ASSIGNMENT-3: PRIOR RANKING OF DOCUMENTS

ANANNYA MATHUR(2019TT10953)

TOPICS COVERED:

1. Computing a similarity graph between each pair of documents in the collection using a similarity function.
2. Using the above similarity graph, calculated the PageRank scores of every document.

1. SIMILARITY COMPUTATION:
   1. Term overlap:

$$Sim_{jaccard}(d_1, d_2) = \frac{|TermSet_{d_1} \cap TermSet_{d_2}|}{|TermSet_{d_1} \cup TermSet_{d_2}|}$$

The TermSet(di) is the set of terms that appear in a document after stemming them using Porter stemmer. We were expected to retain the stopwords.

   2. TF-IDF Similarity:

$$\forall_{d_i \in D,\, j \in V} \qquad tf_{i,j} = 1 + \log_2(f_{i,j})$$

$$\forall_{d_i \in D,\, j \in V} \qquad idf_j = \log_2\left(1 + \frac{|D|}{df_j}\right)$$

$$Sim_{cosine}(\vec{d}, \vec{d'}) = \frac{\vec{d} \cdot \vec{d'}}{||\vec{d}|| \times ||\vec{d'}||}$$

Where V is the vocabulary size (after stemming using Porter stemmer), D is the document collection, dfj is the number of documents that jth term appears in, fi,j is the frequency of jth term in a document di, and each entry of the vector representation of a document di, di vector consists of the product tfi,j × idfj as given above. Hence the length of all di vector is the same and equal to |V |.

2. Computing PageRank
   Maintained an undirected, weighted graph between documents where weight was associated with their corresponding similarity values. A

damping factor of 0.85 was used. Since we were free to use any external library for computing PageRank scores, NetworkX was used.

Approach taken-

To cut down on computation resources, built and maintained a dictionary of every word in the collection along with their term frequencies in every document and their inverse document frequencies. The constructed dictionary was used to compute the tf-idf similarity between documents.

To maintain the graph, iterated through every pair of documents, computed their similarity scores according to the method mentioned and added the two documents with their scores both to the results file and to the graph using G.add_edge(doc1,doc2,weight=sim), where G=nx.Graph() [nx= NetworkX].

Once the iteration over documents was completed, calculated the PageRank scores of documents using nx.pagerank(G,alpha=0.85).

Results:

1. Using the term overlap method, the top 20 documents obtained were:
   {"sci.electronics/54247": 0.00017695160601003757,
   "sci.med/59271": 0.00017268299100788205,
   "sci.med/59454": 0.00017263098629670507,
   "talk.religion.misc/84349": 0.00017030712939422458,
   "sci.med/59407": 0.00016939858856125672,
   "alt.atheism/54160": 0.0001692949268095993,
   "sci.electronics/54263": 0.0001677013572577521,
   "comp.sys.ibm.pc.hardware/60804": 0.00016748431132934368,
   "comp.sys.ibm.pc.hardware/60807": 0.0001672788313061959,
   "rec.sport.baseball/104999": 0.00016727462339863927,
   "talk.politics.guns/54554": 0.00016711390604117225,
   "rec.autos/103727": 0.00016682746278940823,
   "sci.electronics/54164": 0.00016614519670257933,
   "comp.graphics/39040": 0.00016609475983108524,
   "comp.sys.ibm.pc.hardware/60741": 0.00016602390729912364,

"comp.sys.mac.hardware/52443": 0.00016589129125978895,
"talk.religion.misc/84281": 0.00016571603004353363,
"sci.electronics/54208": 0.00016546178953426863,
"talk.politics.guns/54839": 0.00016538973585689018,
"comp.os.ms-windows.misc/10781": 0.00016515910447985463}

2. Using the tf-idf similarity, the top 20 documents retrieved were:
{"talk.politics.misc/178908": 0.00027019794138631034,
"talk.politics.misc/179058": 0.00026645828314927363,
"talk.politics.misc/179073": 0.00026142565728376856,
"sci.crypt/15812": 0.0002597349347825795,
"soc.religion.christian/21496": 0.00025804236798088995,
"talk.politics.mideast/77198": 0.00025783267079827494,
"alt.atheism/53639": 0.00025669350647066814,
"talk.politics.mideast/77195": 0.00025566671964323906,
"talk.religion.misc/84223": 0.00025354980397442745,
"alt.atheism/53538": 0.00025514894229138936,
"soc.religion.christian/21577": 0.00025510773465177114,
"soc.religion.christian/21458": 0.00025506429882677663,
"alt.atheism/54233": 0.00025017309216296 7,
"soc.religion.christian/21536": 0.00024932623720237896,
"talk.politics.misc/179034": 0.00024882568256 68481,
"talk.politics.mideast/77186": 0.00024781544 68254387,
"talk.politics.guns/55087": 0.00024779156693746 62,
"talk.religion.misc/84079": 0.00024717045106066414,
"soc.religion.christian/21597": 0.00024550687393827 97,
"talk.politics.misc/178786": 0.00024544835594481567}

CONCLUSIONS:

In the term-overlap method, it can be observed that the top-scoring documents have one thing in common- science. Articles retrieved which are based on religion(like talk.religion.misc/84349) have strongly based their arguments on science. Articles like rec.autos/103727 discuss “brake fluid”-

which is connected to science; articles like talk.politics.guns/55087 connect propane explosion and ammunition- again a strong topic of science.
In the tf-idf similarity method, it is observed that the top articles retrieved discuss religion and connect politics with religion. The article sci.crypt/15812 connects politics with human psychology- which again happens to be a strong area of religion.
Observing the above results, it can be said that the tf-idf method is more efficient. It produces a strongly connected set of articles at the top positions.