Com S 435/535 Programming Assignment 4
Information retrieval
Authors: Yijia Huang and Mohammad Wardat

Q1) Pick 5 distinct queries.
Q2) For each query, list top 10 files along with TPScore, VSScore, and Relevance Score.

1- One query with exactly one word

Query	history		
Top 10 Document	Relevance Score	TPScore	VSScore
List_of_Minnesota_Twins_broadcasters.txt	0.0557642188699479	0.0	0.13941054717486975
Baseball_at_the_Central_American_and_Caribbean_Games.txt	0.05089150553928377	0.0	0.12722876384820941
List_of_Major_League_Baseball_players_with_a_home_run_in_their_final_major_league_at_bat.txt	0.04449475351491726	0.0	0.11123688378729314
1982_San_Diego_Padres_season.txt	0.04039178859793635	0.0	0.10097947149484086
1964_Pittsburgh_Pirates_season.txt	0.03973716817001274	0.0	0.09934292042503186
1981_San_Francisco_Giants_season.txt	0.0381241859562155	0.0	0.09531046489053875
1976_Major_League_Baseball_season.txt	0.035826490698863525	0.0	0.0895662267471588
Baseball%27s_Seasons.txt	0.03277764325466199	0.0	0.08194410813665497
1981_Montreal_Expos_season.txt	0.029142565140803968	0.0	0.07285641285200992
1967_Pittsburgh_Pirates_season.txt	0.029039793731427913	0.0	0.07259948432856977

2-One query with exactly two words

Query	history museum		
Top 10 Document	Relevance Score	TPScore	VSScore
Heinz_History_Center.txt	1.2449422651178832	2.0	0.11235566279470799
Museum.txt	1.243049647792123	2.0	0.10762411948030748
Jackie_Autry.txt	1.229156392618461	2.0	0.07289098154615252
Studs_Terkel.txt	1.2228522168567055	2.0	0.057130542141763795
Holgu%C3%ADn.txt	1.2172373589240972	2.0	0.043093397310243316
National_Baseball_Hall_of_Fame.txt	1.216304299647267	2.0	0.04076074911816749
National_Baseball_Hall_of_Fame_and_Museum.txt	1.216304299647267	2.0	0.04076074911816749
Baseball_Hall_of_Fame.txt	1.216304299647267	2.0	0.04076074911816749
Major_League_Baseball_Hall_of_Fame.txt	1.216304299647267	2.0	0.04076074911816749
Century_of_Progress.txt	1.2147783910149919	2.0	0.03694597753747978

3- One query with exactly 3 words

Query	baseball game season		
Top 10 Document	Relevance Score	TPScore	VSScore
1964_Major_League_Baseball_season.txt	0.97884473	1.5	0.197111824
1956_Major_League_Baseball_season.txt	0.963943376	1.5	0.15985844
1976_Major_League_Baseball_season.txt	0.960819078	1.5	0.152047695
2006_New_York_Yankees_season.txt	0.959068079	1.5	0.147670197
2002_New_York_Yankees_season.txt	0.95776484	1.5	0.144412099
1936_Major_League_Baseball_season.txt	0.957599668	1.5	0.14399917
New_York_Yankees_award_winners_and_league_leaders.txt	0.957010887	1.5	0.142527218
2003_Major_League_Baseball_season.txt	0.956529102	1.5	0.141322754
1997_New_York_Yankees_season.txt	0.956527675	1.5	0.141319188
1991_New_York_Yankees_season.txt	0.9557166	1.5	0.139291501

4- Two queries with more than 3 words

Query	1925 Chicago Cubs season		
Top 10 Document	Relevance Score	TPScore	VSScore
1925_Chicago_Cubs_season.txt	0.881739167	1.333333333	0.204347917
Sparky_Adams.txt	0.838034446	1.333333333	0.095086116
George_Gibson_(baseball).txt	0.235053073	0.363636364	0.042178138
1908_Major_League_Baseball_season.txt	0.21363801	0.114285714	0.362666453
Billy_Meyer.txt	0.188202372	0.285714286	0.041934501
1925_in_baseball.txt	0.18461972	0.078431373	0.34390224
1905_Chicago_Cubs_season.txt	0.183135568	0.210526316	0.142049447
Chicago_Cubs_award_winners_and_league_leaders.txt	0.181737115	0.114285714	0.282914216
1907_Chicago_Cubs_season.txt	0.181563398	0.210526316	0.13811902
1922_Chicago_Cubs_season.txt	0.180195844	0.210526316	0.134700135

Query	Robert Morris University for		
Top 10 Document	Relevance Score	TPScore	VSScore

Duquesne_University.txt	0.617115155	1	0.042787887
Pittsburgh_metropolitan_area.txt	0.613359495	1	0.033398738
Robert_Morris_University.txt	0.359523625	0.5	0.148809062
Chicago.txt	0.310143923	0.5	0.025359808
John_Jay.txt	0.309570173	0.5	0.023925433
Pittsburgh.txt	0.233826687	0.363636364	0.039112173
Peoria,_Illinois.txt	0.195185512	0.307692308	0.02642532
Wagner_College.txt	0.149230201	0.210526316	0.057286028
Sports_in_Pittsburgh.txt	0.146691478	0.210526316	0.050939223
Matt_Morris_(baseball).txt	0.125779512	0.114285714	0.143020208

Q3) Do you think the rankings produced are acceptable?

All the outputs are reasonable, except the last one. Since the last query contains a proposition "for", then it may misleading the main topic of the original query. The actual file we want from the last query is "Robert_Morris_University.txt". By taking look at the TPScore and VSScore of the file in row 3. Its VSScore is outstanding in the top 10 files, but TPScore is just half of the highest one which impacts the relevance score most. The main reason is that the shortest distance between words "University" and "for" in the top-2 files is less than the one in the actually wanted file. Even there is less relationship between those two words, the distance still impacts the relevance score a lot.