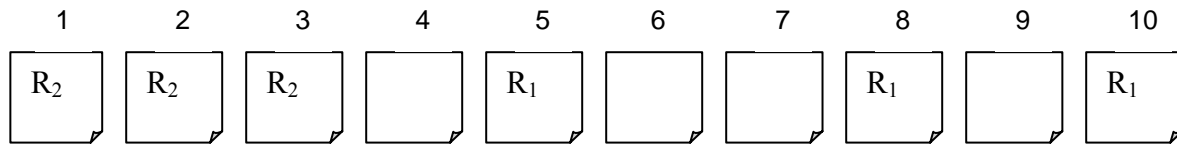
Midterm Exam
Information Retrieval (11-741)
March 5, 2009

Answer all of the following questions. Each answer should be thorough, complete, and relevant. Points will be deducted for irrelevant details. Use the back of the pages if you need more room for your answer.

The exam should take you about 70 minutes to complete. The points are a clue about how much time we think each question should take to answer. We assume about 1.5 points per minute, so a 10-minute question is worth 15 points. Plan your time accordingly.

Good luck.

1. Suppose the top 10 documents returned for a query are as shown below. Relevance is indicated on a three-point scale: 2 (very relevant), 1 (relevant), and 0 (not relevant). In the diagram below, non-relevant documents have blank labels.



- a) Calculate the Average Precision for the ranking. Show your work. It is not necessary to provide a single number – the calculation for this data is sufficient. **[8 points]**

Answer:

$$\begin{aligned} \text{AvgPrec} &= [(1/1) + (2/2) + (3/3) + (4/5) + (5/8) + (6/10)] / 6 \\ &= (1 + 1 + 1 + 0.8 + 0.625 + 0.6) / 6 \\ &= 0.8375 \end{aligned}$$

For Average Precision, there is no distinction between R₂ and R₁. They are equally important.

It was depressing how many students saw multi-valued relevance assessments and automatically thought NDCG.

- b) Calculate the Reciprocal Rank value for the ranking. Show your work. It is not necessary to provide a single number – the calculation for this data is sufficient. **[5 points]**

Answer:

$$\begin{aligned} \text{RR} &= 1/1 \\ &= 1.0 \end{aligned}$$

For Reciprocal Rank, there is no distinction between R₂ and R₁. They are equally important.

- c) Measurements for different queries are usually combined, to improve the reliability of the measurement. Describe how Average Precision for different queries is combined. Describe how Reciprocal Rank for different queries is combined. **[7 points]**

Answer:

The Average Precision is computed for each query. The mean of the Average Precision values for a set of queries is Mean Average Precision (MAP).

The Reciprocal Rank is computed for each query. The mean of the Reciprocal Rank values for a set of queries is the Mean Reciprocal Rank (MRR).

2. Suppose you want to use ‘top docs’ lists (called ‘champion’ lists in the textbook) to improve the speed of query evaluation for terms that are very frequent in the document collection. You know that top docs lists only make sense for terms with very long inverted lists, for example, terms that occur at least 2,000 times in the corpus. How can Zipf’s Law be used to estimate the number of terms that occur at least 2,000 times in a corpus? [8 points]

Answer:

Zipf’s Law is $\text{Rank} \times \text{Frequency} = \text{Constant}$, where $\text{Constant} \approx \text{CollectionSizeInWords} / 10$.

Rank the vocabulary in descending order of frequency. You want to know the rank of the word that occurs 2,000 times in the corpus.

$$\begin{aligned}\text{Rank}_{\text{ctf}=2000} \times 2000 &= \text{CollectionSizeInWords} / 10 \\ \text{Rank}_{\text{ctf}=2000} &= \text{CollectionSizeInWords} / (10 \times 2000) \\ &= \text{CollectionSizeInWords} / 20,000\end{aligned}$$

Many students calculated this value using the *percentage* of terms that have frequency less than 2,000 ($1 - (\text{Rank}_{\text{ctf}=1} - \text{Rank}_{\text{ctf}=2000}) / \text{Rank}_{\text{ctf}=1}$). I accepted this solution, but it isn’t what was asked for, and it was more work than necessary.

3. Stopword removal and stemming are two common techniques to improve a text representation. Describe their effects on the statistics used by a retrieval model. What is the effect on search results if the search engine doesn’t do stopwords removal? What is the effect if it doesn’t do stemming? [15 points]

Answer:

The document length statistic usually includes stopwords, so the statistics used by most probabilistic models are not affected by the inclusion or removal of stopwords. One might make the same assumption for length normalization in the vector space model, although that case is less clear.

Stemming increases the tf and df statistics of most stems. When a query term occurs in a document, its tf weight is probably higher with stemming, and its idf weight is probably lower. For most terms, the increase in tf weight is probably higher than the decrease in idf weight, thus stemming causes query terms to match more strongly.

Most retrieval models give too much weight to stopwords, thus if stopwords are not removed, search results may be dominated by documents that match stopwords. A rare stopword receives a high weight based on its idf. A common stopword has a low idf, but usually a very high tf score, thus a high score overall. Rare stopwords affect few queries, but common stopwords affect most queries.

If the search engine doesn’t do stemming, Precision may be reduced slightly because morphological variants of a query term are not counted in the tf score. Recall will be reduced, because documents that only contain morphological variants of query terms will not be matched. If the search engine does not do stemming, it is the user’s responsibility to include important morphological variants in the query, which most users won’t do (or won’t do well).

4. Provide the formula for the query likelihood retrieval model with Jelenick-Mercer smoothing. Define each component. [8 points].

Answer:

$$P_{JM}(q_i | d) = (1-\lambda) P_{MLE}(q_i | d) + \lambda P_{MLE}(q_i | c)$$

q_i is a query term.

d is a document

c is the corpus

λ is a constant between 0 and 1 that controls the amount of smoothing

P_{MLE} is a maximum likelihood estimate, either $tf/doclen$ or $ctf/corpuslength$

5. Many documents have distinct multiple representations, for example, body text, inlink text, and url text. How are multiple representations incorporated into the query likelihood retrieval model? Provide the formula, and also a brief explanation. [9 points]

Answer:

$$P_s(q_i | d) = \sum_j w_j P_s(q_i | d_j)$$

q_i is a query term.

d is a document

d_j is the j 'th representation of the document

w_j is a constant between 0 and 1 that controls the importance of the j 'th representation. $w_j \geq 0$. $1 = \sum w_j$.

P is a smoothed probability estimate, for example, P_{JM}

The probability of a query term q_i given the document d is a weighted average of its probability given the different representations d_j .

6. PageRank is an important component of web retrieval models. Show how PageRank would be used with a vector space retrieval model, and with a query likelihood retrieval model. Would the differences in how PageRank is used with the two retrieval models cause different effects on user queries, e.g., for long and short queries? Why, or why not? **[15 points]**

Answer:

For the query likelihood model, $p(d|q)$ is rank equivalent to $p(d) p(q|d)$. The prior probability $p(d)$ can be based on a web page's PageRank value, for example, by using training data to learn the probability of relevance for different PageRank values (essentially $p(d) = p(\text{relevant} | \text{PageRank}(d))$).

PageRank is not incorporated directly into the vector space retrieval model. Instead, the vector space retrieval model provides a similarity score $\text{sim}(q, d)$ that can be combined in a linear utility function with some function of PageRank, for example $\text{score}(q, d) = \lambda \text{sim}(q, d) + (1-\lambda) f(\text{PageRank}(d))$.

In the query likelihood model, PageRank has a large effect for short queries, but a small effect for long queries because $p(d) \times p(q_1 | d) \times \dots \times p(q_n | d)$. As the query gets longer, $p(d)$ is just one multiplier among many.

The effect of PageRank on short and long queries in the vector space retrieval model is harder to determine. You might expect that it would have the same effect on both short queries and long queries because cosine correlation normalizes $\text{sim}(q, d)$ by query length.

$$\lambda \frac{\sum_t q_t \cdot d_t}{\sqrt{\sum_t q_t^2} \cdot \sqrt{\sum_t d_t^2}} + (1-\lambda) f(\text{PageRank}(d))$$

However, as query terms are added, they have a larger effect on the numerator than the denominator, so the similarity score increases with query length. Thus, as the query gets longer, PageRank has a smaller effect.

Grading Note: It is not easy to see the effect of PageRank on short and long queries in the vector space. That part of the question was treated as an extra credit problem. **[4 points]**

You needed something like the above answer for full credit.

Partial credit was given for suggesting that PageRank affects short and queries equally in the vector space because cosine correlation normalizes for query length. This is the right kind of analysis, although when you consider actual term weights, you discover that it doesn't work the way you expect.

Partial credit was given for suggesting learning a different λ for each query length. It would be effective, so you get partial credit, but it avoids answering the question that was asked.

7. Similarity Metrics

Let $q = (q_1, \dots, q_n)$ be a query vector, where n denotes the size of the vocabulary, and let $x_i = (x_{i1}, \dots, x_{in})$ be a document vector, where $i \in \{1, 2, 3, 4\}$ denoting the 4 documents. Term frequencies (TF) are used as the term weighting scheme in these vectors.

- a) Fill in the empty slots in the tables below for vector standardization and similarity computation for documents with respect to the query. See the filled slots as examples. [10 points]

Let the query vector be $q = (1, 2, 0)$, (this implies the size of vocabulary $n = 3$).

Table 1. Vector Standardization (Each cell is 0.4 point; total 0.4 x 15 = 6 points)

Document Vector	Doc Length	Average TF \bar{x}_i	L ₂ norm $\ x_i\ _2$	Centered $x_i \mathbf{n} = x_i - (\bar{x}_i, \bar{x}_i, \bar{x}_i)$	Standard Vector $z_i = (z_{i1}, z_{i2}, z_{i3})$
$x_1 = (1, 2, 3)$					
$x_2 = (2, 4, 6)$					
$x_3 = (3, 4, 5)$					
$x_4 = (1, 2, 0)$	3	1	$\sqrt{5}$	$x_4 \mathbf{n} = (0, 1, -1)$	$z_4 = \frac{1}{\sqrt{2}}(0, 1, -1)$

Table 2. Similarity metrics for document ranking (Each cell is 0.4 point; total 4 points).

Document	$q \cdot x_i$	$\cos(q, x_i)$	PCC $r(q, x_i)$
$x_1 = (1, 2, 3)$	5		
$x_2 = (2, 4, 6)$	10		
$x_3 = (3, 4, 5)$	11		
$x_4 = (1, 2, 0)$	5		
Doc Ranking	$x_3 > x_2 > \{x_1, x_4\}$		

Answer:

- Standardization formulae

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^n x_{ik}, \quad z_{ik} = \frac{x_{ik} - \bar{x}_i}{\|z\|_2}, \quad \|z_i\| = \sqrt{\sum_{k=1}^n (x_{ik} - \bar{x}_i)^2} \quad (1)$$

- Dot-product

$$q \cdot x_i = \sum_{k=1}^n q_k x_{ik} \quad (2)$$

- Cosine

$$\cos(q, x_i) = \frac{q \cdot x_i}{\|q\| \times \|x_i\|}, \quad \|q\| = \sqrt{\sum_k q_k^2}, \quad \|x_i\| = \sqrt{\sum_k x_{ik}^2} \quad (3)$$

- PCC

$$r(q, x) = z_q \cdot z_x \quad (4)$$

Table 1. Vector Standardization (Each cell is 0.4 point; totally 0.4 x 15 = 6 points)

Document Vector	Total term count	\bar{x}_i	$\ x_i\ $	Centered $x_i' = x_i - (\bar{x}_i, \bar{x}_i, \bar{x}_i)$	Standard Vector $z_i = (z_{i1}, z_{i2}, z_{i3})$
$x_1 = (1, 2, 3)$	6	2	$\sqrt{14}$	$x_1' = (-1, 0, 1)$	$z_1 = \frac{1}{\sqrt{2}}(-1, 0, 1)$
$x_2 = (2, 4, 6)$	12	4	$2\sqrt{14}$	$x_2' = (-2, 0, 2)$	$z_2 = \frac{1}{\sqrt{2}}(-1, 0, 1)$
$x_3 = (3, 4, 5)$	12	4	$5\sqrt{2}$	$x_3' = (-1, 0, 1)$	$z_3 = \frac{1}{\sqrt{2}}(-1, 0, 1)$
$q = x_4 = (1, 2, 0)$	3	1	$\sqrt{5}$	$q' = x_4' = (0, 1, -1)$	$z_q = z_4 = \frac{1}{\sqrt{2}}(0, 1, -1)$

Table 2. Similarity metrics for document ranking (Each cell is 0.6 point; totally 6 points).

Document	$q \cdot x_i$	$\cos(q, x_i)$	PCC $r(q, x_i)$
$x_1 = (1, 2, 3)$	5	$\frac{5}{\sqrt{5}\sqrt{14}} = \sqrt{\frac{5}{14}} \approx \sqrt{0.36}$	$\frac{-1}{\sqrt{2}\sqrt{2}} = -\frac{1}{2}$
$x_2 = (2, 4, 6)$	10	$\frac{10}{\sqrt{5}2\sqrt{14}} = \sqrt{\frac{5}{14}} \approx \sqrt{0.36}$	$\frac{-1}{\sqrt{2}\sqrt{2}} = -\frac{1}{2}$
$x_3 = (3, 4, 5)$	11	$\frac{11}{\sqrt{5}\sqrt{50}} = \sqrt{\frac{121}{250}} \approx \sqrt{0.484}$	$\frac{-1}{\sqrt{2}\sqrt{2}} = -\frac{1}{2}$
$x_4 = (1, 2, 0)$	5	1	1
Doc Ranking	$x_3 > x_2 > \{x_1, x_4\}$	$x_4 > x_3 > \{x_1, x_2\}$	$x_4 > \{x_1, x_2, x_3\}$

- b) Discuss the major differences of these metrics in ranking documents with respect to a query, such as the tendency to favor longer (or shorter) documents (using the within-document word count as the length measure), the focus on matched proportion instead of absolute term counts, or focus on the variance of term weights from the mean (“center”) of each vector. Overall, which would be the “best” choice for ranking documents in your opinion? Justify your answer. [15 points]

Answer:

Dot-product tends to favor longer (many words) or more verbose (high TF) documents. Cosine, on the other hand, would favor shorter documents among those with the same dot-product values. In other words, cosine similarity focuses on the proportion of matching terms instead of absolute counts. PCC is the only metric (among those three) focusing on variances from the vector-specific means. For example, x_1 , x_2 and x_3 have different lengths and different means, but the same standard vector. Thus using PCC to rank documents, these three documents are indistinguishable.

Cosine or dot-product are more sensible choice than PCC for ranking documents against a query because we want to find relevant documents which typically have more shared words with the query than irrelevant documents. PCC, on the other hand, focuses on local rescaling of term weights in the standardization. It makes sense to remove users’ biases in choosing terminology, but also has the effect of de-emphasizing matching terms among queries and documents, making retrieval less effective.