Internship Report

**Intern Details:**

|  |  |
| --- | --- |
| **Student Name** | **RAJANA VEDA PRADEEP nAIDU** |
| **Roll Number** | **CS15B054** |
| **Company of Internship** | **eclerx** |
| **Profile** | **Software Development** |
| **Internship start date** | **17-5-2018** |
| **Internship duration** | **2 months** |
| **Location** | **Mumbai** |
| **Designation** | **intern** |
| **Internship Manager** | **Mr. NITIN THAKKAR** |
| **Project Name** | **Text infection** |

**Project Overview:**

Name of the project: **TEXT infection**

Languages used: **r , Python**

Frameworks used: **tensorflow ,Keras**

Environments used: **R Studio ,Anaconda**

Tools/Services used: ***Wordnet , datamuse,Tesuras, big .txt taken from Norvig.com***

***Libraries used in R code:***

* rvest ,stringi ,stringr , stringdist, gplots , csv, xml2, regex
* system libraries like utils

***Libraries used in Python code****:*

* numpy ,scipy ,scikit-learn , tensorflow , keras

**Extent of Skills used**:

* Coding in R and in python ,
* Algorithms that I have learned in collage got used here and I learned more algorithms during the course of internship
* The knowledge of Machine Learning and Statistics helped me to learn more things that r related to Machine Learning , like exploring to TensorFlow and Keras
* The previous knowledge of R ,Python , R studio, PyCharm, Sublime helped me during the course of Internship

**Project Details:**

* Text Infection in R ,Word prediction using Python
* Part 1(Code in R) : a)Creating data for corpus file , sorting the words in descending order of frequency . b) Correction of word is based on edit to a word is a deletion (remove one letter) or a replacement (change one letter to another) or an insertion (add a letter) or a transposition (swap two adjacent letters). c) Based on the sorted words taken from the big corpus file , we apply this algorithm to correct the given word .
* Part 2(Code in R) : a)The above word correction is extended to line correction & then to multiple lines ,Hence able to text alignment of all the word in the text ,word by word . b)This thing is further extended to user defined words , user defined has higher priority than the sorted corpus file , while aligning word
* Part 3(Code in R) : a) web scraping in R , while grouping the given set of words into word class , to form the word classes , scraped the dynamic data in R .
* Part4(Code in Python): a) word2vec implementation using Python , using 2 different models with front end Keras & Back end Tensorflow (tf)

**Results :**

* I have done word infection which is further extended to text alignment.
* I had used corpus text file ,for English words and used costumer based input files in same code for improving the results of the test alignment
* Related to word2vec representation of words , I am able to represent word in the form of vector and able to predict the next word in the sentence using n–skip Gram method

**Learning of the student:**

***Technical:***

* Got good exposure to R language and R studios
* Learned Algorithms and statics part that r related to text analysis
* In the period of Internship got good understanding of Natural Language Processing(NLP)
* Got to know more about Keras and TensorFlow and CNKT. Implemented word2vec representation using (continuous bog of words) CBOW & n-skip grams models for predicting context & labels.

***AESTHETIC****:*

* Good modeling of code based on different views
* Writing code in structured way , with code optimizations , alignment and code commenting

**RATIONALE:**

**Software Development Life Cycle (SDLC) learnings**:

* I have learnt that preparing good design is important before coding
* Got to know about corner cases and how to rectify them
* Got to know How to about different kinds of Algorithms and Open source API's that are used in Text Analysis

**Personal Learnings:**

* This Internship helped me to apply my knowledge in computer Science and got exposure to more things that are related to computer science and Engineering.
* I got to know about real time problems that can be occur while solving live problems.
* Along with structured, I had worked with unstructured data which is of customer based text.
* This internship helped me to gain confidence and increased my interest to develop more software tools.
* Had good talk with people who are in similar fields

**Non-revelation of confidential Information**

**T** The above report does not contain any confidential information from eClerx

Mr. NITIN THAKKAR (Project Manager)