Explain your approach to create an analytics solution for understanding the Twitter users with a good data pipeline. (Use the Tweets.csv). Assume the necessary details and state the assumptions. (This is a descriptive question which can be answered in a .doc format or hand written).

1) Analytics is nothing but actionable insights for profits. The skills required for test Analytics can be explained as below:

2) Text Analytics is the process of converting unstructured text to structured form and applying statistical/AI methods to discover hidden patterns, predict the classes/labels as required. In addition, summarization to understand the text faster is also a part of text Analytics.

3) The process of Text Analytics consists of 4 stages:

In the first stage, the unstructured data is preprocessed and converted into standard format for analysis.

In the second stage, the different models are developed for various requirements as clustering, classification, information extraction, Topics detection and summarization.

In the third stage, the discovered knowledge is presented as the requirements in the fourth stage. The developed models will be used for scoring unseen data.

The fourth step requires first and second step similar to third step of presentation.

4) Requirements for text Analytics can be summarized as follows:

A. Ability to read text data and preprocess.

B. Classify the text automatically.

C. Group the text/documents by similarity.

D. Detect the topics in texts.

E. Analyze the social media for sentiment.

F. Analyze the social media for network/communities

5) Different preprocessing steps are as follows:

i. Tokenizing ii. Stemming iii. Lemmatization iv. POS tagging v. N-gram tagging vi. Stop words vii. Normalization viii. Text Cleaning-punctuation, symbols, gaps etc. ix. Text enrichment-emoticons, exclamations x. Lexicons for Knowledge based analysis of texts. xi. Bow format for text analysis. xii. TFIDF format for text analysis xiii. Word2Vector/Embedding format for Text Analysis xiv. Text frequency Analysis-unigrams, bigrams xv. Semantic analysis xvi. Storing processed texts xvii. Web scraping xviii. Visualizing Text analysis results xix. Creating pipelines for Analysis.

6)

The way to tokenize is using split and replace.

Different step for Text Analysis:

A) Access the data.

B) Next step is clean.

C) Tokenization (using replace and split)

D) Analytics is nothing but count and compare.

Normalization here is nothing but convert everything into uppercase or convert everything to lowercase. Normalization should be done before stop words removal .otherwise say if we have a word xyz.it can be in the form say for example XYZ,Xyz,xyz,Xyz,xyZ.normaliztion should be done before stop words removal otherwise duplicates may be present for word xyz and that might not be removed by SWR i.e. stop words removal.

7) The steps of Text Analytics can be summarized as below:

i. Corpus ii. Access iii. Clean iv. Tokenize v. Normalize vi. Stop words vii. Text Analysis-explore(Tabular,Charts-graph,Word cloud)

The first step in Text Analytics will be to explore. Any exploration in Text Analytics should start with Tabular presentation and then the graphical presentation followed by the Word cloud.

After completing the steps as discussed above, we get vocabulary or word vectors.

We finally get BAG OF WORDS as a result of doing the process as discussed above.The hierarchy to be followed is Sentiments then topics and then corpus.The value of Text Analytics is in CORPUS-how strong is your corpus. Corpus is nothing but Repository. The first thing in following explore is count then compare then concentration and then the fourth one is trend.Finally we should get the refined corpus from the steps followed as shown below:

8) Different approaches that will be available for text Analytics will be RULE BASED and ML BASED.

ML based-it tends to optimize the classification accuracy.

Rule enforced algorithms -It doesn’t learn the algorithms at one shot. It makes the mistake and learns and then makes another mistake and then learns more. LEXICON BASED ALGORITHMS will come under rule based.

Various categories of learning can be categorized as below:

A) Lexicon Based learning; B) Machine Learning; C) DEEP Learning; D) Rule enforced Learning or self Learning

In ML based:

One is Machine learning-independent models, Linear Models, Non Linear Models;

second is Deep Learning; third is Reinforced Learning.

Again Text Analytics can be :

A) Supervised Learning- Neural Networks, Naive bayes,SVM,Deep Learning, Machine Learning comes under this category.

B) Unsupervised Learning- SVD,LSA,LDA,KMeans,,SOM,principal component Analysis,Lexicon,Reinforced Learning comes under this category.

Next thing we need to look at tools/Technologies-Python,SKLEARN,NLTK,SPARK AND KERAS.

Next we need to understand the Task needed to be done which can be categorized into 3 types-

A)classification-In classification it can be sentiments, opinions , FAKE NEWS, Hate speech,

B)Topic Detection or Topic Modeling

C) SUMMARIZATION

9)The pipeline in text Analytics-

First it will be problem Definition-

then it will be CORPUS CREATION-under Corpus creation comes Automated ANNOTATIONS.

then it will be Text PREPROCESSING-Access,cleaning,Tokenization,STOP WORDS REMOVAL(SWR),Normalization,POS-Tagging,Lemmatization,Unigrams,BIGRAMS,TRIGRAMS,N-GRAMS comes under this.

Then we need to EXPLORE-TABULAR, Graphical, Word cloud Analysis, Word Frequency, and Word Chart.

Then it will be Modelling-Classification,Topic Modelling,Summarization comes under these.

Then it will be Evaluate-Accuracy, confusion Matrix, Region under curve, ROC will come under these. Then DEPLOY.