Project 2 – Comp 479

Name: Marie Gobran  
id: 27739060

***Implementation of BM25:***

Steps in process\_query.py:

1. Tokenize query
2. Get an accumulated result set of all query tokens
3. Evaluate each document in the result set based on the bm25 algorithm
4. Sort the result set based on scores

***Test queries results:***

1. **Democrats' welfare and healthcare reform policies**

['179', '181', '236', '263', '570', '740', '874', '926', '942', '1175', '1210', '1318', '1611', '1867', '1947', '1969', '2190', '2197', '2223', '2278', '2352', '2913', '3019', '3064', '3065', '3246', '3364', '3492', '3511', '3752', '3885', '3928', '4575', '4603', '4736', '4871',….]

2. **Drug company bankruptcies**

['7', '8', '9', '12', '17', '19', '21', '25', '33', '39', '56', '58', '70', '84', '91', '116', '119', '122', '123', '125', '127', '131', '134', '146', '147', '149', '150', '153', '158', '180', '184', '191', '194', '206', '214', '236', '247', '265', '272', '276', '279', '281', '316', '326', '333', '334', '366', '369', '371', '376', '378', '381', '385', '390', '391', '403', '408', '409', '420', '438', '439', '441', '454', '463', '465', '471', '474', '477', '479', '482', '486', '493', '503', '504', '510', '520', '524', '532', '537', '541', '544', '549', '552', '564', '585', '591', '599', '604', '609', '612', '626', '627', '630', '631', '633', '639', '643', '651', '652', '670', '674', '677', '682', '706', '720', '721', '727', '732', '736', '737', '747', '753', '757', '761', '762', '763', '764', '766', '778', '788', '805', '807', '819', '820', '825', '845', '852', '853', '863', '865', '882'…..]

3**. George Bush**

['28', '43', '178', '196', '255', '286', '342', '871', '1022', '1502', '1667', '1729', '2356', '2733', '2766', '2802', '3014', '3366', '3390', '3563', '3938', '4098', '4764', '4853', '5157', '5334', '5459', '5574', '5631', '5711', '6062', '6065', '6296', '6423', '6564', '6656', '6744', '6882', '6896', '6940', '6989', '7186', '7326', '7469', '7471', '7765', '8009', '8053', '8252', '8554', '8748', '8765', '9002', '9150', '9247', '9342', '9755', '10400', '10682', '10725', '11380', '11624', '12002', '12009', '12209', '12261', '12460', '12605', '12709', '12720', '12806', '12967', '13246', '13507', '14809', '14914', '15284', '15342', '15363', '15405', '15439', '15880', '16090', '16115', '16199', '16218',…]

I certify that this submission is my original work and meets the Faculty's

Expectations of Originality

Name: Marie Gobran

ID: 27739060

date: 12/11/2018