CAB431 Assignment 2

n9720316

Patric Marchant

2020

Contents

[Statement of Completeness 2](#_Toc41239625)

[Q1 2](#_Toc41239626)

[Q4 2](#_Toc41239627)

[Q7 3](#_Toc41239628)

[README 3](#_Toc41239629)

# 

# Statement of Completeness

|  |  |
| --- | --- |
| Q1 | Yes |
| Q2 | Yes |
| Q3 | Yes |
| Q4 | Yes |
| Q5 | Yes |
| Q6 | Yes |
| Q7 | Attempted |

# Q1

**Procedure TrainingSetDiscovery(Q, U)**

//outputs: list of documents in the training set U which are relevant (1) or non-relevant (0) based on a query Q

1. #document and query preprocessing

Let docs = {}

For each document d in U

Get document ID

Find contents of <text> markup

Get tokens/terms and their frequencies

Append (document ID, {term: freq, …}) into docs

1. let topicStatement = Q

let output = {}

for all di in docs do

let found = 0

if topicStatement.title in di

found = 1

print(num + “ “ + di.title + “ ” + found)

output[num].di.title = found

return output

# Q4

Similarly to the baseline bm25 implementation used in Q3, more helper functions (i.e “testRanking” and “w4”) have taken from the CAB431 workshops, both of these functions are used in this question

**Procedure IFTrainingAndTesting(D = D+ ∪ D-, U)**

//outputs: relevant features in the collection of documents

Parse benchmark files for each “topicassignment101-150” into list

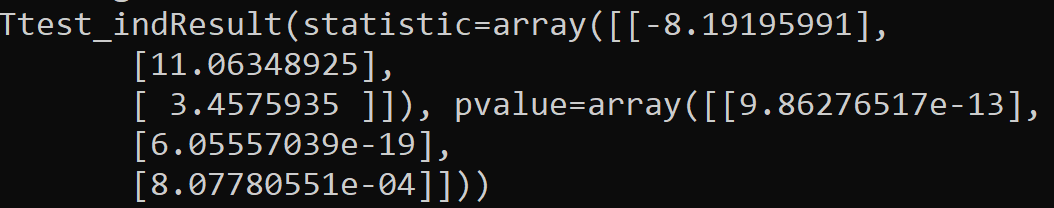
For each coll x in R\_num\_list

Calculate training weights using “w4” function with coll, benchmark[coll] and theta as arguments and store output in trainingWeights[x]

Calculate test ranks using “testRanking” function with coll, trainingWeights[x] as arguments and sort/return output

# Q7

Conducting a T test on both arrays each containing the Precision, Recall, and F1 scores for all 50 documents proved unsuccessful, with the function running, but the values being incorrect. This error was most likely caused by incorrect calculations in earlier functions.



# README

The file structure needed to compile the assignment2.py script requires helper functions from the workshops (coll and df) as well as the porter2 stemming packages and given files and folders from blackboard. 2 additional folders, baselineresults and IFranks also need to exist for the script to store written files in.

The assignment script can be run easily by unzipping the “code.zip” file, cd into the root folder where it is located, installing scipy using pip, and running “python assignment2.py”

