**Advance Predictive Modelling:**

1. **How will you treat text having short cut words (like bcz, u, thr etc…)?**

TextBlob is the library for processing textual data. It provides a consistent API for diving into NLP tasks such as part-of-speech tagging, noun phrasing extraction, etc. So by using this library we can remove the shortcut words considering them as noun phrasing.

From my point of view we can also replace the shortcut words with exact words and then lemmatization.

From another point of view these shortcut words mostly are stop words or common words. So that can be removed.

1. **Write R and python code to replace “bcz” with “because” in whole text?**

Both languages are having very easy procedure to replace these short words with real words.

**Python:** We can use regex library, through which we can define types of pattern of strings we want to change.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

import regex as re

def remove\_words(my\_line):

new\_line =''

compiler\_bcz = re.compile(r"bcz")

for i in my\_line.split():

elif i in compiler\_bcz.findall(my\_line):

new\_line = new\_line + ' ' + 'because'

else:

new\_line = new\_line + ' ' + i

return new\_line

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**R Programming:**

“gsub()” is a function through which we can substitute various pattern of “because”. Sub is for “substitute and g is for global.

Syntax: gsub (pattern, replacement, string)

Exp: gsub(“bcz”, “because”, “Sun is hot **bcz** it’s a star”)

Output: Sun is hot **because** it’s a star

1. **How do you deal with the English text having Hindi words in between?**

This situation comes under ambiguous word uses. Mostly we people in India use Hindi words in between English words. So to handle this situation we have to understand the word.

Hindi is based on Devanagari script and English is is based on Roman script. So the meaning may differ completely.

So POS tagging analysis will be placed to understand the word. Then look up the sentiment scores from sentiwordnet of each languages (Hindi) based on their language tag.

Here we have to do sematic analysis which will find synthetic structure from the level of phrases, clauses, sentences. Then we search for any old verb or noun related to that word makes any meaningful sentences.

To find out the meaning there are sentiments (-ve, +ve and nutral). To find polarity of sentences a scoring algorithm is used.

And if the word is found to be meaningful then it comes under either in +ve or –ve directory.

Finally convert the words of features using ngrams with their corresponding sentiment scores.

**4. Write R code to connect with this public API -** [**http://www.omdbapi.com**](http://www.omdbapi.com)

There are various types of libraries through which we can connect to API.

Libraries: “httr”,”jsonlite”

Exp:

Library (httr)

New\_data<- GET (“<http://www.omdbapi.com>”)

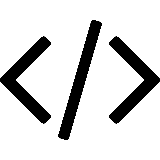
We can apply query, header and content as required from the API by changing the GET method content.

1. **What are the different methods to deploy a model into production system?**

Different method to deploy model:

**Method 1:** Data scientist created model by using Python or R. Then it goes to the Development team to redeploy in the front end using Java, C++. Then configure the model according to the API and publish it.



****

****

**Method 2:** Model directly can be used applying on Serialized Object then python code.

Serialized Object

(Python)

EndPont



