

Advanced BigData Analytics

Final Presentation

Group 05

Mubeen, Rashmi, Sandeep

Problem Statement

- Sentiment analysis of text is a challenging subject due to complexity of language and human expressions.
- NLP provides a way to make a machine determine the sentiment of a review based on the text.
- To analyze the sentiments represented in movie reviews from IMDB using NLP

Approach



Data set

- IMDB dataset – two sets:
 - labeledTrainData

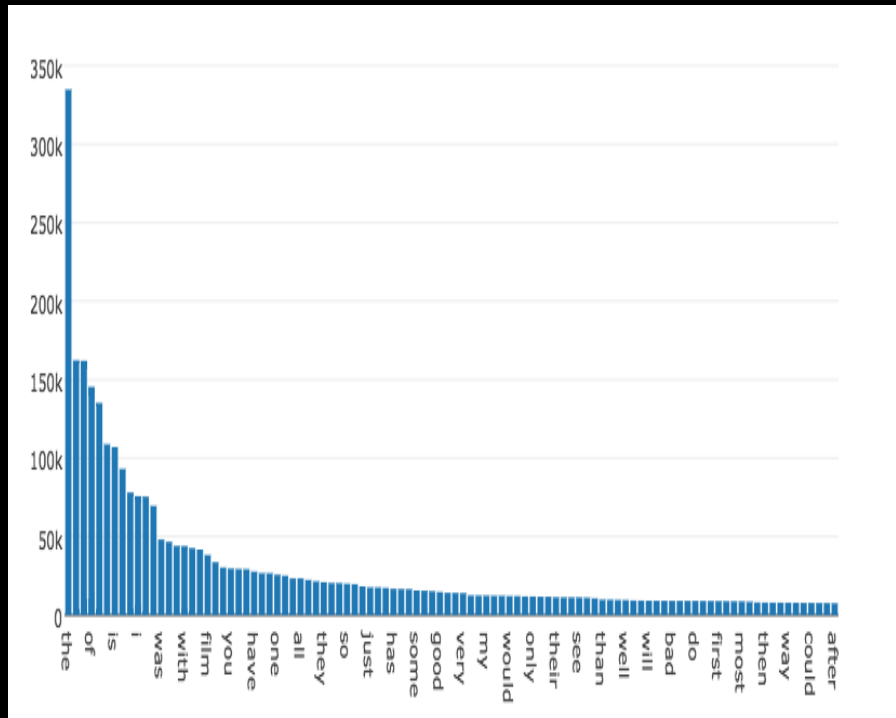
Column name	Data type	Description
ID	String	Unique Identifier
Sentiment	Number	0 Or 1 for neg and pos
Review	String	Review string

- UnlabeledData

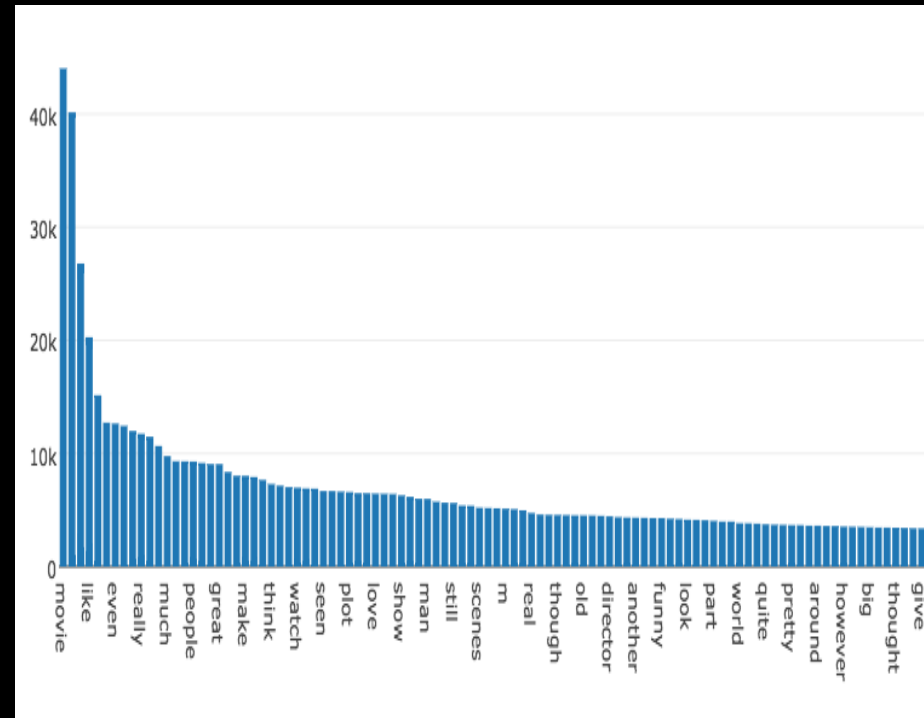
Column name	Data type	Description
ID	String	Unique Identifier
Review	String	Review string

Exploratory Data Analysis

Word Counts

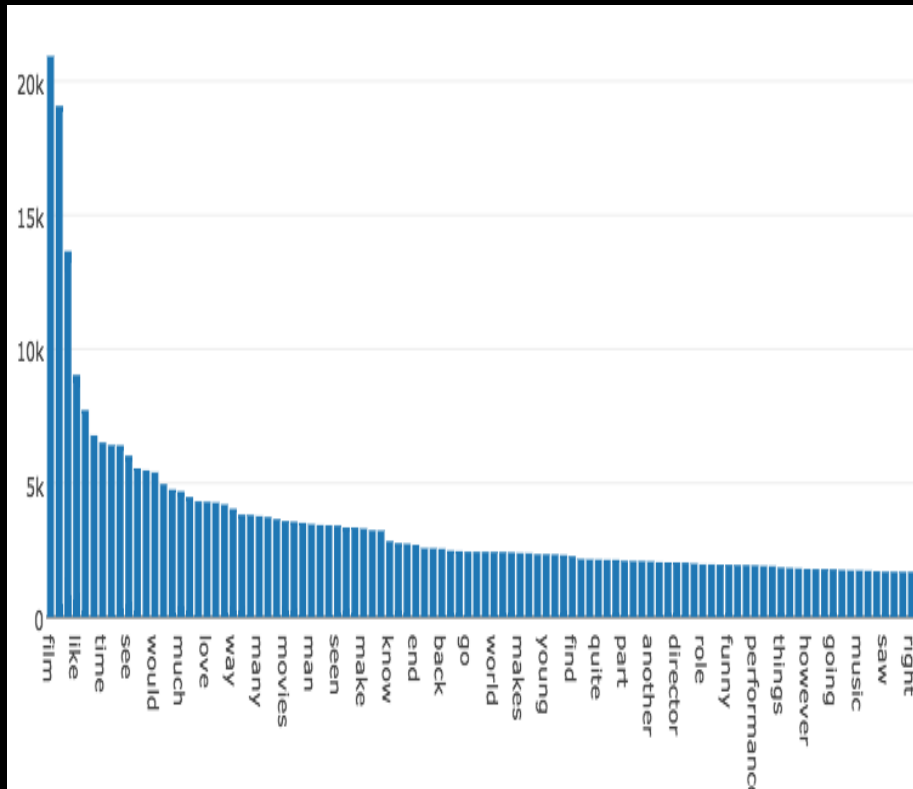


After Removing Stop Words

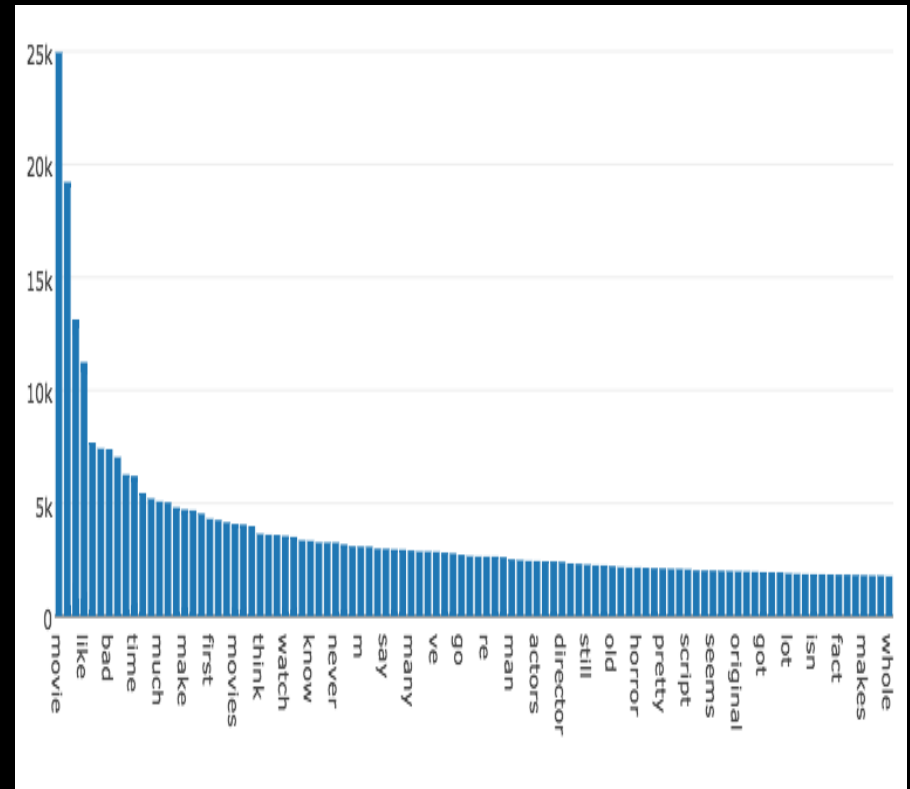


Exploratory Data Analysis

Positive Reviews



Negative Reviews



Feature Engineering

- TF_IDF
 - Creating vectors for text depending upon the occurrence and repetitions
- Word2Vec
 - Creation of vectors by training a model to understand the semantics of the text.

TF-IDF

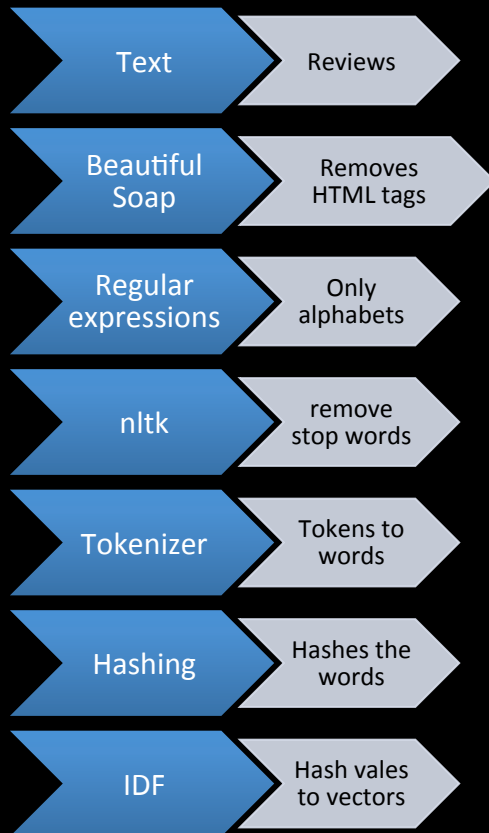
- Term Frequency:
 - The weight of a term that occurs in a document is simply proportional to the term frequency.
- Inverse Document Frequency
 - The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.

Word2Vec

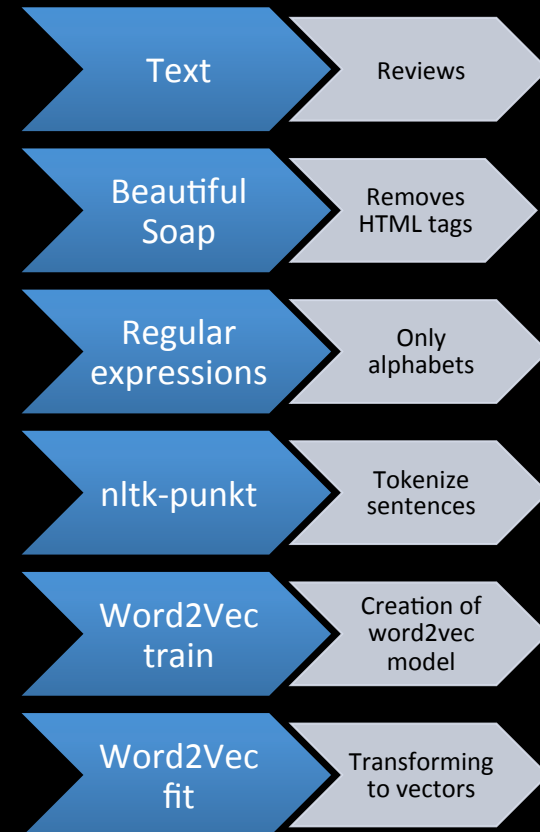
- Takes an word corpus as an input and generates vectors for each word.
- The model has to trained with text data for construction of the vocabulary and to learn vector representation of words.

Feature Engineering

TF-IDF



Word2Vec



Training Algorithms

- Naïve Bayes
- Random Forest
- Logistic Regression
- K-Means

Results

Algorithm	TF-IDF	
Naïve Bayes	Accuracy: 72%	Recall: 71.5
Random Forest	Accuracy: 55.6%	
Logistic Regression	Accuracy: 73.9%	AuROC: 0.8006
ML-Logistic Reg.	Accuracy: 75.6%	AuROC: 0.82.75
LR-LBFGS	Accuracy: 74.89%	

Algorithm	Word2Vec
Random Forest	Accuracy: 61.22%
LR-LBFGS	Accuracy: 84.31%
K-Means	WSSSE: 3347.68

Model Selection

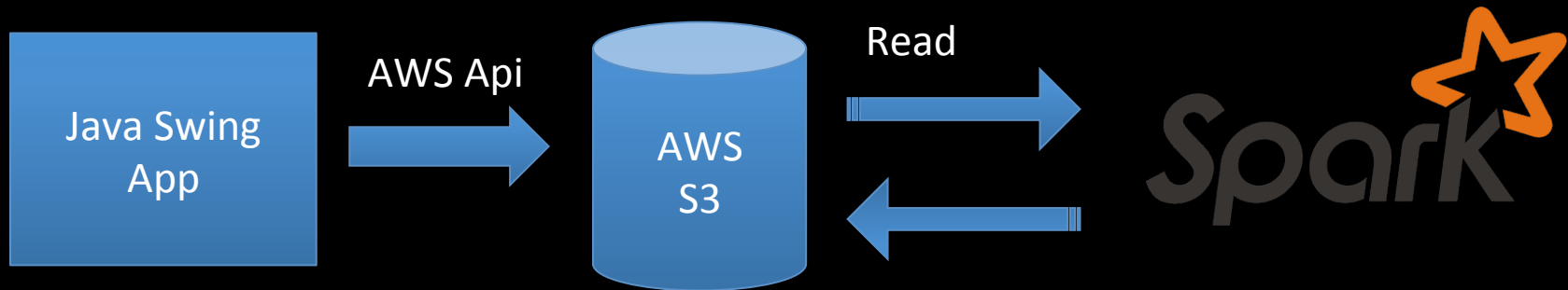
- Cross-Validator was used to determine the optimum Feature count and regression parameter value with TF-IDF as Feature engineering model.
- The AuROC value for Logistic Regression improved unto 84%

Model Selection


- The final mode selected was:
 - Word2Vec for feature engineering
 - # of vectors: 300
 - Classification Algorithm: Logistic Regression with LBFGS
 - Clustering Algorithm: K-Means with cluster = 2

Interface Model

- Java swing application for the UI
- Spark Read for connectivity to spark
- Amazon S3 for persistence
- AWS api for connectivity to S3



UI

 **AWS** ▾ **Services** ▾ **Edit** ▾

Rashmi ▾ Global ▾ Support ▾

Upload


Create Folder

Actions ▾

NonePropertiesTransfers

↻

[All Buckets](#) / [spark-sentimentanalysis](#)


	Name	Storage Class	Size	Last Modified
<input type="checkbox"/>	 file.txt	Standard	20 bytes	Sat Aug 22 11:25:50 GM

Transfers


✕

☐ Automatically clear finished transfers

✓ Done

✖ Delete: 

✓ Done

✖ Delete:  Deleting result from spark-sentimentanalysis

UI



A UI mockup of a review form. The form is contained within a light blue rectangular area. At the top left of this area are three small colored circles (red, yellow, green). The word "Review" is positioned to the left of a large white text input field with a blue border. The input field contains the text "This is a bad review" with a cursor at the end. Below the input field is a rounded rectangular button labeled "Submit".

Review

This is a bad review

Submit

Lessons Learnt

- Word2Vec model creation's accuracy increases with the vector size but it also increases amount of time it takes for the training and transformation
- The time and accuracy is also influenced by the text training dataset to the word2Vec model