SM in AI Project

Team 51

Document Classification using Topic Modelling

PROJECT DEFINITION :-

- Classify various documents into different clusters based on the Algorithms present in topic modelling.

SCOPE OF THE PROJECT

- Implementation of Topic modelling algorithms LDA , LSI , pLSI .
- Comparing accuracy of above algorithms in different cases.

Why Topic Modelling ?

- Bag-of-words model is commonly used in methods of document classification where the frequency of occurrence of each word is used as a feature for training a Classifier.
- Synonyms are captured into same features.
- Consumes large amount of data because of large feature data sets.
- Completion of process takes huge amount of time.

Why Topic Modelling ?

- Classifies different documents into clusters having set of unique features.
- Few features are used for training a classifier.
- Reduces computation time as the number of features is limited.

LDA
Latent Dirichlet Allocation

LDA

- LDA represent documents as mixtures of topics that split out with certain probability.
- Probabilistic Model with interpretable topics.

LDA : ASSUMPTIONS

- The number of words in the document follows Poisson Distribution.
- Assumes that a document is a mixture of a topic.

LDA :- ADVANTAGES

- We can infer the content spread of each sentence by a word count.
- We can derive the proportions that each word constitutes in given topics.
- Algorithm will check and update topic assignments.
- Handles synonyms up to some extent .

LDA : DISADVANTAGES

- Linear Model , not the best solution to handle nonlinear models
- Deciding on the number of topics is based on heuristics and need some expertise

LSA

Latent Semantic Analysis

- It is also known as Latent Semantic Indexing(LSI) literally means analyzing documents to find the underlying meaning or concept of those document.

Assumptions

LSA

- Words are assumed to have only one meaning
- Concepts are represented as pattern of words that usually appear together in documents
- Documents are represented as ‘bag of words’ ,order of the words in document not matters but frequencies matter

Advantages

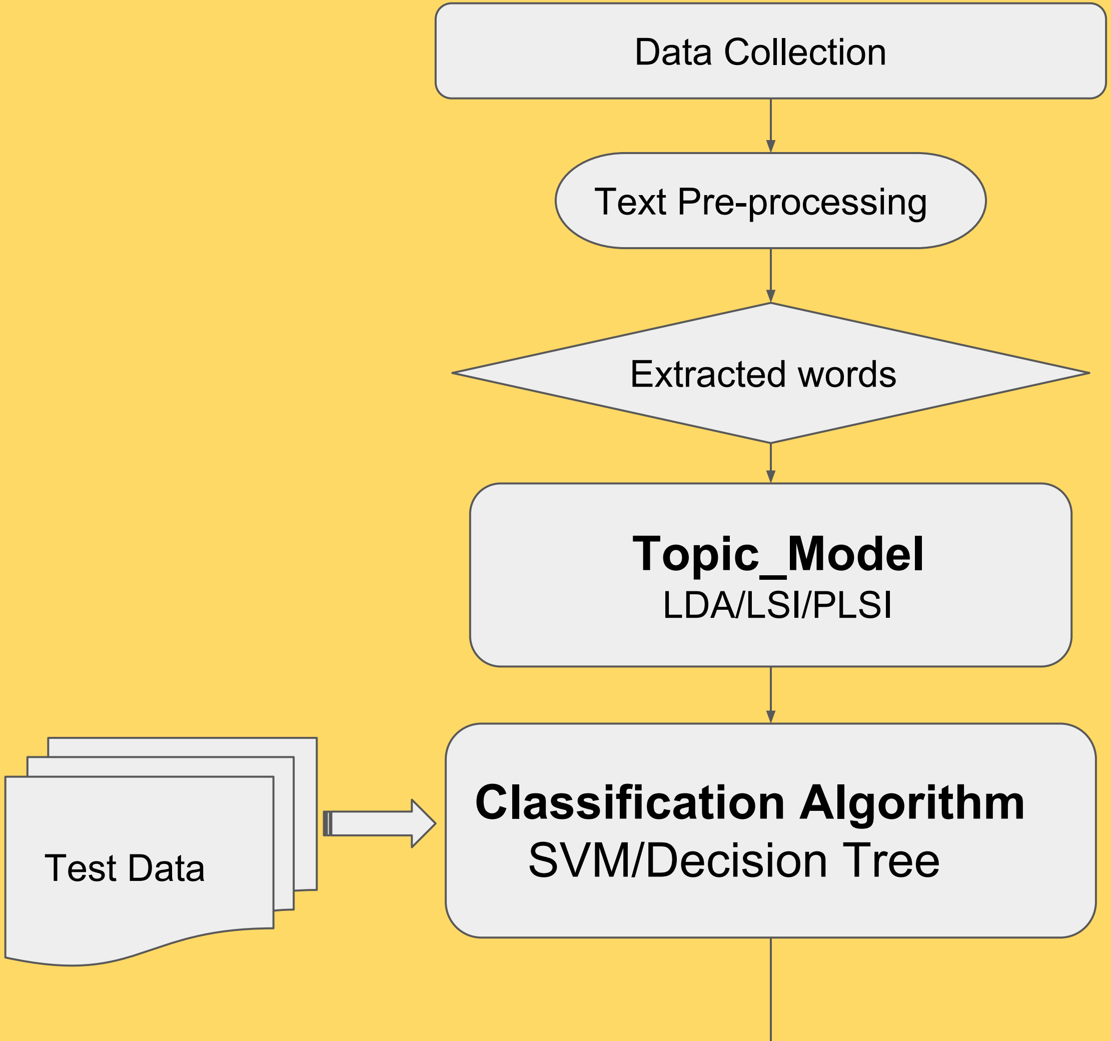
LSA

- Synonymy : LSA can handle synonymy problem(two different words having same meaning)
- The concept space has vastly lower dimension compared to original matrix

Disadvantages

LSA

- It can’t handle polysemy(words with multiple meaning)
- As it uses mathematical technique Singular value decomposition(SVD),it is computation expensive



Data Collection

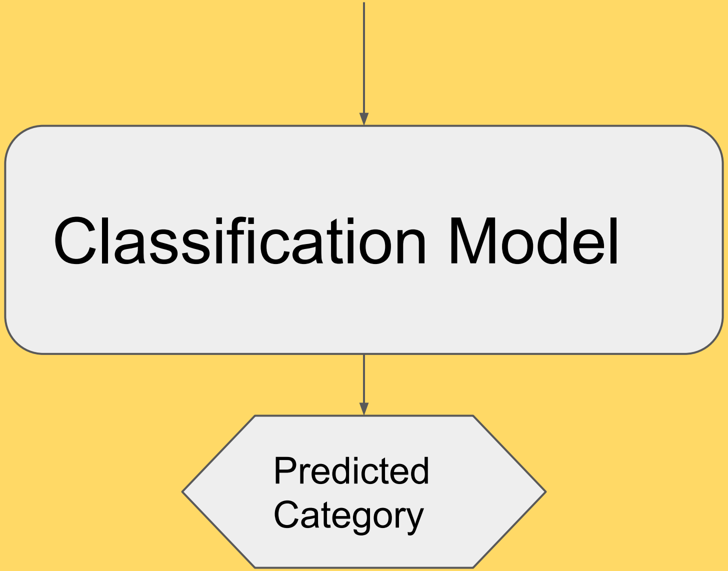
Text Pre-processing

Extracted words

Topic_Model
LDA/LSI/PLSI

Classification Algorithm
SVM/Decision Tree

Test Data



Classification Model

Predicted Category

TOOLS REQUIRED :

- Programming Languages :- R , Python .
- Java API for LDA,LSA
- Data collection(dmoz data)

PROJECT PLAN :

- 24 th Sep - Initial implementation of algorithms.
- 26 th oct - Classification of documents with accuracy.