**B9DA104\_1819\_TMD3**



|  |  |  |
| --- | --- | --- |
| **Course Title** | **:** | **Master’s in Data Analytics** |
| **Lecturer** | **:** | **Mr. Amir Sajad Esmaeily** |
| **Module** | **:** | **B9DA104 Machine Learning** |
| **CA Title** | **:** | **NLP with R** |
| **Student Name** | **:** | **Satyajit Saha** |
| **Student No** | **:** | **10505118** |
| **Date** | **:** | **July 2019** |

Definition of NLP:

* Natural Language Processing is a set of techniques through which computers and people can interact.
* It Basically refers to the language used by humans to communicate with each other. This communication can be verbal or textual.
* NLP is a field that brings together computer science, artificial intelligence and linguistics.
* If one would have to break down NLP to one equation it would be:
* NLP= computer science + Artificial Intelligence + Computational linguistics

Brief History:

* NLP is a field that started in the 1950’s (And the goal is to make machines understand our language.)
* In 1970 many programmers began to write “**conceptual ontologies**” with real world information into computer-understandable data
* In late 1980 there was revolution in natural language processing with introduction of Machine learning algorithm
* In 2010 representation learning and deep neural network style machine learning method became widespread in Natural Language Processing.

Application of NLP in real life:

* Natural Language Processing (NLP) is present in our daily lives.
* NLP is present in:
  + Auto –Predict (Google predict automatic search)
  + Sentiment analysis
  + Auto – correct (Google Keyboard)
  + Speech recognition (Google Web Speech)
  + Report generation in our Analytical tools
  + Spam filter (Google filter spam emails separately)
  + Question Answering (IBM Watson’s answer a query)

Mechanism of NLP:

When Text is composed of speech teste Text conversion. The mechanism of NLP involves in two process

1)Natural Language Understanding **(NLU**): It can understand the meaning of the given text

2)Natural Language Generation **(NLG):** It’s the process of Automatically producing text from structured data in a readable format with meaningful phrases and sentences

* NLG is a subset of NLP

Steps in NLP

* Tokenization
* Stemming
* Lemmatization
* Named Entry Recognition
* Chunking

Tokenization

* First step in NLP.
* It involves breaking a text documents into piece so machine can understand.
* Understand each word in respect to the sentences
* Produce a structural description on an input sentence

Stemming:

* Normalise words into it’s base form or root form
* It takes common prefix or suffix to detect key words
* Stemming is often done by an additional plug-in component to the index processing and a number of such components exist both commercial and open source.

Lemmatization

* It takes morphological analysis of words
* Groups together different inflected forms of a word, called Lemma
* Somehow similar to stemming as it maps several words into one common root
* Output should be proper word

Named Entry Recognition

* At the simplest named entities are people, place and things (products) mentioned in text document.
* NER is also known as entity chunking / extraction
* Popular technique used in information extraction to identify and classify the segments
* We have also trained NLP algorithm to recognize non standard entities such as tree or type of cancer.

Chunking

* Picking up group of information and grouping them into bigger pieces
* It works on top of POS tagging
* Part of speech tagging means identifying each token’s part of speech (noun, verb, adjective etc)
* It used POS tagging as input and provide chunk as output
* We need to correctly identify Parts of Speech in order to recognize entities, extract theme and process sentiment.
* Lexalytics has a highly robust model that can POS tag with > 90% accuracy.

Implication of NLP

* Technologies that used to build NLP pipelines are:

1) R (Covered topic)

2) Python

* + Natural Language Toolkit
  + Spacy

About Dataset

We have extracted the dataset from Kaggle. Below is the link-

<https://www.kaggle.com/vigneshwarsofficial/reviews/downloads/reviews.zip/1>

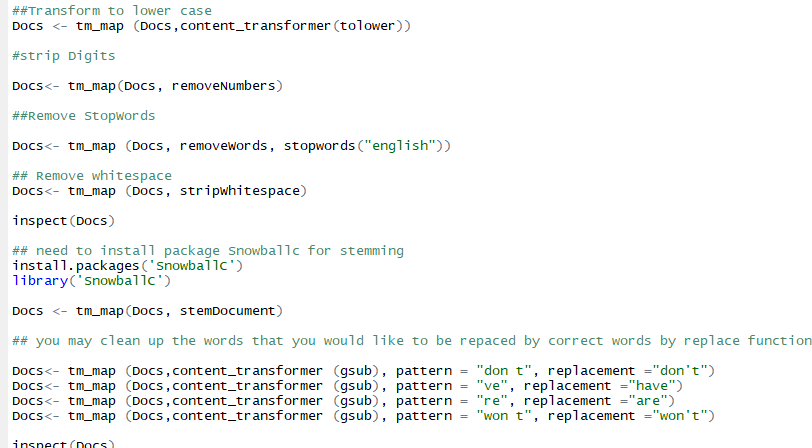
Most of the Restaurants ask reviews to the customers and based on the reviews the restaurant can improve the customer satisfaction. So Reviews plays a vital role for the successful growth of the restaurant.

The dataset consists of 1000 rows and 2 columns. Review Column consist of customer reviews and like column consist of 0 and 1. If the review is positive, 1 and if negative, 0.

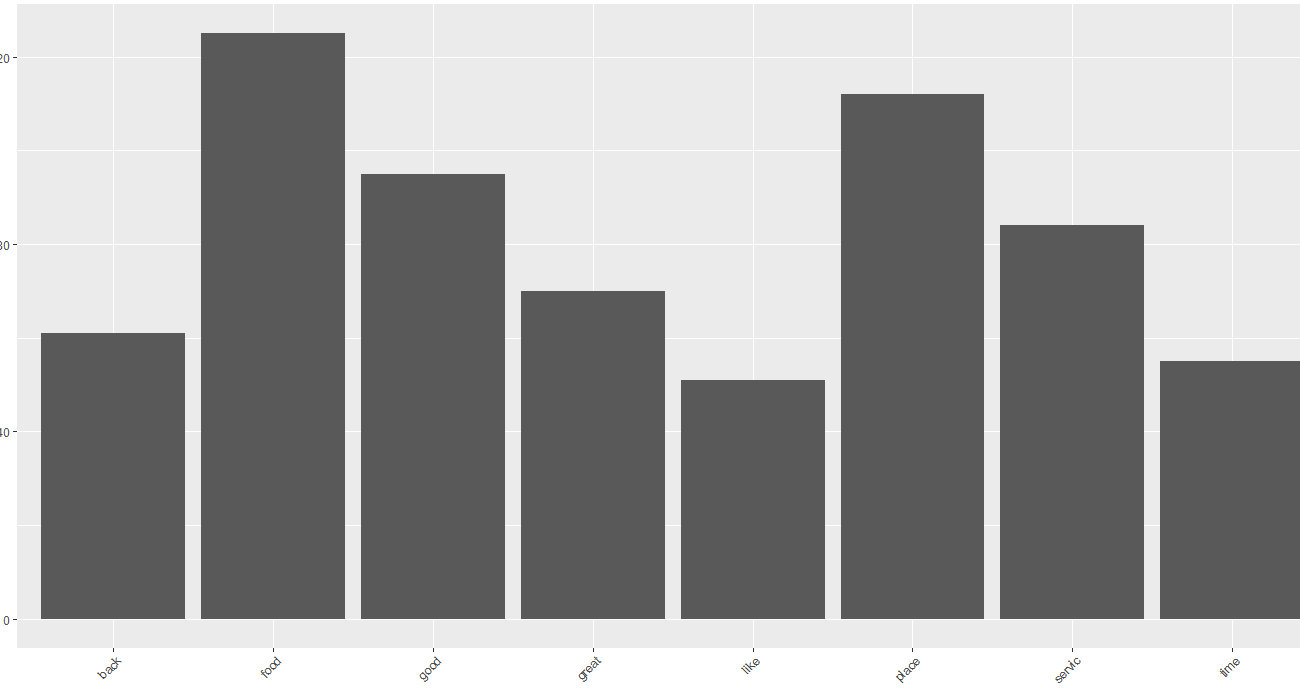
Scripting

I have made the script and described above each step in R programming language and executed the same.

As per algorithm is concerned, we have chosen Random forest classification for above dataset. I also attached R script with this document.



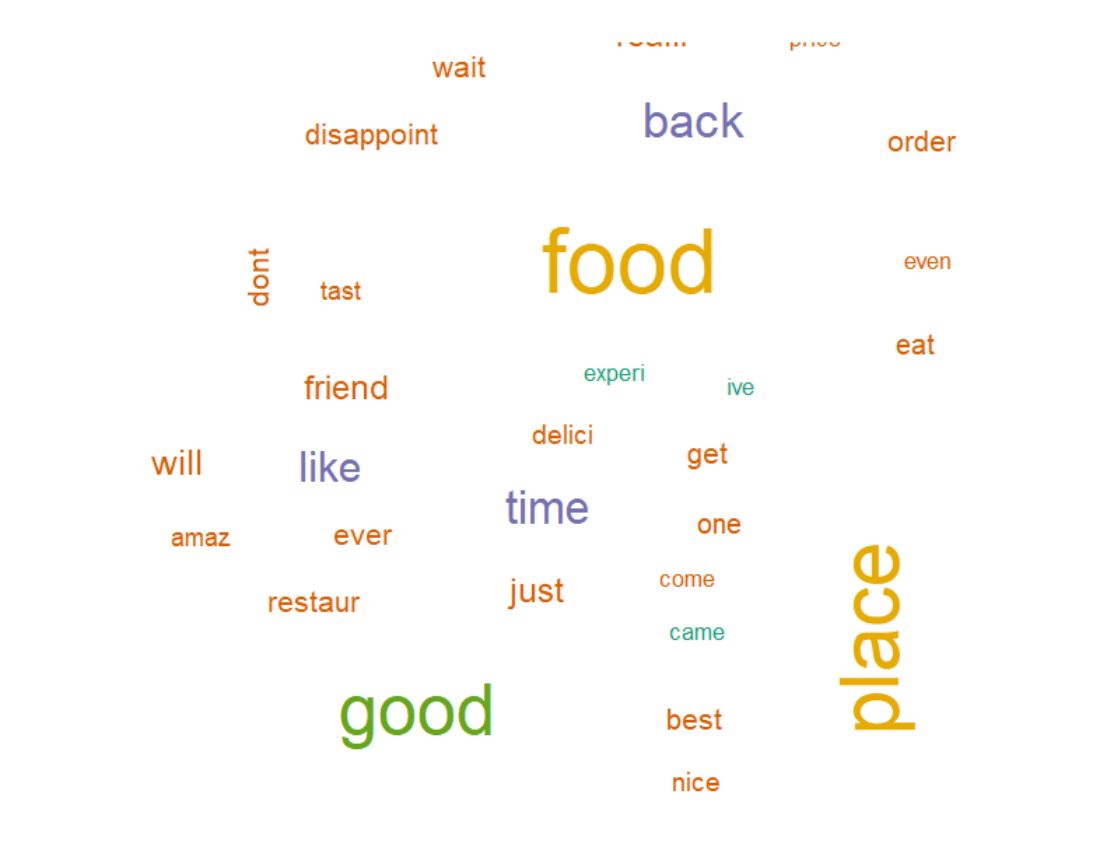
Analysis



Here we have used ggplot2 to make bar plot for mapping the words which has been used many times

Therefore, we can see that “food” has been used maximum numbers.

Word Cloud:



Here we have made word cloud to represent the words which have been used many times. Here also we can see that “food “is being used maximum times.

**Conclusion:**

Language is messy and complex. Meaning varies from speaker to speaker and listener to listener. Machine learning can be a good solution for analysing text data. In fact, it’s vital – purely rules-based text analytics is a dead-end. But it’s not enough to use a single type of machine learning model. Certain aspects of machine learning are very subjective.