CSCE 470

Programming Assignment #1

Ranking Models VSM and BM25F

Jason Gilman

09-28-2020

Introduction

This report will describe the design choices undertaken when implementing the VSM and BM25F ranking models. An explanation as to why the parameters work, as well as their effect on the models' overall effectiveness will be given.

NDCG Results

The following NDCG scores were given when running the development and training data sets on the two ranking models.

NDCG Scores		
	Dev	Train
VSM	0.862726196	0.819440342
BM25	0.849953738	0.813266812

Parameter Values

The following parameter values were used in the design process to tune the ranking models. The weights given to the title and body of a document were used to tune the VSM and BM25 ranking models, while B_{title} , B_{body} , λ, λ', K_1 were only used for the BM25 model.

Parameter Values							
	weightTitle	weightBody	B_{title}	B_{body}	λ	λ'	K_1
VSM	0.3	0.6	-	-	-	-	-
BM25	0.9	0.1	0.9	0.1	0.9	0.9	10

Design Rationalization

Overview

For my design of the VSM, tuning the parameters to offer more weight to the body resulted in greater NDCG scores, while giving more weight to the title offered better scores in the BM25. This is most likely because the BM25 is sensitive to term frequency and document length. This offers a greater opportunity to consider the title when ranking the documents with BM25. In contrast, the VSM model does not consider document length, so the body should be considered more when ranking the documents.

BM25 Parameters

The BM25 specific parameters offer increased tuning of the model. A high value of K_1 contributes to the score greatly. This is why I opted to use a high k value, there was a direct correlation to an increase in score to an increase in k_1. The B_{title} and B_{body} values are used when normalizing term frequencies in the BM25 model. Depending on the field of the term frequency, they give weight to terms in the fields. The λ and λ' values are used when ranking with non-textual features. The λ value accounts for rescalings in the ranking model. The λ' value is included in a function that is chosen depending on the feature you want to account for.

TF and V_j Functions

In my implementation of the two ranking functions, sub-linearly scaling the term frequencies produced greater NDCG scores. As the raw term frequency increases, the normalized term frequency results in a greater score. Therefore, I chose to scale the term frequencies. The V_j function was determined to be the log of λ' concatenated with the document page rank. In the V_j function, I wanted to account for the page rank, and the logarithmic function is known to work well with the BM25 model.

Run Instructions

```
gradlew.bat build && java -jar build/libs/pa1.jar rank data/signal.dev vsm && java -
jar build/libs/pa1.jar ndcg ranked.txt data/rel.dev out.txt
```

```
gradlew.bat build && java -jar build/libs/pa1.jar rank data/signal.dev bm25 && java -
jar build/libs/pa1.jar ndcg ranked.txt data/rel.dev out.txt
```