# Experiment A7

Title- Text Analytics

Aim- 1. Extract sample document & apply following document preprocessing method. Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.

2. Create representation of document by calculating Term frequency & Inverse document frequency.

Theory- Text analysis also called text analytics refers to the representation, processing & modelling of textual data to derive useful insights.
Text databases are rapidly growing due to the increasing amount of information available in technique electronic form. Traditional information retrieval techniques becomes inadequate for the increasingly popular & essential inadequate for the increasingly popular & essential part of data mining. The discovery of proper patterns & analyzing the text document from the huge volume of data is a major issue in real-world application

In text analysis, we take the help of corpus. A corpus is a large collection of text used for various purpose in natural language processing. few example of corpus are shakespeare Brown corpus, Google N-grams corpus.

Steps in Text Analysis-

① Parsing - It is the process that takes unstructured data (text) & imposes a structure for further analysis. The unstructured data could be plain, text file weblog, xml, etc parsing deconstructs the provided text and renders it in more structured way.

② Search & retrival - It is the identification of the documents in a corpus that contains search items like specific words, phrases, topic or entities, search items are generally called key terms.

③ Text mining - It uses the terms & indices produced by the previous two steps to discover meaningful insights pertaining to domains or problems of interest. Clustering & classification techniques can also be applied to derive insights.

Text - Pre-Processing Techniques

① Tokenization - It is the process of splitting a text into a list of tokens. They work by seperating the words using punctuation & spaces. It does not discard the punctuation.

Example -

```
from nltk.tokenize import word_tokenize
text = "Hello world!"
word = tokenize(text)
```

output - ['Hello', 'world', '!']

② Pos tagging - Part of speech tagging explains how a word is used in sentence It is used to extract relationships between words. It tells if a word is noun, pronoun, objective, etc.

Input :- "You gave me a car"

Output - [('You', 'PRP'), ('gave', 'VBD'), ('me', 'PRP'), ('a', 'DT') ('car', 'NN')]

PRP - Personal Pronoun  DT - Determine
VBD - verb Past tense  NN - Noun

③ Stop words removal - Stopwords are common words that are mostly irrelevant for text analysis, like 'a', 'or', 'the', 'is', 'for', etc.

Example - Input - I a an apple
Output - I are apple

④ Stemming - It is reducing words to their base or root form by removing a few suffices from the words. It is text normalization technique the Poster stemming algorithm is mostly used for this. However stemming does not provide correct form.

| Input : | Acting | Books | learning | obesity |
|---|---|---|---|---|
| output : | Act | Book | learn | obes |

* Term frequency (TF): It measures the frequency of a word in a given document. It highly depends on the length of the document & the generality

of words:

$$Tf = \frac{\text{count of the term in document (T)}}{\text{total number of term in document}}$$

⑤ Lemmetization - It is similar to stemming technique that aims to get the base words. But unlike stemming that might produce improper words. The lemmetization process does not only turn the suffics but also use lexical knowledge base.

| | | | |
|---|---|---|---|
| Input: | learning | caring | causes |
| output: | learn | care | cause |

* Inverse document frequency (IDF) - It is measure of the importance of word. IDF provides weightage to each word based on its frequency in the corpus. D.

$$IDF = \log\left(\frac{\text{Total no of document N in Groups e}}{\text{number of documents containing t}}\right)$$

Conclusion -

The representation of sample document & text - Preprocessing has been implemented with success.

## Experiment A8

Title - Data visualization I

Aim - 1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows & contains information about the passengers who boarded the information titanic ship. Use the seaborn library to see if we can find any pattern in data.
2. write a code to check how the price of ticket (column name : 'fare') for each passenger is distributed by plotting a histogram.

Theory -

- Data visualization is actually a set of data points & information that are represented graphically to make it easy & quick for user to understand.
- Data visualization is good if it has a clear meaning, purpose & is very easy to interprete without provide an accessible way to see & understand trends, outliers patterns. In data by visual effects or elements such as chart, graph & maps.

Characteristics of Effective Graphical visual -

1) It shows or visualizes data very clearly in an understandable manner.

2) It encourage viewers to compose different pieces of data
3) It closely integrates statistical & verbal descriptions of dataset.
4) It also helps in identifying area that needs more attention & improvement.
5) Using graphical representation, a story can be told more efficiently. Also, it requires less time to understand picture than it takes to understand textual data.

Types of data visualization-

① comparitive plots- They are used for comparing the data points. They are - i. Column & big charts
ii. Line chart
iii. Area chart
iv. Bubble chart
v. Pie chart

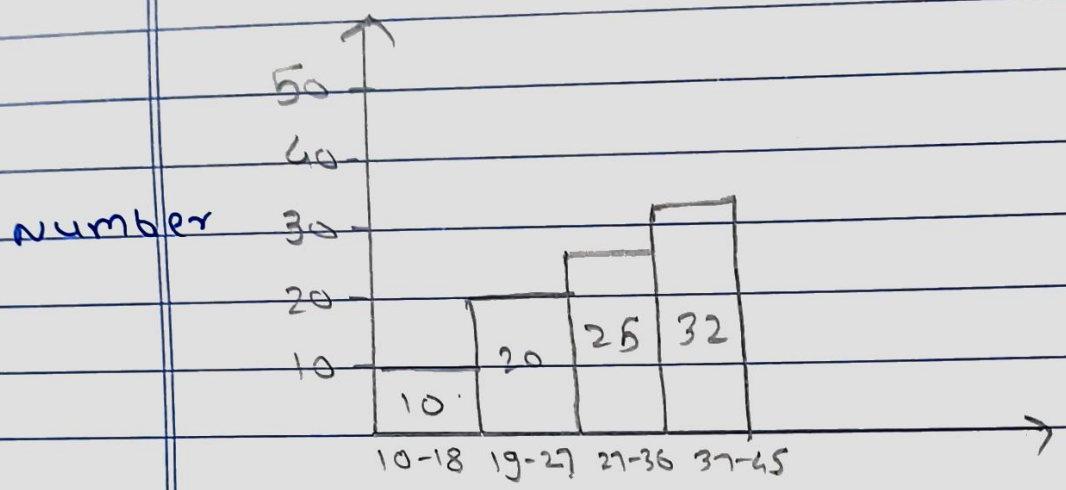
② Statistical Plots- They are useful to show results of statistical analysis.
They are - i. Histogram
ii. Tree map
iii. Bar plot
iv. waterfall chart

③ Topology Plots- It uses geometric structures to show relationships of connections between datapoints in your dataset.
They are - ① Linear Topology
② Graph

④ Spatial Plots- They are logical views for visualizing data which represents a map, location, shape etc.
They are - ① Raster Surface
② Point map
③ Heat map

Histogram-
A histogram is used to represent the frequency of data in a dataset as well its distribution. It looks similar to a column chart but is different as it plots the frequency for each distribution rather than the actual value of any datapoint - itself.
Example -

| Age group | Number of People |
|---|---|
| 10-18 | 10 |
| 19-27 | 20 |
| 28-36 | 25 |
| 37-45 | 32 |




Seaborn Library-

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization & exploratory data analysis.

Seaborn work easily with dataframe & the Pandas library.

Conclusion -

We use the seaborn library to see if we can find patterns in data & to check price distribution.

# Experiment - A9

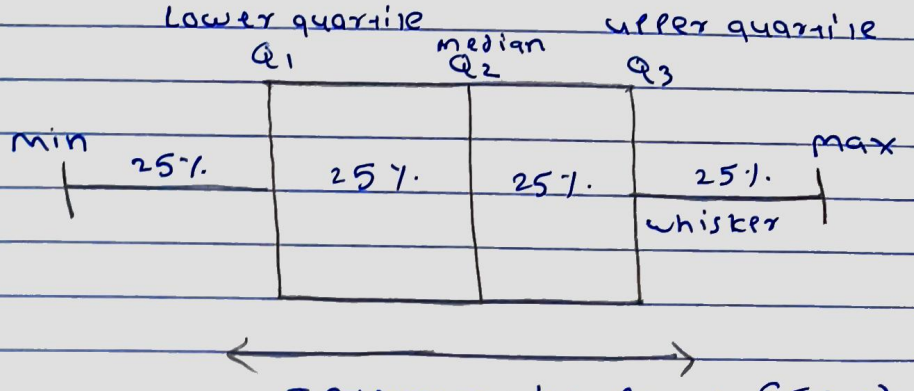
Title- Data visualization II

Aim- Use the inbuilt dataset 'titanic' as used in the above Problem. Plot a bar plot for distribution of age with respect to each gender along with the information about whether they survived or not ('sex' and 'age'). Write observations on the inference from the above statistics.

Theory: Box plot or whisker diagram shows how data points in a dataset are distributed this plot is quite interesting as it shows several statistical values of once

It depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

Components of Box Plot:



a) minimum - It is the minimum value in the dataset excluding the outliers.

b) First quartile ($Q_1$) - 25% of the data lies below the first or lower quartile.

c) median ($Q_2$) - It is the mid-point of the dataset. Half of the values lie below the medium and half above.

d) Third quartile ($Q_3$) - 75% of the data lies below the third or upper quartile.

e) maximum - It is the maximum value in the dataset excluding the outliers.

f) Interquartile Range (IQR) - The area inside the box (50% of the data) is known as the IQR.

$$IQR = Q_3 - Q_1$$

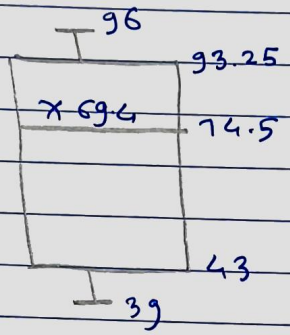
g) Outliers - They are the data points below and above the lower and upper limit.

The box plots can have skewness and the median might not be at the center of the box.

Example - Below is the distribution of students and their score in science

| Student | Ajay | Sunil | Ravi | Pinky | Rajy | Shiva | Raj | Rinky | Ramesh | Adam |
|---|---|---|---|---|---|---|---|---|---|---|
| marks | 65 | 54 | 96 | 94 | 93 | 85 | 84 | 39 | 44 | 40 |

The box plot for the same is:




The values shown by the box plots are:

| | |
|---|---|
| minimum | 39 |
| I Quartile | 43 |
| mean | 69.4 |
| median | 74.5 |
| II quartile | 74.5 |
| III quartile | 93.25 |
| maximum | 96 |

Uses of a Box Plot.

→ They provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed it is, it is skewed or not.

→ The median gives the average value of the data.

→ They show if data is skewed or not
i) If median is at center - normal distribution

ii) If median lies closer to I Quartile: Positive skew
iii) If median lies closer to III Quartile: negative skew

Box Plots give an idea about the outliers which are points numerically distant from the rest of the data.

Conclusion:

Hence, we Plot a base plot for the distribution in the given dataset, successfully.

# Experiment - A 10

Title : Data visualization III

Aim: Download the Iris flower dataset or any other dataset into a Dataframe. Scan the dataset and give the inference as

1. How many features are there and what are their types?
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a boxplot for each feature in the dataset Compare distributions and identify outliers.

Theory: Data visualization is actually a set of data points and information that are represented graphically to make it easy and quick for any user to understand.
It is good if it has a clear meaning, purpose and is very easy to interpret, without requiring context.

Tools of data visualization provide an accessible way to see and understand trends, outliers and patterns in data by using virtual effects or elements such as chart, graphs and maps.

Characteristics of effective graphical visualizations
a) It shows data clearly in an understandable manner.

b) It encourages viewers to compare different descriptions of data set.

c) It closely integrates statistical and verbal descriptions of data set.

d) It helps in identifying areas that need more attention and improvement.

Types of data / features:

1) Numeric data: It consists of all types of numbers. Such features can be measured by any numerical variables. eg. Age, scores, etc

2) Qualitative data: It consists of words, pictures and symbols, basically non-numeric features eg. colours

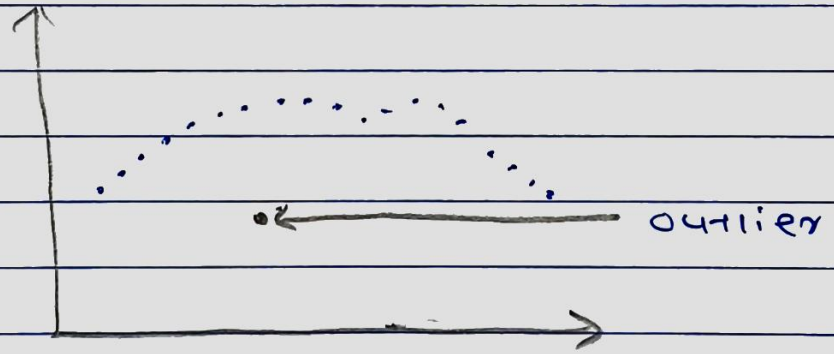
3) Nominal data: It is used for labelling variables without any type of quantitative value. It is just a name of feature without applying any order eg. Gender, ethnicity, etc

4) ordinal-data - This shows where a numbers is in order It is the data which is placed into some kind of order by ~~which~~ their position on a scale
eg. letter grades: (A, B, C); Rank ($1^{st}$, $2^{nd}$, $3^{rd}$...)

Outlier:

It is a data field that deviates significantly from the rest of the data objects and behaves in a different manner. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.

An outlier cannot be termed as noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.




Conclusion: The histograms and box plots for each feature has been plotted and compared

Group B
Assignment - 1

Sahil Gadge
19C0021
TE Comp 1
DSBDAL

Aim: write a Code in JAVA for a simple WordCount application that counts the number of occurances of each word in a given input set using the Hadoop MapReduce framework on local standalone setup.

Algorithm:

1. Start
2. Take the input data or file to be processed
3. Splits the incoming data into the smaller pieces (Split)
4. maps data by performing action like filtering, grouping and sorting
5. Combines the data to improve the performance by reducing amount of data transfered across network
6. shuffles and sorts the output from all the mappers
7. Reduce aggregates the output of mappers using reduce() function
8. End

Steps:

Step 1: Create a text file in your local machine and write some text into it.

$ nano data.txt

Step 2: check the text written in the data.txt file

$ Cat data.txt

Step 3: Create a directory in HDFS, where to kept text file.

$ hdfs dfs -mkdir /test

Step 4: Upload the data.txt file on HDFS in the specific directory.

$ hdfs dfs -put /home/data.txt/ test.

Step 5: write the MapReduce Program using eclipse

Result:

The word count application is implemented in Hadoop MapReduce

Group B

Assignment - 2

Sahil Gadge
19C0021
TE Comp 1
DSBDAL

Aim: Design a distributed application using Map Reduce which process a log file of a system.

Algorithm:

1. Start
2. Flume collects streaming log data into HDFS from various HTTP sources and application servers
3. Data stored in HDFS which was collected by Flume.
4. Pig parses these log files into the structured format
5. Schema will defined by hive to the structured data and schema will be stored in hive metastore
6. It picks any visualization supporting hive connectivity among Tableu and Hunk
7. End

Steps:

Step 1: Flume collects streaming log data into HDFS from various HTTP sources and application servers.

Step 2: The data collected by flume is stored into the HDFS

Step 3: Pig Parses these log files into structured format through various UDF's

Step 4: Hive or Hcatalog will define schema to this structured data and schema will be stored in hive metastore.

Step 5: It picks any visualization supporting hive connectivity among Tableu and Hunk.

Result: We have designed distributed application using MapReduce which process a log file of a system and resultant output is verified

Group B
Assignment-3

Sahil Fadge
19C0021
TE Comp I
DSBDAL

Aim: Locate dataset (eg. sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.

Algorithm:

1. Start
2. It exceeds in 3 stages
   2.1 Map stage process the input data which is in array or file format
   2.2 Shuffle stage removes duplicate values and the data will be sorted or shuffle.
   2.3 Reduce stage aggregates the values come from previous stages against the corresponding keys
3. End

Steps: Step 1: Download the dataset for various cities in different years.

Step 2: Make a project in eclipse and in that project create Java class and put code in it.

Step 3: Start the hadoop daemons
start-dfs.sh

Step 4: move the dataset to the hadoop HDFS.

Step 5: Run the Jar file with below command
hadoop jar /jar file_location/dataset-location in-HDFS /output_file-name

Step 6: Move to localhost: 50070/, under Utilities select Browse file system and download it.

Step 7: See the result in downloaded file.

Result: Located the dataset working on weather data which reads text input files and finds the average for temperature, dew point and wind speed.