1111111111111111111111111111111111111111111111111111111111

National University of Computer & Emerging Sciences
FAST-Karachi Campus
Information Retrieval (CS317)
Quiz#2 (Sol)

Dated: March15, 2018 Marks: 30
Time: 25 min.
Std-ID: ____________________

Question NO. 1

What do we mean by Vector Space Model? What are some of the draw backs it has? [5]

Vector Space Model (VSM) is an abstract model for information retrieval. In this model, text-features like term, phrases are extracted from a collection and a vector space is created from these controlled vocabulary terms |V|. Each document is represented as a vector in $R^{|V|}$ space with coefficient represents weight for each term. Query is also represented as the vector in the same space. The retrieval of documents, is through computing similarity (distance) between documents and query in the same space

Drawbacks:

1. The order in which the terms appear in the document is lost in the vector space representation.
2. Theoretically assumed terms are statistically independent.
3. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".

Question NO.2

Consider a corpus of three documents, which comprises of Vocabulary = {w_1 w_2 w_3 w_4 w_5 w_6 w_7}, assume that the subscript dictate order of dimension:

d_1= {w_3 w_1 w_3 w_2 w_3 w_5}
d_2= {w_1 w_2 w_6 w_4}
d_3= {w_2 w_1 w_6 w_7}

Using the term frequency (TF) and Inverse Document Frequency (IDF) as log (N/df $_t$), where N=3, provide the document vectors in vector space model for the given documents.? [5]

	D1	D2	D3	IDF	tf*idf for D1	tf*idf for D2	tf*idf for D3
W1	1	1	1	0	0	0	0
W2	1	1	0	0	0	0	0
W3	3	0	1	0.477121	1.43136376	0	0.477121255
W4	0	1	0	0.477121	0	0.477121255	0
W5	1	0	0	0.477121	0.47712125	0	0
W6	0	1	1	0.176091	0	0.176091259	0.176091259
W7	0	0	1	0.477121	0	0	0.477121255

Vectors for documents:
d1: (0,0,1.431,0,0.477,0,0)
d2: (0,0,0,0.477,0,0.176,0)
d3: (0,0,0.477,0,0,0.176,0.477)

Question NO.3

Compare and contrast the following pairs of terms. [10]

Free Text Search	Parametric search
Free Text search are the search mechanism by which a query written in Natural Language without any structural constraints are used to search a document collection. User is free to pose any query.	In parametric search, a structural query is used to know the search parameters based on meta/non-meta descriptors. User is bound to provide few parameters-value for selected parameters.
Precision Critical Task	**Recall Critical Task**
Time matters a lot in these tasks Tolerance to missed documents Redundant and large number of resources Example: web search	Time matter less Non-tolerance to missed documents Less redundant very few resources Example: legal/patent search
Static snippets	**Dynamic snippets**
A static summary of a document is always the same, regardless of the query that hit the document. Generally, first few bytes are extracted from the document and stored as page descriptor / summary.	A dynamic summary is a query-dependent attempt to explain why the document was retrieved for the query at hand. Specific algorithms like KWIC to prepare a document summary on the fly using query terms.

Question No. 4

a. An information retrieval system returned 150 documents from a collection of 400 under a given query. There were 70% non-relevant documents in the retrieved result. The collection contains 35% relevant documents. Find the Precision and Recall for this system. [5]

	Rel	Non-Rel
Ret	45	105
Non-Ret	95	155

Total Returned Docs. =150;
70% non-relevant returned are 150X0.7= 105
The collection contains 35% relevant that is 400X0.35 =140 relevant documents, the query returned 45 relevant (150-105); 155 Non-relevant were not retrieved.
Precision = 45/150 = 0.3
Recall= 45/140 = 0.321

b. What is the F1 score for the system in part a.? [2.5]

F1 = (2*P*R)/ (P+R) = (2*0.3*0.321) / (0.3 + 0.321) = 0.1926 / 0.621 = 0.310

c. List some of the limitations of Normalized Discount Cumulative Gain (NDGC)? [2.5]

Normalized DCG metric does not penalize for bad documents in the result.
NDCG may not be suitable to measure performance of queries that may typically often have several equally good results.
NDCG does not penalize for missing documents in the result.