Lab Class IR

Submission by January 14, 2024,

Exercise 1 : Datasplits

When training and evaluating a ranking model, the dataset is usually separated into three “splits”, **train-**, **test-**, and **validation split**.

(a) What is each of these splits used for?

(b) Why is the data split?

(c) Why is there a separate test- and validation split?

Exercise 2 : Significance Testing

(a) What is significance testing used for in the context of evaluating ranking models?

(b) Imagine, you are comparing the effectiveness of many ranking models for statistical significance. For three of these, Student’s t-test expresses statistical significance. Can you reject the null hypothesis for these models?

Exercise 3 : Hypothesis Testing

You tested your hypothesis: *“On english text, removing all vowels from queries and documents after stemming does not decrease ranking effectiveness in terms of nDCG@5.”* and get an effectiveness degradation of 0.12. Student’s t-test gives you a p-value of $p = 42\%$.

(a) What is the null hypothesis?

(b) What is your result? Can you accept or reject the null hypothesis?

Exercise 4 : Abstract Ranking Model

We introduced an abstract model of ranking, where documents and queries are represented by features. What are some advantages of representing documents and queries by features? What are some disadvantages?

Exercise 5 : Abstract Ranking Model

Documents can easily contain thousands of non-zero features. Why is it important that queries have only a few non-zero features?

Exercise 6 : Inverted Index

Indexes are not necessary to search documents. Your web browser, for instance, has a “Find” function in it that searches text without using an index. Also the UNIX tool `grep` does not use an index. When should you use an inverted index for text search? What are some advantages of using an inverted index? What are some disadvantages?

© Hagen/Merker 2024

Exercise 7 : Inverted Index

We have seen many different ways to store document information in inverted lists of different kinds. What kind of inverted lists might you build if you needed a very small index? What kind would you build if you needed to find mentions of cities, like Los Angeles or São Paulo?

Exercise 8 : Wildcard indexing

How may a search engine that uses an n-gram inverted index be modified to support these wildcards:

- Token-Wildcard *?* that can match any token (e.g., *to ? or not to be*)
- Character-Wildcard * that can match any character in a token (e.g., *in*m*ion* should match among others *information*)

Which components need to be changed and how?

Exercise 9 : Index Compression

Come up with at least two concrete approaches to compress singly linked posting lists. What are advantages or disadvantages of each algorithm? *There are no concrete right answers since this question is about creative thinking.*