**Information Retrieval Final Exam UID: Sai Sri Malempati (U00973516)**

## 1. Vector Space model

Given the document collection D = {D1, D2, D3}

D1 => Betty Botter bought some butter

D2 => But the butter's bitter

D3 => The bitter butter makes the batter bitter

Assume that the stopword list contains the words {some, but, the}. The words will be stemmed, i.e., {bought -> buy, butter's -> butter, makes -> make}

**1.1 Show the uncompressed dictionary and the posting lists including the raw tf and idf values: raw tf is the raw term count and raw idf is (number of documents)/(document frequency). The terms are sorted in the dictionary order and the posting list is sorted by document id.**

The uncompressed dictionary after removed stopwords and stemmed in the below table:

| Term | Document ID |
|---|---|
| batter | 3 |
| betty | 1 |
| bitter | 2 |
| bitter | 3 |
| bitter | 3 |
| botter | 1 |
| butter | 1 |
| butter | 2 |
| butter | 3 |
| buy | 1 |
| make | 3 |

The posting lists including the raw tf and idf values is represented as below:

| Term | Posting list as : list of (DocumentID's, term frequency(tf)) | IDF (number of documents/ document frequency) |
|---|---|---|
| batter | [(D3,tf=1)] | 3/1 = 3 |
| betty | [(D1,tf=1)] | 3/1 = 3 |
| bitter | [(D2,tf=1) , (D3,tf=2)] | 3/2 = 1.5 |
| botter | [(D1,tf=1)] | 3/1 = 3 |
| butter | [(D1,tf=1) , (D2,tf=1) , (D3,tf=1)] | 3/3 = 1 |
| buy | [(D1,tf=1)] | 3/1 = 3 |
| make | [D3 ,tf=1)] | 3/1= 3 |

**1.2 If the terms in the query are sorted by IDF = log (raw idf), what are the terms likely used in scoring for the query: "bitter butter in the batter"? Briefly justify the answer.**

The terms likely used in scoring for the query: "bitter butter in the batter" are bitter, butter, in, batter. 'The' is in stopwords, so it gets removed from the query in preprocessing step. Terms in the query are represented in the below table which are sorted by IDF = log (raw idf). IDF of term 'in' is 0.

| **Query term** | **Log(raw IDF)** |
|---|---|
| butter | log(1)=0 |
| in | 0 |
| bitter | log(1.5)=0.17609125905 |
| batter | log(3)=0.47712125472 |

**1.3 If a query-independent document quality scoring scheme g(d) gives 1, 1.5, 0.5, for D1, D2, and D3, respectively, all posting lists will be sorted by the quality score in descending order and the cutoff value for the champion list is 0.8. Let the document vector weighting function be g(d) + raw tfidf based cosine similarity. What are the relevance scores and the final ranking list of the documents for the query "bitter batter"?**

The Document vectors based on weighting function created are :

First Document vector ( D1) = [0,0]

Second Document vector (D2)= [1.5,0]

Third Document vector (D3) = [3,3]

The Weighting schema for Query vector is raw tf-idf. The query vector (q)= [1.5,3]

We need to calculate scoring function g(d) + cosine-similarity(q, d) for all the documents, where g(d) is the document quality scoring schema given as 1,1,5,0.5 for D1,D2,D3 respectively.

Scoring function for D1= g(d) of D1 + cosine-similarity(q,d1)=1+0=1

Scoring function for D2= g(d) of D2 + cosine-similarity(q,d2)=1.5+0.447=1.947

Scoring function for D3= g(d) of D3 + cosine-similarity(q,d3)=0.5+0.9486=1.4486

Relevance score of D1=1 , Relevance score of D2=1.947 , Relevance score of D3=1.4486

The ranking list is :: D2>D3>D1 but D3 have quality score 0.5 which less than cutoff value of 0.8

Therefore, Final Ranking list of Document for the query is :: D2>D1

## 2. Learning to Rank

### 2.1: What are the advantages of pairwise learning-to-rank algorithms?

for the loss function, pairwise algorithms consider two documents i.e. a pair , with these pair of documents it does the optimal ordering and compares with the actual truth. The main goal of this ordering algorithm is to minimize the wrong ordering of documents relative to actual truth.

**Advantages :**

- The prediction of relative order instead of class label or relevance score is much near to the ranking nature.
- Pairwise learning to rank algorithm can be used on any binary classifier and works like binary classifier as well.
- Pairwise learning to rank algorithm minimize the number of inversions in the ranking results.
- It actively trains the learning to order items.
- Training process is not complex while compared to other ranking algorithms.
- Improper classification in raking which causes vector differences is minimized by the loss function in Pairwise learning to rank algorithm.
- pairwise learning to rank algorithm helps many online advertisings in ranking problems.

**References:**

https://wwwconference.org/www2009/pdf/T7A-LEARNING%20TO%20RANK%20TUTORIAL.pdf
learning-to-rank
microsoft-research-paper-ranking

### 2.2: If you are asked to evaluate the ranking quality of two learning-to-rank algorithms, X and Y, how will you design the experiment? Briefly describe the key steps.

Below are the steps to design for evaluating the ranking quality of two learning-to-rank algorithms, X and Y :

1. consider a collection of documents file and some queries file as well.
2. Have the ground truth for the all the query and documents i.e., the correct order of relevant documents .
3. Perform the preprocessing of removing stop words , Tokenization and stemming for all the documents
4. Create an inverted index with terms after preprocessing.
5. Perform Feature Extraction for dimensionality reduction and increase accuracy.
6. Generate training file and train X and Y models.
7. Apply cross validation on the models.
8. Run the queries on the model and save the results by comparing with ground truth.
9. Based on above results, compare and calculate MAP, Precision, DCG, NDCG, IDCG for the models X and Y.
10. Perform the Wilcox or t-test to get the p-values based on which we can evaluate the ranking quality.

**References:**

Lecture content.
evaluation-of-ranked-retrieval-results.html

## 3. Document Clustering

### 3.1: Can normalized mutual information (NMI) be used to determine the optimal number of clusters? Briefly justify your answer.

Clustering is the process of grouping the data which belongs to same type. optimal number of clusters means best value for number of clusters which are required for clustering the dataset. To identify the optimal number of clusters is one of major common issue in clustering. We use generally, k number of clusters for a flat clustering . There are different cluster evaluation methods like elbow, silhouette, gap statistics, Bayesian inference criterion, mutual Information .
Normalized mutual information is measure of mutual dependency between two variables. optimal number of clusters reaches when the clusters data remains same . There will be no mutual dependency in the clusters data which say that normalized mutual information becomes neutral at this point.
Hence Normalized mutual information cannot be used for determining the optimal number of clusters.

**References:**

Lecture content

NMI.pdf

evaluation-of-clustering

### 3.2: If you want to use cosine similarity as the similarity measure in clustering, can you still use the kmeans clustering algorithm? Briefly justify your answer.

In k-means clustering algorithm, we generate “k” number of clusters for given dataset. To find the centroids in k-means for datapoints in a cluster we use different types of distance measures. There is no single measurement which suits for all datasets in k-means. Inorder to find the best measurement for the dataset we need to use Principal Component analysis.

K-means, generally uses Euclidean distance as the similarity measure in clustering. Euclidean distance is used as similarity measure in k-means for smaller datasets. Cosine similarity measure is used for to find the angles or similarity between any two vectors. For larger size datasets or high dimensional datasets, cosine similarity is used as the similarity measure in k-means clustering, in that case k-means is called as spherical k-means. In spherical k-means we need to modify the calculation of distance with cosine similarity and centroids are normalized to unit length.

**References:**

Lecture content

http://datamining.rutgers.edu/publication/ICDM07K-means.pdf

https://www.researchgate.net/publication/4202779_Efficient_online_spherical_k-means_clustering

## 4. Document vectors

**We have discussed a simple method in the class for deriving a document vector with word vectors (i.e., word2vec). Please check the literature to summarize other possible methods for representing documents with vectors. Compare and discuss the methods you have found.**

Below are other possible methods for representing documents with vectors.

**Vector Model:**

Vector space model have another name as model of term vector. Vector space model is algebraic model which represents any documents (say text documents) in the form of vectors. In vector model the Queries and Documents are represented in form of vectors. Each dimension represents terms of document or query. if the term exists in document then the value is non-zero in vector otherwise its zero. We have different weighing techniques for these values like tf-idf, tf. Term weights are not binary in this model. This model allows to retrieve results by ranking documents which are partially matched .

**Disadvantages:**

- Search results on long documents are poor because of poor similarity values.
- There can be bad precision if the search words are not matched completely with terms.
- Order of terms appeared in document is lost in vector model.
- Statistical independency of the terms.
- The semantic information not preserved for the terms in vector model.
- Higher dimensional documents are complex to represent in vector model.

**Latent Semantic Indexing:**

Latent semantic indexing is a retrieval method and indexing which uses singular-value decomposition(SVD) for finding correlation of terms and concept of terms. In Latent Semantic Indexing, the terms are grouped based on the concept of words in similar context have similar meaning. Latent Semantic Indexing has multivariate statistical technique which is an application of correspondence analysis.

**Disadvantages:**

- To select number of features is problem in Latent semantic Indexing.
- Polysemy is captured partially in Latent Semantic Indexing.
- The Computation in costly in Latent Semantic Indexing.
- SVD representation consumes lot of space which cases storage problems.
- SVD representation of Latent semantic Indexing is more meant for normally distributed data

**Paragraph Vector:**

Paragraph vector is an extension of word2vec model. Paragraph vector represents the entire document in d-dimensional space latent which is an unsupervised learning approach. Distributed memory and distributed bag of words were two learning techniques for word vectors. Distributional memory model is for paragraph to vector model. Paragraph vector models uses paragraph id inorder to predict context of the terms. In Distributed bag of words, ordering of words does not matter. Paragraph vector will work on variable length of data documents.

**Disadvantages**:

- The computation is expensive compared to other models.
- The Paragraph vector is not considering the order of words.so it does not perform well in all cases.

**Latent Dirichlet Allocation:**

Latent Dirichlet allocation model is method of probabilistic clustering. The word Dirichlet means topics or context. So Latent Dirichlet allocation is a method which provides relationship of terms and Dirichlet.

The relationship is calculated using posterior probability(p[t/d]) for every term(t) in Terms and document(d) in Documents. Latent Dirichlet allocation assumes few topics where topics are nothing but set of words. It maps all documents to the topics which is done by mapping the words of the documents to the topics assumed. LDA ignores the ordering of words in the document. The document vector representation in LDA uses bag of words representation of documents. Latent Dirichlet Allocation is used in many recommendation applications like amazon recommendations, e-commerce etc.

**Disadvantages:**

The following are the limitations of LDA:

- The number of topics used in LDA must be known earlier
- The topics are fixed in LDA which results in if there any words in document which are relevant but does not fall under the topics known are ignored.
- In LDA, correlation between the topics are not taken to consideration
- In LDA the structure of sentence is not being modelled.

**References:**

Lectures content

wiki-Vector_space_model

wiki-Latent_semantic_analysis

doc2vect-paragraph2vec

LDA-towardsData

limitation-of-latent-dirichlet-allocation

Doc2Vec

## 5. Conversational search

**Read this paper "Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval", and write a research summary that include (1) The research problem and its significance, (2) The research challenges this paper wants to address, (3)A brief description of the approach, (4) Unique contributions, (5) Strengths and weaknesses**

**(1) The research problem and its significance:**

The main problem of this research paper is about identifying the desirable properties of conversational search . Conversational search is something different from traditional way of information retrieval. In conversational search have more like human and system interactions using natural language where it considers the entire string of words in query by returning natural language results. In this, the retrieval system is made to learn the properties based on the past works so that I can return best match results for information retrieval.

**(2) The research challenges this paper wants to address:**

The research challenge that paper address to building the computational model for conversational information retrieval. Based on user needs it has built conversational search system including the feedback from user which has to be in natural language. Any of existing conversational search systems are not good at understanding the user need and even asking for clarifications from users. In this

research it addresses the above problem by building a conversational search system which is self-learning based on users search data, handling conversations from both user and system back-and-forth. This conversational search system understands the context of the user as well.

**(3) A brief description of the approach:**

The Theoretical approach of conversational search defined in paper are mainly choice of interaction and action selection. The approach starts with user by giving statements for which the system takes the search actions. In this approach, instead of just retrieving the information it learns the retrieval with process of considering the earlier conversation with user. The system is ingested with earlier user data. In this approach, the user is requested for feedback and response is saved and used by the retrieval system for to give best response based on user inputs and within context of search. Here the main goal is getting the information what was needed by user based on conversation constructed. It is much less concerned with the way of representing the data. This research approach theoretical model is to satisfy the user exceptions with adding attributes from conversations.

**(4) Unique contributions**

The unique Contribution of the research paper is :

- In point of conversational search of data retrieval, it says that every attribute is desirable and important.
- The conversational search theoretical approach, it gives partial results in conversation to the user.
- The theoretical model approach which permit users to satisfy the specified properties, so this proving the practical implementation of conversational search.
- Any conversational search model should satisfy the four desired properties of User Revealment, Mixed Initiative, Memory and Set Retrieval, System Revealement. These properties provides a framework for to design the conversational search models.
- The paper shows best use cases which uses the five properties of theoretical model to retrieve desirable and user expected results.
- The research paper describes about the cost of function in decision making of retrieval process

**(5) Strengths and weaknesses**

The data which is not certain in the system is removed from retrieval process in conversational search. This reduces the cost of uncertain search in conversational search. The feedback mechanism in the conversational search shows strong impact n the result received. If the user gives relevant feedback, then the response for search results in near future will be more reliable to the excepted one. The research paper proposes and describes theoretical model of interaction choice and selection action. approach and their five desirable attributes to achieve the conversational search.

The conversational search proposed in the paper expects to learn from the user feedback. In some cases, if the user provides an irrelevant feedback then the retrieval gives wrong results. The conversational search works on the earlier communication feed to the system. If user not aware or not having information about what he wants, then the conversational search system cannot provide the search results. Security or data privacy is one of drawback in conversational search. In conversational search the system learns on the data or feedback of the user. So, any user who is not interested in sharing their data will not be interested in conversational search.