Jean-Loup Johnston

40006259

Comp 479

Winter 2019

Sample runs

I am using b = 0.5 and k = 5

I tried running an And query for several variations of **Democrats’ welfare and healthcare reform policies** but had no documents returned. Making all the words singular and even variations with a term removed returned nothing. So I did an OR query and let the ranking return the most relevant ones first:

['21577', '15630', '17560', '1595', '17697', '19173', '18731', '2916', '6940', '14976', '5093', '11204', '18878', '12326', '19134', '1999', '15791', '20423', '6276', '4268', '219', '7467', '20686', '4006', '17940', '18161', '19829', '2132', '18683', '3266', '951', '5868', '19929', '7219', '18271', '19877', '7806', '13257', '7375', '8072', '9347', '12750', '13671', '16171', '19660', '6052', '15661', '12806', '9704', '13548', '10230', '11749', '3551', '16777', '2493', '2417', '4130', '8310', '9688', '3366', '20091', '16092', '1908', '225', '8139', '8307', '20023', '10212', '16778', '2068', '7401', '7549', '5235', '8592', '8662', '7031', '12456', '12471', '1971', '18449', '4882', '2883', '2891', '9248', '19133', '12552', '20890', '18357', '17934', '18597', '20805', '9096', '17740', '20449', '20094', '19157', '20061', '2011', '19453', '19039', '18469', '3289', '535', '16827', '1900', '7019', '6581', '18159', '1760', '7433', '3502', '3670', '3124', '18328', '10251', '16226', '18722', '20829', '6329', '12879', '17915', '5659', '7216', '20461', '15873', '21561', '4059', '6079', '1088', '12438', '862', '10664', '2242', '6323', '2594', '672', '8289', '9226', '17023', '2802', '16196', '3978', '1188', '6600', '32', '55', '20795', '16048', '19508', '4508', '8635', '16935', '16957', '8667', '17423', '7651', '20631', '1895', '16540', '3113', '20311', '15799', '4287', '7025', '12455', '1969', '4396', '19017', '3899', '16544', '12961', '7493', '12970', '17452', '19756', '18130', '2140', '19147', '7551', '18941', '17242', '3619', '15140', '8151', '11463', '21158', '6606', '16168', '10125', '2507', '15666', '19220', '11456', '16292', '61', '8137', '28', '7595', '19115', '18163', '9944', '54', '9055', '10633', '1857', '10355', '8885', '12828', '17394', '19495', '11955', '17438', '8242', '21455', '13241', '17953', '11592', '8135', '4891', '19600', '19740', '17185', '19321', '20335', '6108', '12640', '1563', '9659', '226', '6426', '2965', '18516', '17307', '19525', '14270', '17217', '17271', '17202', '21525', '2979', '17598', '1954', '3411', '18507', '12431', '9011', '2156', '19102', '11998', '12415', '2541', '2980', '7892', '20828', '5891', '19834', '21119', '5160', '5288', '19824', '18868', '7314', '3752', '17350', '4854', '17747', '821', '234', '16375', '1318', '12856', '11062', '4727', '4625', '14825', '14905', '20462', '15694', '16006', '12844', '4815', '428', '20678', '12689', '11314', '5227', '9774', '3847', '17087', '2848', '20727', '15191', '2415', '6039', '12718', '16749', '9673', '4587', '6415', '14749', '12497', '15543', '209', '18734', '3030', '1105', '17402', '17409', '8413', '11320', '5774', '8835', '15408', '15456', '16770', '17803', '18598', '20860', '17038', '14874', '9694', '8240', '17420', '4940', '13722', '2223', '15369', '8038', '3064', '17070', '13260', '7633', '20911', '3492', '2052', '7537', '7610', '6746', '1226', '7524', '17528', '3511', '2078', '19014', '18885', '17106', '1416', '1175', '6431', '5453', '10780', '19591', '12744', '16759', '9442', '6351', '2387', '20756', '3928', '19976', '16789', '6927', '17443', '12623', '17017', '14950', '4575', '20907', '20925', '17269', '11082', '18681', '11881', '13869', '10800', '15364', '11163', '19291', '21391', '13178', '15763', '7360', '10134', '19845', '18041', '13045', '15718', '7343', '17196', '6127', '3364', '4736', '17216', '17277', '20027', '6954', '19398', '10661', '10774', '12440', '12612', '11387', '15224', '570', '20780', '21508', '6744', '17477', '16103', '14911', '5943', '16701', '8982', '16963', '4491', '860', '8390', '7877', '1376', '18568', '10100', '3198', '17382', '8608', '8676', '9180', '9549', '18', '1957', '3537', '21542', '2190', '18443', '15154', '15382', '19015', '2427', '5717', '5944', '12598', '17381', '2913', '18009', '5432', '6796', '18472', '9462', '1947', '874', '8176', '5973', '19019', '17058', '1210', '5516', '8141', '6099', '20868', '3246', '15525', '12519', '10995', '12986', '742', '18609', '18985', '9701', '18642', '21273', '6392', '7310', '10670', '7032', '15738', '9873', '13637', '20863', '926', '942', '11820', '13323', '10936', '10049', '1786', '14880', '11770', '16051', '10605', '7155', '12039', '10486', '16214', '181', '14809', '10344', '17363', '17293', '17300', '9445', '6472', '18265', '17296', '2118', '17126', '19543', '5318', '15552', '18726', '14766', '2278', '9795', '17119', '17395', '17397', '11327', '7254', '17201', '13908', '8203', '5285', '7099', '13592', '18776', '5850', '3562', '18835', '8427', '9639', '8743', '10905', '5169', '14419', '7105', '8747', '238', '8117', '17312', '20865', '11600', '19947', '18148', '16198', '18625', '263', '5271', '18420', '3885', '1867', '17750', '12602', '237', '12773', '9307', '18403', '10350', '308', '9782', '18005', '8076', '19016', '6410', '6657', '12942', '13752', '5542', '1611', '236', '18455', '740', '2352', '11175', '17305', '9788', '16784', '4603', '179', '9848', '5206', '15957', '19000', '11214', '17410', '19604', '10255', '5345', '8069', '8210', '17207', '7628', '15678', '19211', '13244', '11732', '13657', '3019', '9758', '13046', '2197', '20500', '12673', '5759', '6190', '7062', '4871', '11198', '12025', '3065', '8980', '14826', '6395', '17358', '19285', '6896', '6722', '13034', '17368', '9470', '16200', '15470', '335', '11768', '10771', '10689', '15442', '17450', '18299', '18086', '16937', '15549', '7405']

AND query for **Drug company bankruptcy**:

['3091', '2127']

No stemming was done. I tried the query with the term “bankruptcies” but had no documents returned, so I changed it to Bankruptcy and had 2 results.

AND query for **George Bush**:

['20891', '4008', '8593', '16780', '3560', '20719', '7525', '20860', '8500', '2796', '965', '854', '5405']

A simple query to find documents containing both terms. The limitation is that we don’t know the position of terms. Ideally we would want the terms next to each other in returned documents.

**President George bush**

Returned the same results but ranked differently. Perhaps the ranking is better with this query.

It is difficult to verify the ranking because none of the returned articles list George Bush in the people tag.

['20891', '8593', '4008', '16780', '3560', '20719', '7525', '20860', '2796', '8500', '965', '854', '5405']

Project report

My first step in this project was to analyze and understand the bm25 equation and find the requirements. I then looked at my previous SPIMI algorithm and found which values I needed for the equation. I gathered a dictionary of all document lengths while reading through the reuters files, and stored this dictionary in a file on disk. From these lengths I could find the average document length. At the same time I made a dictionary of all term frequencies for every document and also stored that in a file on disk. Then after merging I went through the final index and found the document frequency of every term, which I stored in a file on disk. Now that I had all the arguments for the bm25 equation I could implement it in the query file.

When a user queries, the query function runs then passes the returned list to bm25 to rank it. The final list can then be sorted by rank and printed to the user. The bm25 equation gets all the required variables from the saved documents on disk.

From this assignment I learned how to apply a ranking equation to a list of documents. There a more variables in the ranking than I would have guessed before starting this project. I also became more familiar with the python programming language by doing this assignment.