IR&TM Review I

Part 1: Questions

Chapter 1

Question

Why don't we use grep for information retrieval?

Question

Why don't we use a relational database for information retrieval?

Question

Google does not always interpret the query as a boolean conjunction of its terms. Give examples.

Question

What is a term-document incidence matrix?

Question

In constructing the index, which step is most expensive/complex?

Question

Complex Boolean retrieval systems like Westlaw use many operations that go beyond strictly Boolean operators. Name some of them.

Chapter 2

Question

Define the number of types/tokens in a sentence.

Question

What is the number of types and the number of tokens in this verse?

Fischer 's Fritz fischt frische Fische Frische Fische fischt der Fritz

Use space for tokenization. Do not count lowercase and uppercase as different types.

Question

An IR system can normalize terms by defining equivalence classes. E.g., "suit" and "suits" could be in an equivalence class. What is the limitation of this model in IR?

Question

What is tokenization?

Question

Give an example in English were tokenization is nontrivial

Question

Give an example for German were tokenization is nontrivial.

Question

What is a stop list?

Question

What is lemmatization? Give an example.

Question

What is stemming? Give an example that is not also a lemmatization example.

Question

Name a particular stemmer.

Question

Give an example of a pair of words that a typical stemmer would put in one equivalence class and we would expect improved performance of the IR system.

Question

Give an example of a pair of words that a typical stemmer would put in one equivalence class and we would expect decreased performance of the IR system.

Question

Name two data structures that support phrase queries.

Question

Name a data structure that supports proximity queries.

Chapter 3

Question

Which data structures are typically used for locating the entry for a term in the dictionary?

Question

Which data structure is best used for locating the entry for a term in the dictionary if the collection is static?

Question

Which data structure is best used for locating the entry for a term in the dictionary if prefix search must be supported?

Question

Which special strings are stored in the permuterm index for the word "car"?

Question

What sequence of letters is looked up in the permuterm index for the following wildcard queries?
X, X*, *X, *X*, X*Y

Question

What is the difference between the regular inverted index used in IR and the k-gram index?

Question

Give an example of a query that cannot be corrected using isolated-word spelling correction.

Question

Define Levenshtein edit distance.

Question

Define Damerau-Levenshtein edit distance.

Chapter 5

Question

Give the formula for Zipf's law.

Question

Give the formula for Heaps' law.

Chapter 6

Question

What is the feast or famine problem?

Question

Define the Jaccard coefficient

Question

What is the bag of words model?

Question

What is the advantage of idf weighting compared to inverse-collection-frequency weighting?

Question

What is the tf-idf weight of term t in document d?

Question

What is the relationship between term frequency and collection frequency?

Question

Why don't we use Euclidean distance of tf-idf vectors to rank documents with respect to a query?

Question

Write down the formula for cosine similarity between query q and document d.

Question

Explain the notation ddd.qqq

Chapter 7

Question

What is the advantage of pivot normalization compared to regular cosine normalization?

Question

What is document-at-a-time processing?

Question

What index organization does document-at-a-time processing require?

Question

What is term-at-a-time processing?

Question

What data structure does term-at-a-time processing require that document-at-a-time processing does not require?

Question

What is a tiered inverted index?

Question

Name two criteria that can be used for deciding as to whether to put a document d in tier 1 of a tiered index.

Chapter 8

Question

Name three criteria for evaluating a search engine.

Question

What are the components of an information retrieval benchmark?

Question

What is the difference between the concepts of query and information need?

Question

Define precision

Question

Define recall

Question

Define F_1

Question

What is the harmonic mean of two numbers?

Question

Why is F_1 defined as the harmonic mean?

Question

What is an easy way of maximizing the recall of an IR engine?

Question

What is an easy way of maximizing the precision of an IR engine?

Question

What is a precision-recall curve?

Question

An evaluation benchmark ideally should tell us for any document-query pair whether the document is relevant to the query. Why is Cranfield the only collection that actually satisfies this desideratum?

Question

Define the kappa measure

Question

What is the minimum and maximum of the kappa measure?

Question

What is the significance of kappa being less than / greater than 0?

Question

What is A/B testing?

Question

What does marginal relevance measure?

Question

What distinguishes a dynamic from a static summary?

Question

What is a simple heuristic for computing a dynamic summary if you can display n characters?

Chapter 9

Question

What is the difference between adhoc retrieval and relevance feedback?

Question

Give the mathematical definition of the centroid

Question

In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a 'Find pages like this one" search correspond to?

Question

Why is relevance feedback not used by most search engines?

Question

What is the difference between relevance feedback and manual query expansion?

Question

Give an example of ineffective automatic query expansion

Question

Search engines log the sequence of queries that a user issues during a session. How can this be exploited for query expansion?

Question

Search engines log the documents users click on in response to a query. How can this be exploited for query expansion?

Chapter 12

Question

Describe the three steps of the basic language modeling approach to information retrieval. Given: a collection of documents d_i and a query q.

Question

We defined the multinomial distribution as follows.

$$P(d) = \frac{L_d!}{Review_{t_1,d}!Review_{t_2,d}!\cdots Review_{t_M,d}!}P(t_1)^{Review_{t_1,d}}P(t_2)^{Review_{t_2,d}}\cdots P(t_M)^{Review_{t_M,d}} \quad (1)$$

What is the meaning of $Review_{t_1,d}$ in this formula? What is the meaning of $P(t_1)$ in this formula?

Question

What is the basic idea of the language model approach to adhoc IR?

Question

How is length normalization performed in the vector space model vs the language model approach to IR?

Question

What is the number of classes in the Naive Bayes approach to classification? What is the number of classes in the language model approach to IR?

Chapter 13

Question

What is the machine learning approach to text classification?

Question

Give one advantage and one disadvantage of rule-based classifiers compared to machine-learned classifers.

Question

What is bad about maximum likelihood estimates of the parameters $P(t|c)$ in Naive Bayes?

Question

What is the time complexity of training a Naive Bayes classifier and why?

Question

What is the main independence assumption of Naive Bayes?

Question

What is feature selection?

Question

What is feature selection used? Give the two main reasons?

Question

In words: what is the meaning of mutual information when used for features selection in text classification?

Chapter 14

Question

Write down the formula for Rocchio classification.

Question

Give three examples of linear classfiers and one example of a nonlinear classifier.

Chapter 15

Question

What is the definition of a linear classifier?

Question

For linearly separable problem, how many different linear decision boundaries are there that separate the two classes of the training set perfectly?

Question

Which decision boundary does the linear SVM choose?

Question

What is a support vector?

Question

How does an SVM classify a test set point in the margin

Chapter 16

Question

What is the difference between classification and clustering?

Question

Why is result set clustering useful?

Question

What is hard/soft clustering?

Question

Does K-means always converge and why?

Question

Does K-means find the global optimum? Why (not)?

Chapter 18

Question

Write down the matrix equation that defines SVD. The term-document matrix C should be the left side of the equation and a product of matrices the right side.

Question

We learned that LSI can improve retrieval because it gets rid of details. What is meant by that?

Chapter 19

Question

What was the Goto model and what was bad about it?

Question

What are the advantages of a search engine ad compared to other types of ads (radio, television, newspaper)?

Question

What is the problem with duplicates and near duplicates in terms of relevance to the user?

Question

What is the advantage of shingling/sketches for near duplicate identification compared to computing tf-idf similiarity scores between documents?

Question

How can we eliminate exact duplicates in information retrieval?

Chapter 20

Question

What is politeness?

Question

What is freshness?

Chapter 21

Question

When using PageRank for ranking what assumptions are we making about the meaning of hyperlinks?

Question

What is a Google bomb? Give an example

Question

Why is PageRank a better measure of quality that a simple count of inlinks?

Question

What is the meaning of the PageRank q of a page d in the random surfer model?

Question

What is ergodicity and why is it important for PageRank?