

India ML Hiring Hackathon 2019

Problem Statement:

Identify Key aspects of a Review.

Format of a dataset:

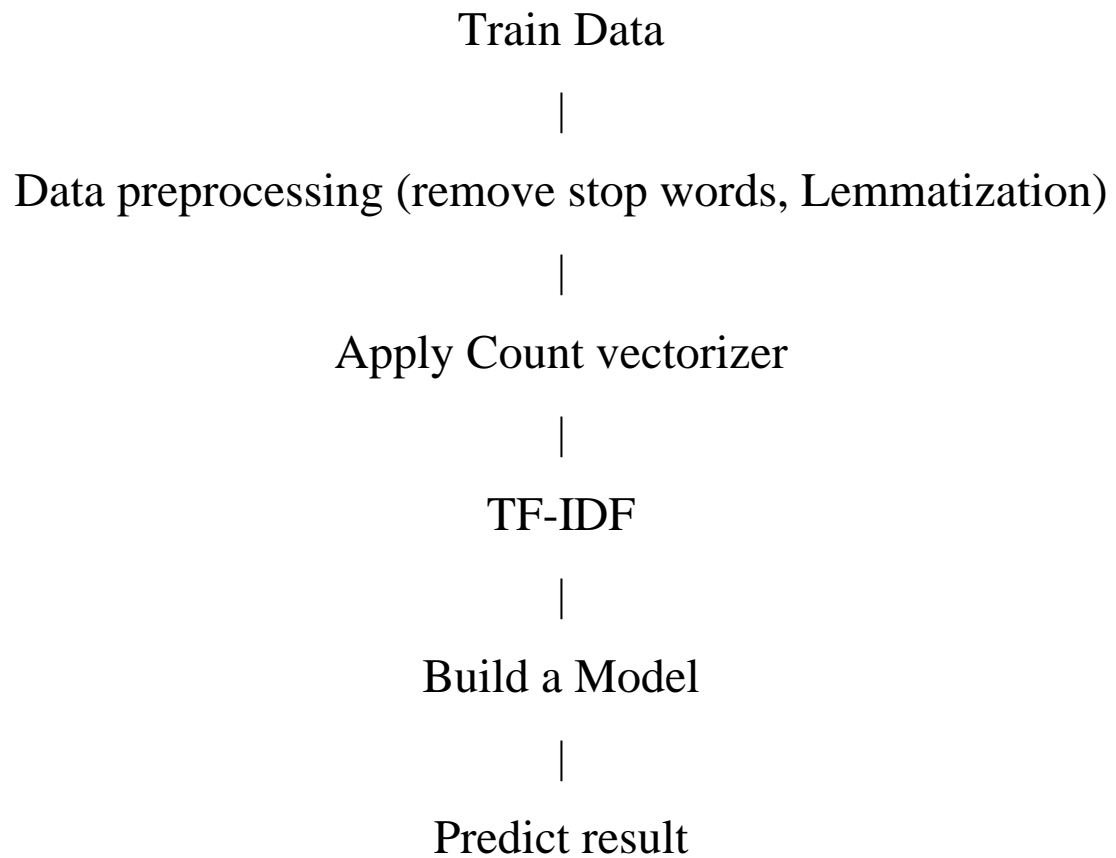
Train data:

| Variable | Definition |
|--------------|-------------------------------|
| Review Text | User's reviews |
| Review Title | Title of the reviews |
| Topic | Topic of the reviews (Target) |

Test data:

| Variable | Definition |
|--------------|----------------------|
| Review Text | User's reviews |
| Review Title | Title of the reviews |

Approch:



- **Data preprocessing:**

- Missing values: train and test data does not have any missing data.
- Stopwords: For the purpose of analyzing the data and building NLP model stopwords does not add much value to the meaning of the document.
- Lemmatization: It is the process to convert word to its base form.

- **Count vectorizer:**

- Reviews contains a series of words. To run the machine learning algorithm, we need to convert text into numerical feature vectors.
- segment each review into words and then count a number of times each word occurs in data and finally assign each word an integer id. This is known as feature vector.

- **TF – IDF:**

- Just by counting number of words in the dataset will give more weightage to common words. To avoid that I am using TF-IDF.
- TF-IDF (Term frequency inverse document frequency): is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling

- **Build models:**

- Build different Machine learning models with preprocessed train data.

- **Predict Results:**

- Predict the result of different models with test dataset.
- Compare results of different models.
- Make a final submission whose accuracy is more.

Model improvement:

- Improve accuracy of the model by tuning parameters using GridsearchCV.
- Merge Review Text and Review Title columns and train the data.

Other Approaches:

- Deep learning- Models with pretrained Embeddings
- Rasa NLU - Use Rasa NLU's intent classification