AIT 580-FINAL PROJECT

by

Suhasini Konimeti Naresh Kumar

G01224512

(0)Title of the Project:

Data Analysis on Healthcare dataset.

(1)Introduction:

The project is primarily based on the prediction of category of the patient based on the corresponding measure using Natural Language Processing (NLP). In the given dataset the measure is the types of issue that the patient has based on their age and the disease they are affected with or their current check up routine. Category describes the type of medication that the associated person comes under, it includes health outcomes, prevention and unhealthy behaviors. This project helps in detecting the category once the measure is given so that the medical case can be redirected to the organization associated with the category recognized so that the public is given timely response.

(2)Nature of the Data Curation:

The CDC and the Robert Wood Johnson Foundation are partnering together on the 500 Cities project for better healthcare. The aim of the 500 Cities Project is to provide small area estimates for city-and census tract-level chronic disease risk factors, health outcomes, and medical preventive care use for the United States' largest 500 cities. Such small estimates of the region would allow cities to do so. The purpose of the 500 Cities Project is to provide city- and census tract-level small area estimates for chronic disease risk factors, health outcomes, and clinical preventive service use for the largest 500 cities in the United States. These allow the local health departments to understand the burden and geographic distribution of health-related variables in their jurisdictions and assist them in planning public health interventions in an efficient way.[1]

The Centers for Disease Control and Prevention (CDC) collaborates to form the experience, data, and tools that folks and communities ought to shield their health through health promotion, prevention of unwellness, injury, incapacity, preparedness and prevention for new health threats.[2]

The Robert Wood Johnson Foundation(RWJF) is the nation's largest philanthropy which is devoted to health. The RWJF is functioning aboard others to make a national culture of health. This foundation places the well-being at the top of every aspect of life.[3]

The data was collected for the betterment of the public by the government, so it is not biased. One of the pros of the dataset is that it is vast. Since it has many entries, it is efficient in attaining the highest accuracy possible for the prediction. The main con of the dataset is that it has many unwanted data variables which are not used in the prediction process, these unused variables were removed while preprocessing the data set. The data set also contained missing value in its variable entry named CityName. These null entries were also removed in order to prevent their intervention in accuracy of the prediction model.

(3)Questions:

The questions of interest are as follows:

- Is it possible to predict the category of the patient using the measure?
- What is the emerging health problem ?

Hypothesis:

- The health issues that occurred in 2016 are related to the health issues that occurred in 2017.

Considering measure to be the primary input, once it is provided the category is predicted. When the category is predicted the organization should also alert the professionals working under the category with the patient's record. This also helps in redirecting the patient to that category so that the patients are provided with efficient and timely treatment.

Under this case both the organization providing the medication and the public seeking the medication guidance are benefitted. This aids the organization in providing targeted medication in order to help the public have an efficient medication for the betterment in health and also analyze which is the disease that is emerging, so that they can make the public take protective steps to prevent the disease. It also benefits the public with better health.

(4)Requirements and Resources needed:

The software that were used in this project are Anaconda, Jupyter notebook, Python, numpy, panda, tensor and plotly.

Anaconda Navigator is a desktop graphical user interface (GUI) enclosed in Anaconda distribution that permits to launch applications and simply manage conda packages, environments, and channels while not exploiting the command-line commands. It is accessible for Windows, macOS, and Linux. Data scientists usually use multiple versions of the packages and use multiple environments to separate these totally different versions. The command-line program conda is a package manager that is associated with a setting manager. This helps data scientists make sure that every version of every package has all the dependencies it needs and works properly.[4]

Jupyter Notebook is an interactive computational environment, within which code execution, rich text, mathematical and statistical functions , plots and media can be combined.[5]

Python is a high-level, general artificial intelligence language. Python permits programming in Object-Oriented and Procedural areas. Python programs are usually smaller than alternative programming languages like Java. Python language is being employed by most tech-giant corporations like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc. Strength of the Python is the large library which might be used for the subsequent Machine Learning, interface Applications (like Kivy, Tkinter, PyQt etc.), internet frameworks like Django(used by YouTube, Instagram, Dropbox), Image process (like OpenCV, Pillow), internet scraping (like Scrapy, BeautifulSoup, Selenium), check frameworks, Multimedia, Scientific computing, Text processing etc.[6]

NumPy is a basic package for scientific computing with Python. It contains a strong N-dimensional array object, refined (broadcasting) functions and tools for desegregation C/C++ and algebraic language code. It's helpful in working algebra, Fourier rework, and random range capabilities. Capricious datatypes permits NumPy to integrate with a large variety of databases.[7]

Pandas could be a Python package providing quick, flexible, and communicatory information structures designed to operate with structured and unstructured information in a simple and intuitive way. It is a high-level building block for doing sensible data analysis in Python. Pandas are turning into the foremost powerful and versatile data analysis / manipulation tool accessible in any language.[8]

TensorFlow is an open supply software system library for top performance numerical computation. Its versatile design permits computation across a spread of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices. Originally developed by researchers and engineers from the Google Brain team from Google’s AI organization, it comes with sturdy support for machine learning and deep learning and is employed across several different scientific domains.[9]

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.[3] SciPy makes use of Matplotlib. [10]

Hardware resources that were used includes dell laptop with a 64-bit operating system, 16.0GB RAM and i7 processor.

The pre-processes that were needed to make the data suitable for usage were removing the unwanted variables. The variables that were used for the project were Year, StateAbbr, StateDesc, CityName, GeographicLevel, Category, UniqueID, Measure, PopulationCount, GeoLocation, CategoryID, MeasureId and Short_Question_Text. Again, from the newly created data frame with the above specified variable, the null values were removed. The null values might intervene in the accuracy of the prediction hence these values are removed. Since the dataset used is a real-world data set there is no issue in the originality of the data. The dataset is vast with a greater number of observations which makes the data efficient for prediction. If the dataset had a smaller number of entries that would result in any contingency in the training data set.

(5)Descriptive analysis:

The dataset is a complete dataset for the five hundred Cities project that was released in 2018. This dataset includes 2016, 2017 model-based small space estimates for twenty seven measures of chronic unwellness out of that 5 is because of unhealthy behaviors, 13 with health outcomes , and 9 cases with the utilization of preventive services. The information was provided by the Centers for Disease Control and Prevention (CDC), Division of Population Health, medical specialty and surveillence Branch. The project was funded by the Robert Wood Johnson Foundation (RWJF) in conjunction with the Centers for Disease Control and Prevention Foundation. It is a first-of-its kind effort to disperse information on an outsized scale for cities and for tiny areas inside those cities. It includes estimates for the five hundred largest American country cities and around twenty-eight thousand census tracts inside these cities. These estimates may be accustomed in determining the rising health issues and to develop and implement effective, targeted public healthcare and disease prevention activities. As a result, the tiny space model cannot discover effects because of native interventions, users are cautioned against victimization . Information sources generate these measures that embodies Behavioral Risk Factor Surveillance System (BRFSS) data (2016, 2015), Census Bureau 2010 census population data, and American Community Survey (ACS) 2012-2016, 2011-2015 estimates. Because some questions are only asked every other year in the BRFSS, there are 4 measures (high blood pressure, taking high blood pressure medication, high cholesterol, cholesterol screening) from the 2015 BRFSS that are the same in the 2018 release as the previous 2017 release.[11]

The attributed used in the revised dataset were Year, StateAbbr, StateDesc, CityName, Category, UniqueID, Measure, PopulationCount, GeoLocation, CategoryID, MeasureID and Short_Question_text.

Out of all the attributes used Year and PopulationCount are of the data type integer and the other columns comes under object since python is an object-oriented programming language.

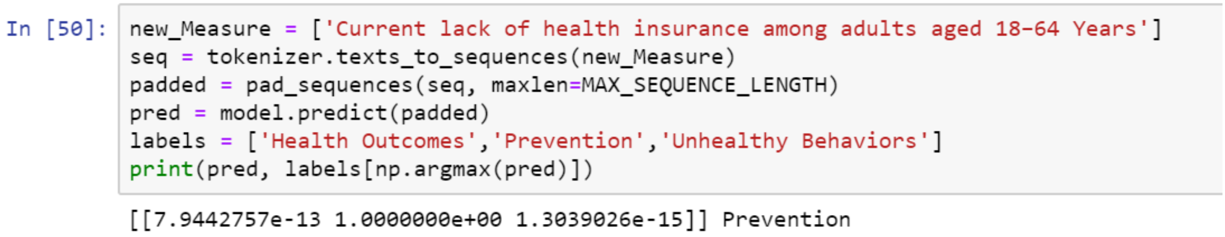
Recurrent Neural Network(RNN) is a sort of Neural Network wherever the output from previous is fed as input to this step. In ancient neural networks, all the inputs and outputs are freelance of every different, however in cases like once it's needed to predict succeeding word of a sentence, the previous words are needed and hence there is a necessity to recollect the previous words. Therefore RNN came into existence, that solved this issue with the assistance of a Hidden Layer. The most vital feature of RNN is Hidden state, that remembers some information of a couple of sequence. RNN has a "memory" that remembers all information concerning what has been calculated. It uses constant parameters for every input because it performs constant task on all the inputs or hidden layers to provide the output. This reduces the complexness of parameters, not like different neural networks.[12]

Natural Language processing , typically shortened as NLP, is a branch of Artificial Intelligence that deals with the interaction between computers and humans. The ultimate objective of NLP is to browse, decipher and understand the human languages in a manner that's valuable. Most NLP techniques have faith in machine learning to derive information from human languages.[13]

(6)Results:

Question 1:

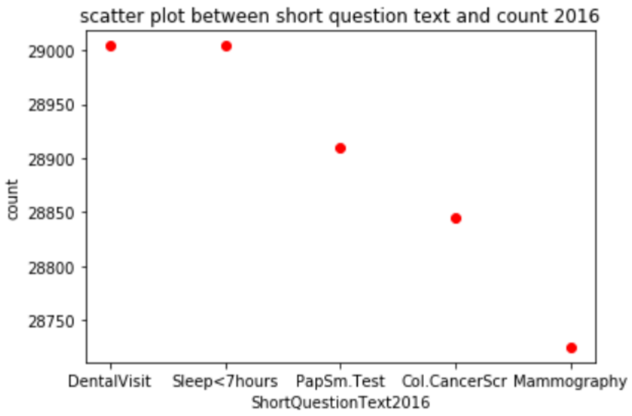
Is it possible to predict the category of the patient using the measure?



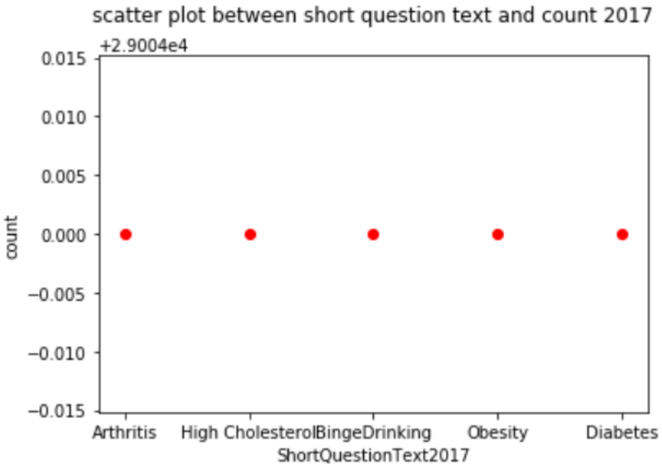
```
In [50]: new_Measure = ['Current lack of health insurance among adults aged 18-64 Years']
         seq = tokenizer.texts_to_sequences(new_Measure)
         padded = pad_sequences(seq, maxlen=MAX_SEQUENCE_LENGTH)
         pred = model.predict(padded)
         labels = ['Health Outcomes','Prevention','Unhealthy Behaviors']
         print(pred, labels[np.argmax(pred)])

         [[7.9442757e-13 1.0000000e+00 1.3039026e-15]] Prevention
```

The above picture is the prediction model result using Sequential TensorFlow in Natural language processing. This aids in detecting the category of the organization that should deal with the case in order to provide targeted healthcare.

Question 2: What is the emerging health problem ?



The diagram above shows the scatter plot that depicts the emerging disease in the year 2016. The health care issues that were prominent in 2016 are Dental visit, Sleep < 7hours, Pap Smear Test, Colorectal cancer screening and Mammography.



The above diagram shows the emerging disease in the year 2017.The emerging health problems of 2017 are Arthritis, High cholesterol, Binge Drinking, Obesity and Diabetes.

Hypothesis:

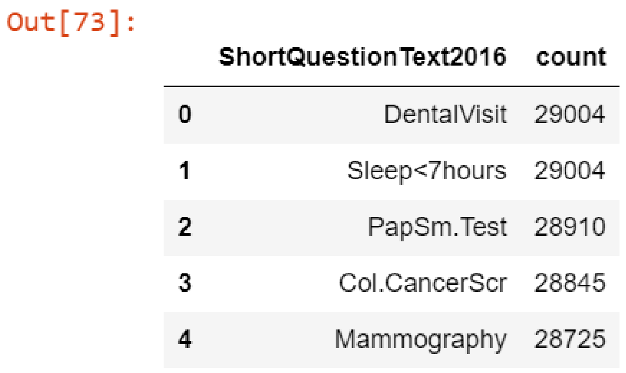
The health issues that occurred in 2016 are related to the health issues that occurred in 2017.

Hypothesis Outcome:

The health issues that occurred in 2016 are not related to the health issues that occurred in 2017.

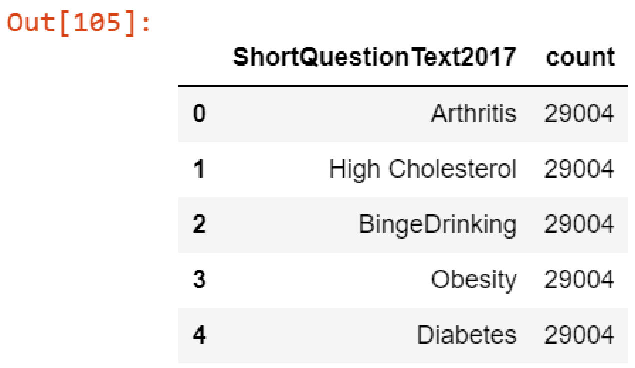
Justification:

From the two pictures it can be concluded that the health issues in 2016 is not related to the health issues in 2017. This is because in 2017 the diseases that appeared did not prevail in 2016 and even the screening and testing for the diseases were not conducted.



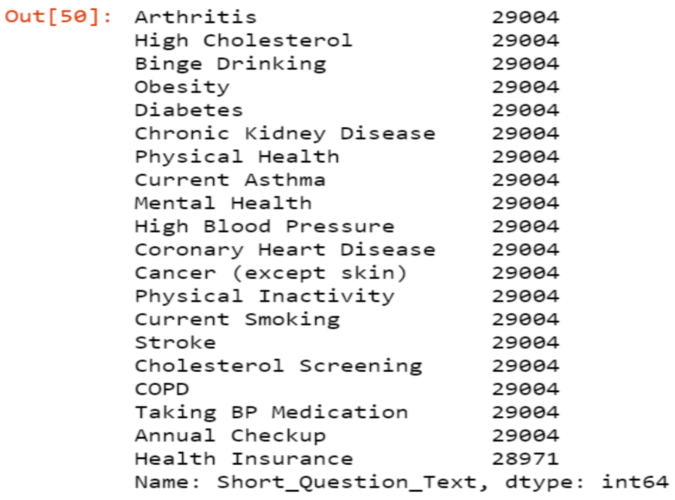
Out[73]:

	ShortQuestionText2016	count
0	DentalVisit	29004
1	Sleep<7hours	29004
2	PapSm.Test	28910
3	Col.CancerScr	28845
4	Mammography	28725



Out[105]:

	ShortQuestionText2017	count
0	Arthritis	29004
1	High Cholesterol	29004
2	BingeDrinking	29004
3	Obesity	29004
4	Diabetes	29004



```
Out[50]: Arthritis                     29004
         High Cholesterol              29004
         Binge Drinking                29004
         Obesity                       29004
         Diabetes                      29004
         Chronic Kidney Disease        29004
         Physical Health               29004
         Current Asthma                29004
         Mental Health                 29004
         High Blood Pressure           29004
         Coronary Heart Disease        29004
         Cancer (except skin)          29004
         Physical Inactivity           29004
         Current Smoking               29004
         Stroke                        29004
         Cholesterol Screening         29004
         COPD                          29004
         Taking BP Medication          29004
         Annual Checkup                29004
         Health Insurance              28971
         Name: Short_Question_Text, dtype: int64
```

The above image gives the overall view of the emerging health issues in the year 2017. All the issues had the same number of counts in case of 2017 except for health insurance.

(7)Conclusion and Future Work: Therefore, the classification model for the prediction of category with an accuracy of 0.86 by using measure as an input is built successfully by using Natural Process Language(NLP) with sequential TensorFlow.

This helps in easy access of targeted and improvised medication to the public. The emerging health issues were also found so that protective measures can be taken in order to prevent the public becoming a victim.

Future Work could be made on designing an app for the patients to keep track of their appointment. Their vitals can also be updated so that if there is any disturbance in their body can be known and actions can be taken in an early stage in order to avoid fatal diseases.

(8)References and Citations:

[1] "500 Cities Project: Local Data for Better Health." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, December 5, 2019. https://www.cdc.gov/500cities/index.htm.

[2] "Centers for Disease Control and Prevention: USAGov." A. Accessed December 16, 2019. https://www.usa.gov/federal-agencies/centers-for-disease-control-and-prevention.

[3] "About RWJF." RWJF, November 5, 2019. https://www.rwjf.org/en/about-rwjf.html.

[4] "Anaconda Navigator¶." Anaconda Navigator - Anaconda documentation. Accessed December 16, 2019. https://docs.anaconda.com/anaconda/navigator/.

[5] "The Jupyter Notebook¶." IPython. Accessed December 16, 2019. https://ipython.org/notebook.html.

[6] "Python Programming Language." GeeksforGeeks. Accessed December 16, 2019. https://www.geeksforgeeks.org/python-programming-language/.

[7] "NumPy¶." NumPy. Accessed December 16, 2019. https://numpy.org/.

[8] "Pandas." PyPI. Accessed December 16, 2019. https://pypi.org/project/pandas/.

[9] "Tensorflow." PyPI. Accessed December 16, 2019. https://pypi.org/project/tensorflow/.

[10] "Matplotlib." Wikipedia. Wikimedia Foundation, November 21, 2019. https://en.m.wikipedia.org/wiki/Matplotlib.

[11] "500 Cities: Local Data for Better Health, 2019 Release." Data.gov. Publisher Centers for Disease Control and Prevention, December 15, 2019. https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-fc759.

[12] aishwarya.27Check out this Author's contributed articles., aishwarya.27, and Check out this Author's contributed articles. "Introduction to Recurrent Neural Network." GeeksforGeeks, October 3, 2018. https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/.

[13] "Paraphrasing Tool: Free Article Rewriter, to Rewrite Sentences." Duplichecker.com. Accessed December 16, 2019. https://www.duplichecker.com/article-rewriter.php.

(9)Terms Definition:

LSTM: LSTM stands for Long Short-Term Memory. It is an architecture of Recurrent Neural Network. It is used for deep learning.

Keras: Keras is a high level open-source library created for the fast experimentation with neural network.

NLTK: NLTK stands for Natural Language Toolkit. It is used for tokenization and word count. NLTK is used to understand human language with packages that a machine understands.