Project 2 - Report

Jiancong Lu and David Margolin

MapReduce Design:

Sub Problem 1 (TF-IDF Matrix)

Step 1

Map (to stripes by word)
"docId2 w1 w4 w3" -> (docId1, [w1, w2, w3, w1]) ->
(w1, {"docId1": 1/4})
(w2, {"docId1": 1/4})
(w3, {"docId1": 1/4})
(w1, {"docId1": 1/4})
* also filter out any word that doesn't match gene_xyz_gene

Reduce (combine same words)
(w1, {"docId1": 1/4})
(w1, {"docId2": 1/4})
(w3, {"docId1": 1/4})
(w1, {"docId1": 1/4}) ->

(w1, {"docId1": 2/4, "docid2": 1/4})
(w3, {"docId1": 1/4})

Step 2

Map (to tfidf)
(w1, {"docId1": 2/4, "docid2": 1/4}) -> (w1, {"docId1": 2/4*log(4/2), "docid2": 1/4*log(4/2)})

Sub Problem 2 (Cosine Similarity)

Step 1

Map (compute cosine similarity)
(w2, {"docId1": 1/4*log(4/1)}) -> (0.004545, w2)

Map (sort by key descending)
(0.004545, w2)
(0.000232, w4)
(0.023333, w3) ->

(0.023333, w3)
(0.004545, w2)
(0.000232, w4)

Map (to values)
(0.023333, w3)
(0.004545, w2)
(0.000232, w4) ->

w3, w2, w4

Potential Problems:

- Stripes design uses more memory than pairs (but is faster as the corpus grows)
- Cosine similarity can result in 0/0
- No normalization of data

Top 5 the most similar terms with the pattern gene_xyz_gene to the term "gene_egfr+_gene" from the new data set project2_egfr.txt

1. gene_epidermal_growth_factor_gene
2. gene_egf_gene
3. gene_egf_receptor_gene
4. gene_l858r_gene
5. gene_egfr_kinase_gene