**Text mining for identifying the importance of the research article using NLP and Machine Learning**

**Prerequisites:**

* **Hardware requirements:**

1. Intel i3 processor or above.
2. 4GB RAM.
3. System type 64 bit OS.

* **Software requirements :**

1. Xampp(PHPMyAdmin server)
2. Jupyter Notebook
3. Python (2.7 or above)
4. PyQt
5. Highcharts

* **Links to download software required:**
  + 1. Anaconda : <https://www.anaconda.com/distribution/>
    2. Jupyter Notebook
    3. Xampp: <https://www.apachefriends.org/download.html>
    4. Highcharts: <https://www.highcharts.com/>
    5. Python: <https://www.python.org/downloads/>
    6. PyQt: <https://sourceforge.net/projects/pyqt/files/latest/download>
* **Python Libraries to install:**

1. import requests
2. from Bio import Entrez
3. import urllib
4. import json
5. import pymysql
6. import dateutil.parser
7. import datetime
8. import bs4 as bs
9. from bs4 import BeautifulSoup
10. import random
11. import re
12. import operator
13. import nltk
14. import numpy as np
15. from scipy.sparse.linalg import svds
16. from sklearn.feature\_extraction.text import TfidfVectorizer
17. import matplotlib.pyplot as plt

* **Database in MySql:**

Name of database: meta\_data\_info

Table No.1 :

Name : metadata

Schema :

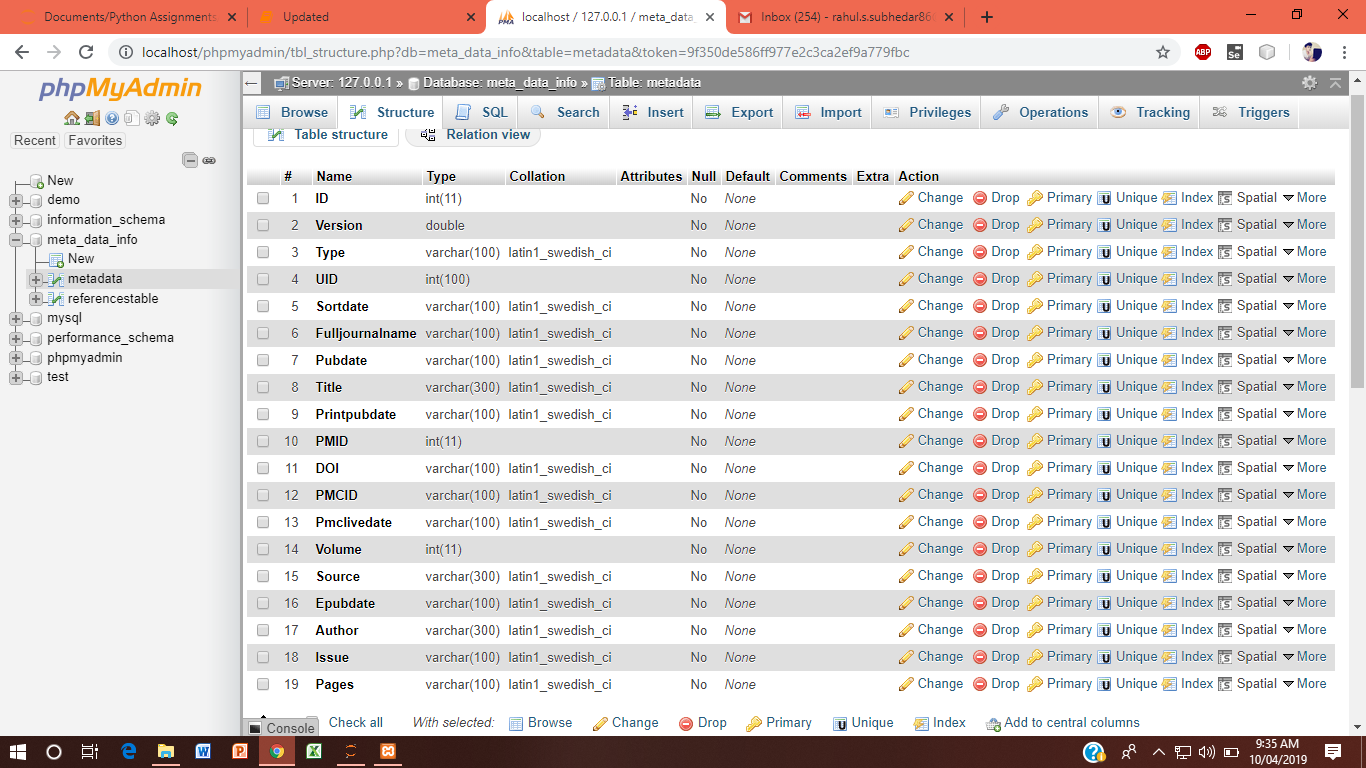
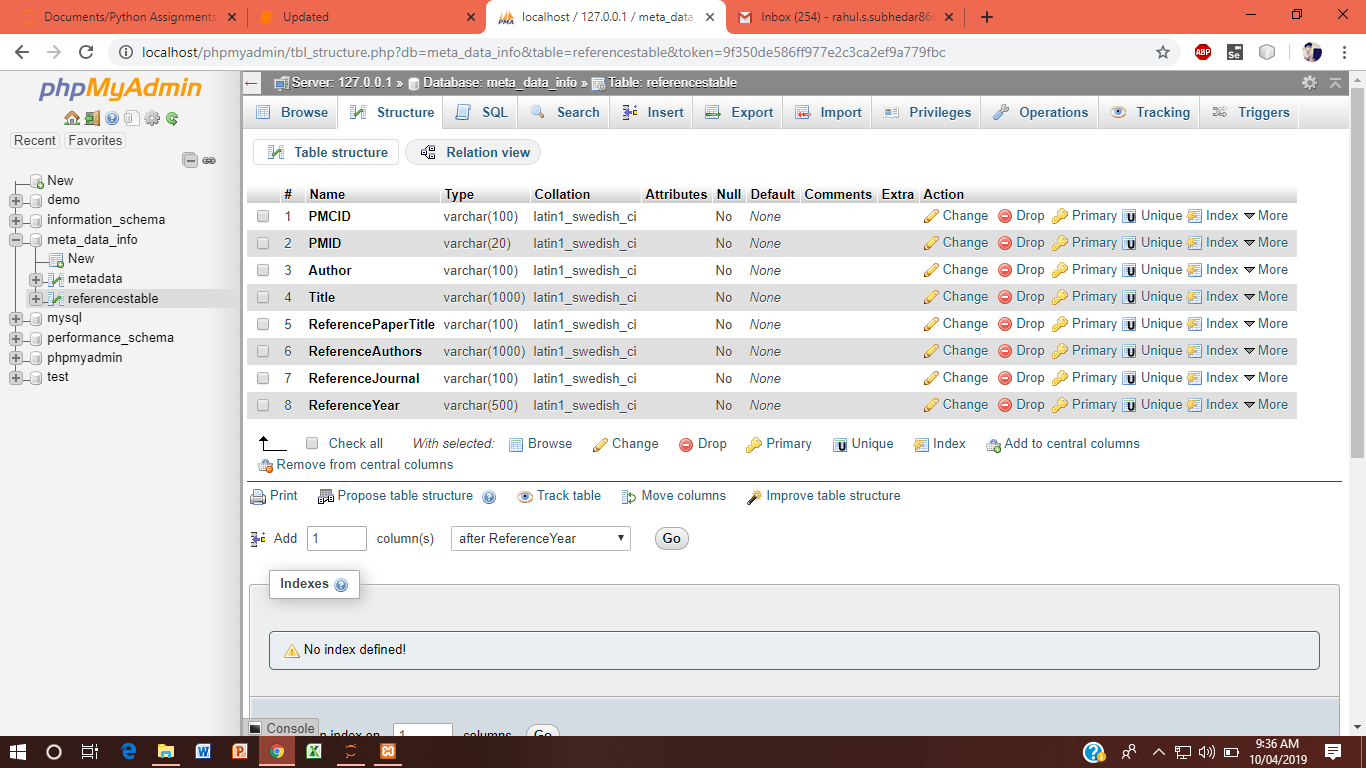
****

Table No.2

Name : references table

Schema:



**Work in Jupyter :**

Steps:

1. Open Jupyter Notebook
2. Go to directory in jupyter notebook where you want to save the project.
3. Create new Python file .
4. Paste the project (which is provided) in Python file .
5. Save and Run the project.

* **Use cases:**

**1.Extraction of Metadata:**

Metadata is data [information] that provides information about other data.

Steps:

1. Run the project which is provided through Jupyter Notebook.
2. Give the Organ name on which you want to research on.
3. This will provide the count of research article available on the PubMed site for the desired organ specified above.
4. The user will give count of articles (i.e 2,5,10,20,etc)wants to research on.
5. This will provide the Metadata ( i.e pmcid,Author,Title and all other available fields)
6. Through the provided metadata the further Summarization and Visualization of data is done .

**2.Importance of Research Articles :**

The Authors who have published most articles those authors articles are considered as important articles for desired organ specified in metadata extraction.

**3.Data Visualization:**

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

Steps:

1. The important research articles metadata which is extracted is stored in the database.
2. Through those entries in database the Data Visualization (graph) is created using Highcharts (software for creating visualization on UI)
3. Which will give user the idea about which articles are important (based on the authors who have published most articles).
4. So, further for user it will be easy to understand and get the summarized data.

**4. Summarization :**

1.  Through data visualization we get important article based on that we are doing summarization of the article.

2.  There are two ways to get the summary of  article

* + - Based on user input keywords
    - Using concept of  bag of words

3.  Summarization Based on user Keywords:

1.  User gives the input as a keyword which has been important for his research.

2.  In the output he gets the sentences summary in which that keyword  is occurs.

4.  Summarization Using concept of  bag of word: The bag-of-words model is simple to understand and implement. It is a way of extracting features from the text for use in machine learning algorithms.