Text Clustering

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances.

The idea is that documents can be represented numerically as vectors of features. The similarity in text can be compared by measuring the distance between these feature vectors. Objects that are near each other should belong to the same cluster. Objects that are far from each other should belong to different clusters.

Essentially, text clustering involves three aspects:

- Selecting a suitable distance measure to identify the proximity of two feature vectors.
- A criterion function that tells us that we've got the best possible clusters and stop further processing.
- An algorithm to optimize the criterion function. A greedy algorithm will start with some initial clustering and refine the clusters iteratively.

Historical Context :

1971 - Vector space model was developed. Text documents are modelled as vectors. Features are considered to be the words in the document collection and feature values come from different term weighting schemes, the most popular of which is the Term Frequency-Inverse Document Frequency (TF-IDF).

1983 - Various clustering methods, including hierarchical and non-hierarchical methods are introduced. They show how clustering can be used to interpret large quantities of analytical data. They discuss how clustering is related to other pattern recognition techniques.

1992 - Partition-based clustering algorithms are developed. Buckshot method selects a small sample of documents to pre-cluster them using a standard clustering algorithm and assigns the rest of the documents to the clusters formed. Fractionation method finds k centres by initially breaking N documents into N/m buckets of a fixed size m > k. Each cluster is then treated as if it's an individual document and the whole process is repeated until there are only K clusters.

1997 – K modes is introduced, incremental update rule is introduced, but the optimal number of clusters is still found out by trial and error.

2008 - Hierarchal algorithm for document clustering are introduced. They use cluster overlapping phenomenon to design cluster merging criteria. The system computes the overlap rate in order to improve time efficiency.

Levels :

Document level: It serves to regroup documents about the same topic. Document clustering has applications in news articles, emails, search engines, etc.

Sentence level: It's used to cluster sentences derived from different documents. Tweet analysis is an example.

Word level: Word clusters are groups of words based on a common theme. The easiest way to build a cluster is by collecting synonyms for a particular word. For example, WordNet is a lexical database for the English language that groups English words into sets of synonyms called synsets.

Steps :

1. Text pre-processing: Text can be noisy, hiding information between stop words, inflexions and sparse representations. Pre-processing makes the dataset easier to work with.

2. Feature Extraction: One of the commonly used technique to extract the features from textual data is calculating the frequency of words/tokens in the document/corpus.

3. Clustering: We can then cluster different text documents based on the features we have generated.

4. Measuring Efficiency

Table 1. Comparison of Clustering Algorithms

ALGORITHM	Time Complexity	Similarity Criterion	Overlap	Advantages	Disadvantages
Single linkage	$O(n^2)$	Join clusters with most similar pair of documents	Crisp clusters	□ Sound theoretical properties □ Efficient implementations	□ Not suitable for poorly separated clusters □ Poor quality
Group Average	$O(n^2)$	Average pairwise similarity between all objects in the 2 clusters	Crisp clusters	High quality results	Expensive in large collections
Complete linkage	$O(n^2)$	Join cluster with least similar pair of documents	Crisp clusters	Good results (Voorhees alg.)	Not applicable in large datasets
Centroid/ Median HAC	$O(n^2)$	Join clusters with most similar centroids / medians	Crisp clusters		Small changes may cause large changes in the hierarchy
K-means	O(nkt) (k: initial clusters, t: iterations)	Euclidean or cosine metric	Crisp clusters	□Efficient (no sim matrix required) □Suitable for large datasets	Very sensitive to input parameters
Single Pass	O(nlogn)	If distance to closest centroid > threshold assign, else create new cluster	Crisp clusters	□Efficient □Simple	Results depend on the order of document presentation to the algorithm
Scatter/ Gather	Buckshot: O(kn) Fractionation: O(nm)	Hybrid: first partitional, then HAC	Crisp clusters	□Dynamic Clustering □Clusters presented with summaries □Fast	□Must have a very quick clustering algorithm □Focus on speed but not on accuracy
Suffix Tree Clustering	O(n)	Sim = 1 if |Bm)Bn|/|Bm| > threshold and |Bm)Bn|/|Bn| > threshold, else Sim = 0	Fuzzy clusters	□ Incremental □Captures the word sequence	□Snippets usually introduce noise □Snippets may not be a good description of a web page

Additional Questions:

1) What are Meta-Heuristic Optimization Algorithms and why are they used in text clustering?

Ans - A metaheuristic algorithm is a search procedure designed to find, a good solution to an optimization problem that is complex and difficult to solve to optimality. It is imperative to find a near-optimal solution based on imperfect or incomplete information in this real-world of limited resources (e.g., computational power and time).

Metaheuristics can often find good solutions with less computational effort than optimization algorithms, iterative methods, and simple greedy heuristics. They give approximate solutions.

Machine learning problems rely heavily on large datasets, and it is difficult to formulate the optimization problem to solve for optimality. Therefore, metaheuristics play a significant role in solving practical problems that are difficult to solve using conventional optimization methods.

2) What are some Meta-Heuristic Optimization Algorithms?

Ans –

a) Particle Swarm Optimization (PSO) algorithm is used as a feature selection technique for text document clustering in. The purpose of using PSO is to select the most informative features for representing each text document, thus improving the KM clustering algorithm performance.

b) Tabu search (TS), is another metaheuristic algorithm, is based on the memory structures and uses local search methods to find a potential solution by checking its neighbours to find a better solution. Generally, local search methods get stuck in suboptimal regions. TS enhances the search process by restricting the same solution's recurrence from coming back to previously visited solutions.

c) The genetic algorithm (GA) is a metaheuristic motivated by the evolutionary process of natural selection and natural genetics. The algorithm combines the fittest survival among string structures with a structured yet randomized information exchange to form a search algorithm. The key

aspect of genetic algorithms is the balance between efficiency and efficacy necessary for survival in different harsh competitive environments.

3) What is Spectral Clustering?

Ans - Spectral clustering is an EDA technique that reduces complex multidimensional datasets into clusters of similar data in rarer dimensions. The main outline is to cluster all the spectrum of unorganized data points into multiple groups based upon their uniqueness.

Spectral clustering is flexible and allows us to cluster non-graphical data as well. It makes no assumptions about the form of the clusters (unlike k means).