Lab Class IR

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 do we need to separate evaluation splits? That is, why do we need separate test- and validation splits?

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 $p = 42\%$, $\alpha = 0.05$.

(a) What is the null hypothesis?

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

Exercise 4 : Vocabulary vs. Terminology

Explain the difference between a "vocabulary" and a "terminology".