SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

INTERNSHIP PROJECT

SEM/YEAR – 8th /4th

SYNOPSIS

PROJECT NAME: - WhatsApp Chat Analyzer

TEAM MEMBERS

[redacted]

ABSTRACT

The most used and efficient method of communication in recent times is an application called WhatsApp. WhatsApp chats consist of various kinds of conversation held among two people or a group of people. This chat consists of various topics. This information can provide a lot of data for the latest technologies such as Machine Learning. The most important things for Machine Learning models are to provide the right learning experience which is indirectly affected by the data we provide to the model. This tool aims to provide in depth analysis of the data which is provided by WhatsApp. Irrespective of whichever topic the conversation is based on, our developed code can be applied to obtain a better understanding of the data. The advantage of this tool is that it is implemented using simple python modules such as pandas, matplotlib, seaborn, streamlit, NumPy, re, emojis and a technique sentiment analysis which are used to create data frames and plot different graphs, where then it is displayed in the streamlit web application which is efficient and less resources consuming algorithms, therefore it can be easily applied to larger dataset.

1. INTRODUCTION

This tool is based on data analysis and processing. The first step in implementing a machine learning algorithm is to understand the right learning experience from which the model starts improving. Data pre-processing plays a major role when it comes to machine learning. In order to make the model more efficient we need lots of data, so we turned our focus primarily on one of the large-scale data producers owned by Facebook which is nothing but WhatsApp. WhatsApp claims that nearly 55 billion messages are sent each day. The average user spends 195 minutes per week on WhatsApp and is a member of plenty of groups. With this treasure house of data right under our very noses, it is imperative that we embark on a mission to gain insights on the messages which our phones are forced to bear witness to. A list that uses pie charts and diagrams to represent the interesting data that it collects after analyzing your WhatsApp chats. You know the drill by now. You will take a backup of your chat and send it to an email id listed on the site.

2. PROBLEM DOMAIN

WhatsApp-Analyzer is a statistical analysis tool for WhatsApp chats. Working on the chat files that can be exported from WhatsApp it generates various plots showing, for example, which other participant a user responds to the most. Communication between people using the internet becomes part of their daily life. People used to communicate with each other using the online chat system to transfer their messages. We propose to employ dataset manipulation techniques to have a better understanding of WhatsApp chat present in our phones. It shows most used emoji and word which repeatedly most times. It tracks our conversation and analyzes how much time we are spending.

3. SOLUTION DOMAIN

Data pre-processing, the initial part of the project is to understand implementation and usage of various python built -in modules. The above process helps us to understand why different modules are helpful rather than implementing those functions from scratch by the developer. These various modules provide better code representation and user understandability. The following libraries are used such as numpy, scipy pandas, csv, sklearn, matplotlib, sys, re, emoji, nltk seaborn etc.

Exploratory data analysis, first step in this to apply a sentiment analysis algorithm which provides positives negative and neutral part of the chat and is used to plot pie chart based on these parameters. To plot a line graph which shows author and message count of each date, to plot a line graph which shows author and message count of each author, Ordered graph of date vs message count, media sent by authors and their count, Display the message which is di not have authors, plot graph of hour vs message count.

4. SYSTEM DOMAIN

- **Software Requirements:** The connections of your software with other libraries:
    a. Streamlit
    b. Numpy
    c. Pandas
    d. Wordcloud
    e. Plost
    f. Pathlib
    g. Collection
    h. Matplotlib
    i. Windows 7 and above

- **Hardware Requirements:**
    1. Any web browser supported device.
    2. HARD DISK: 50 GB or above
    3. RAM: 4 GB or above

- **Supported device types**:
  The software is developed for Windows 32-bit/64-bit or android etc.

- **Nature of the data and control interactions between the software and the hardware:**
  Internet connection

Technology Used

Python

It is an interpreted, high-level general-purpose programming language. Created by Guido Van Rossum and first released in 1991. Its language constructs and objects-oriented approach aim to help programmer with clear, logical code for small and large-scale tools. Python is used for web development (server-side), software development, mathematics, it can be used alongside software to create workflows, it can connect to database systems, it can also read and modify files, it can be used to handle big data and perform complex mathematics and can be used for rapid prototyping, or for production-ready software development.

JSON

Java Script Object Notation is an open standard file format, and data interchange format, that uses humanreadable text to store and transmit data objects consisting of attribute-value pairs and array data types. It is very common data format, with diverse range of applications. Such

as serving as a replacement for xml in ajax systems. Json is a language-independent data format. It was derived from JavaScript, but many modern programming languages include code to generate and parse JSON-format data. The official Internet media type for Json is application/json. Json filenames use the extension (.json). When exchanging data between a browser and a server, the data can only be text. Json is text, and we can convert any JavaScript object into json and json to the server. We can also convert any json received from the server into JavaScript objects. This way we work with the data as JavaScript objects, with no complicated parsing and transactions.

NumPy: NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra and matrices. It is an open-source project and you can use it freely. NumPy stands for Numerical Python.

Matplotlib: For Python and its numerical extension NumPy, Matplotlib is a cross-platform data visualization and graphical charting package. As a result, it presents a strong open-source substitute for MATLAB. The APIs (Application Programming Interfaces) for matplotlib allows programmers to incorporate graphs into GUI applications.

Re: A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression (or if a given regular expression matches a particular string, which comes down to the same thing).

RegEx: A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. RegEx can be used to check if a string contains the specified search pattern.

Streamlit

Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.

5. FEASIBILITY STUDY

- **Technical Feasibility**

The technical feasibility study reports whether there exists correct required resources and technologies which will be used for project development. It is the measure of the specific technical solution and the availability of the technical resources and expertise. In our project we will be using Jupyter notebook(web based application) and VS code(text editor), both of them are open source software's. Along with these various python libraries and will be used. Jupyter is non-profit organization created to develop open-source software, open standards, and services for interactive computing across dozens of programming languages. The idea is to implement a data processing code using python to make better sense of WhatsApp group chat data.

- **Economical Feasibility**

Cost and benefit of the project is analyzed in economic feasibility, that means what will be the cost of final development of the product. This project has no cost in development since all the software and technologies used are open source. This project is not economical as it mainly depends on the analysis of data between two/more devices(phones)

- **Operational Feasibility**

It is to determine whether the system will be used after the development and implementation. In Operational Feasibility degree of providing service to requirements is analyzed .This involves the study of utilization and performance of the product. Our project shows the whole analysis of the chats among people. It can be two people or a group of people and provides various information using charts in easily readable format.

6. CONCLUSION

In conclusion, it can be said that the capabilities of the WhatsApp application and the power of the python programming language in implementing whatever data analysis intended, cannot be overemphasized. This work was to discuss the WhatsApp application and python libraries, to create an analysis of a WhatsApp chat .We propose to employ dataset manipulation techniques to have a better understanding of WhatsApp chat present in our phones.It shows most used emoji and word which repeatedly most times. It tracks our conversation and analyzes how much time we are spending.The system was done with python, and the python libraries that were implemented includes, NumPy, Pandas, Matplotlib and Seaborn. At the end of the work expected results were obtained and the analysis was able to show the level of participation of the various individuals on the given WhatsApp group chat . On serious note this system has the ability to analyze any WhatsApp chat input into it.