Topic modeling and Clustering

by

Raj Nath Patel
CDAC Mumbai

Topic modeling

- A topic model is a type of statistical model for “**discovering**” the abstract "**topics**" that occur in a collection of documents
- Topic models are a suite of algorithms that uncover the “**hidden thematic structure**” in document collections. These algorithms help us **develop** new ways to **search, browse** and summarize large archives of texts
- Topic models provide a simple way to analyze large volumes of **unlabeled** text. A "topic" consists of a **cluster** of words that **frequently** occur together

Clustering

- **Def1**: It deals with finding a structure in a collection of unlabeled data
- **Def2**: The process of organizing objects into groups whose members are similar in some way

Clustering

- **Hard Clustering:** If a certain datum belongs to a definite cluster then it could not be included in another cluster, Eg:K-means
- **Hierarchical Clustering:** A hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted
- **Soft Clustering:** Each point may belong to two or more clusters with different degrees of membership, Eg: Fuzzy C-means **(Topic Modeling)**

Output of Topic modeling

- **Cluster of words**: The list of words related to the same topic
- **Frequency of words**: Frequency of words in given any topic
- **Distribution of Topics**: How relevant a document is with the given topic

Steps in Topic model

- Tokenization
- Stop word removal
- Prepare dictionary
- Convert the data in required format
- Train the model
- Evaluate the model

How?

- Organizing data into clusters such that-
  - High intra-cluster similarity
  - Low inter-cluster similarity
  - Informally, finding natural groupings amongst objects

Possible Applications

- **Marketing**: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records
- **Libraries**: book ordering
- **Biology**: classification of plants and animals given their features
- **WWW**: document classification; clustering weblog data to discover groups of similar access patterns

Clustering algorithms

- **Exclusive Clustering:** If a certain datum belongs to a definite cluster then it could not be included in another cluster, Eg:K-means
- **Overlapping Clustering:** Each point may belong to two or more clusters with different degrees of membership, Eg: Fuzzy C-means
- **Hierarchical Clustering:** A hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted
- **Probabilistic Clustering:** Clustering use a completely probabilistic approach