BITS - Pilani, Hyderabad Campus

CS F469 IR Assignment - 1

Plagiarism Checker

Team Members:

1. Kushagra Gupta 2018A7PS0208H
2. Parveen 2018A7PS0623H
3. S. Sai Naga Shashank 2018AAPS0347H

- Method used to find similarity between a document and a query:

  $\text{Cos}\theta = \vec{D}.\vec{Q}$, where $\vec{D}$ and $\vec{Q}$ represents normalized vectors of the document and the query, respectively, containing 'tf-idf' value of each token

- Pre-processing has been done to find $\vec{D}$ for each doc in the corpus, beforehand.
- As a result we get a Matrix of the form:

	Doc-1	Doc-2	Doc-3	…	Doc-N
Term-1					
Term-2					
Term-3					
…					
Term-M					

Each cell has 'tf-idf' value of the Term-i for Doc-j

M – total number of terms in the whole corpus, N – total no. of documents in the corpus

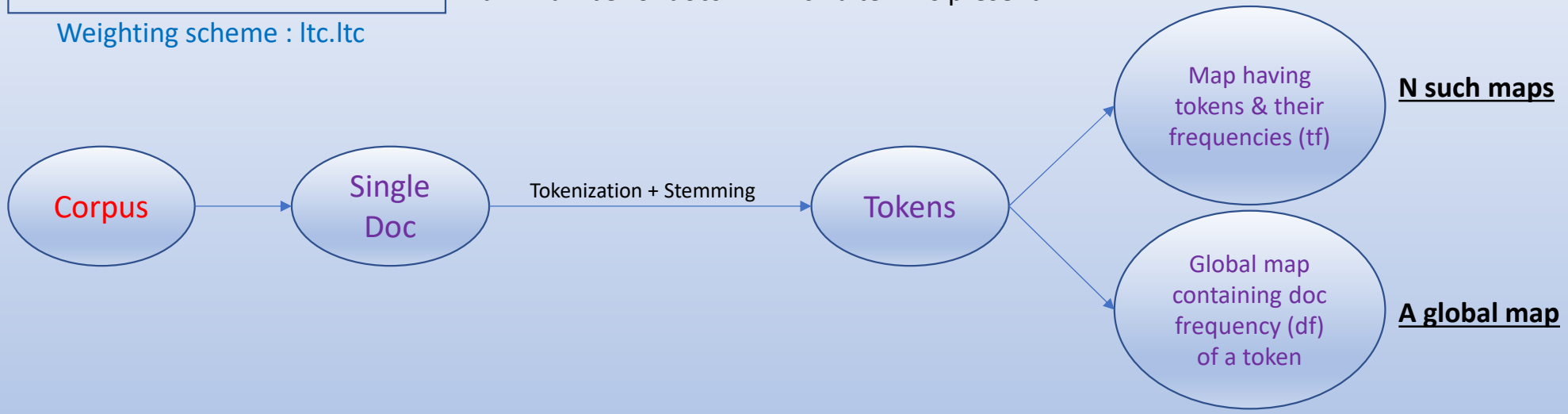
- But, since every doc doesn't contain every term, therefore this matrix is sparse!
- Therefore, instead of this matrix we used maps for each doc, which contains only the 'tf-idf' value of the terms present in that document only (***Memory Efficient***).

Pre-processing:

tf-idf = (1+log(tf)) * log(N/df)

Weighting scheme : ltc.ltc

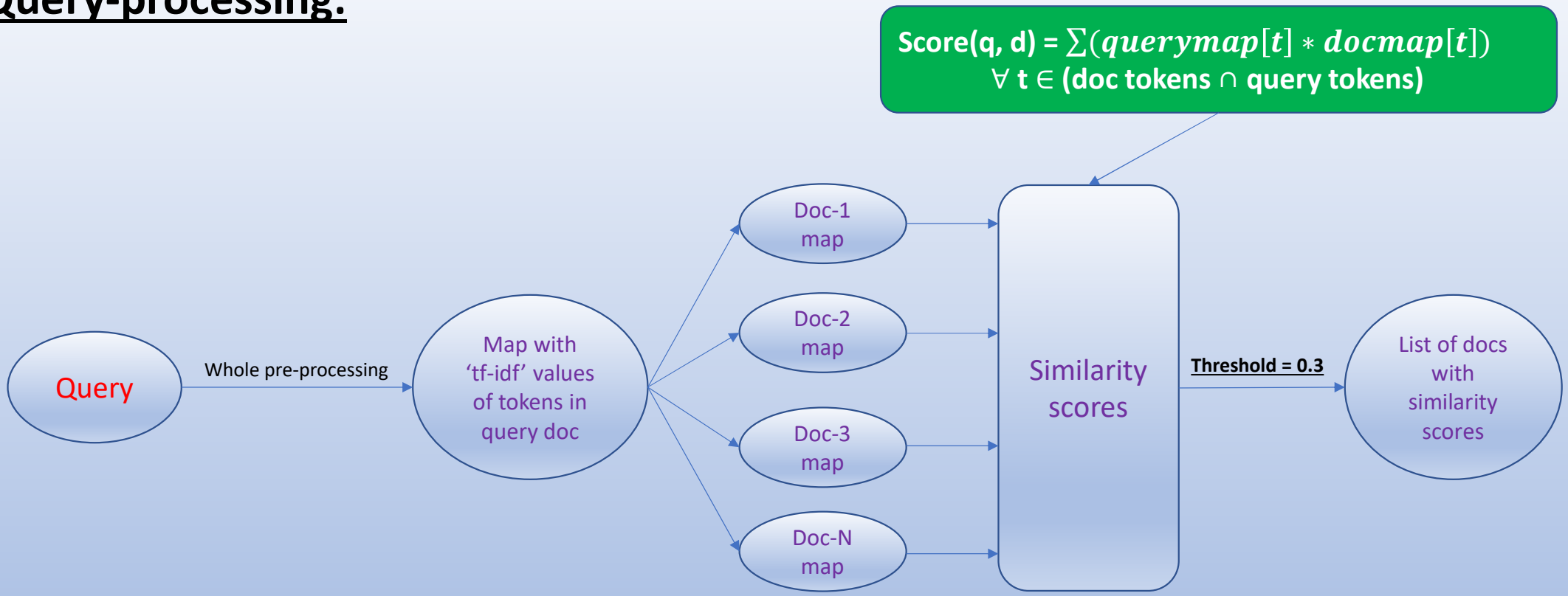
tf – term frequency of a term in a particular doc
df – number of docs in which a term is present



Since only 'tf-idf' values of a doc is required for similarity calculation, not the 'tf' and 'df' values, therefore -

- Global map is updated to have 'log(N/df)' value of a token
- and similarly, N maps are updated to have 'tf-idf' values of its tokens
- All (N+1) maps have been normalized to length = 1

Query-processing:



- Since Query-Map only have tokens of query doc, again it is memory efficient, as we are not required a vector with all tokens of the corpus!
- But Score Calculation may be slow because of the inability of maps to exploit inbuilt faster vector dot product!

Pros & Cons of the Model:

- **Since, it is a 'bag of words' model, therefore words proximity can't be handled here, and it can give False Positives as well.**
  For e.g. "A B C D" and "D B C A" are same for the model

- **But, as paraphrasing has been used to avoid plagiarism, therefore 'bag of words' model is successful in detecting sentence and paragraph paraphrasing.**

- **The model will not work if there is only 1 document in the corpus. Because in that case log(N/df) will be log(1/1) = 0 for each token. As a result, tf-idf and gradually Similarity Score will be 0 for every query!**

- **It will also not work if all docs in the corpus are similar. In that case also for most of the tokens 'tf-idf' will be zero and consequently the Score as well.**

Runtime:

- **Average Indexing time = 140 docs / second**
- **Average query time = 5 milliseconds / 100 docs**
- **Average has been calculated with up to 12800 docs in the corpus having around 150 words each.**