SAMRIDDHI SINHA | 15CE10045

CIVIL ENGG. (B.Tech 4Y)

EDUCATION

Year	Degree/Exam	Institute	CGPA/Marks
2019	B.TECH	IIT Kharagpur	7.44 / 10
2014	All India Senior School Certificate Examination	Central Board of Secondary Examination	88%
2012	All India Secondary School Examination	Central Board of Secondary Examination	10 / 10

INTERNSHIPS AND PROJECTS

Google Summer of Code 2017

• Participating in the on-going **Google Summer of Code 2017**, under the mentorship of **Portland State University**.
• Project involves the Creation of an Natural Language Processing Toolkit for Indian Languages with functionalities for Hindi and Bengali to be implemented over summer.
• Each language have the following functionalities attached to them **tokenizer**, **lemmatizer**, **pos-tagger** and **gender tagger** (if the language had gender rules). **NLTK wasn't used** for creating any of the functionalities above.
• **Phase 1:** Extraction of morphological data from the **Hindi Dependency Treebank** corpus which came in **CoNLL** format. Extracted data from the **Hindi Wikimedia Dumps** using a self-built regex parser. This aided in extendng the size of vocabulary immensely thereby improving **word vector accuracy.** Implemented a **word, sentence and a simple (delimiter based) tokenizer.**
•**Phase 2:** Creation of a **lemmatizer** from a dictionary built from **HDTB**. The lemmatizer implemented a simple Dictionary lookup method backed off by various other functions based on regex and language rules to improve accuracy. For **gender tagging** in Hindi apart from dictionary lookup, **gensim's word2Vec** implemention was used to create **word embeddings. A one-hot encoding** was used to represent gender and the **Multi Layer Perceptron Classifier** was trained to classify gender.
•**Phase 3:** The plans for the POS tagger involves a combination of an **N-gram** and a **rule-based** tagger for Hindi, For other languages a sequence tagger based on **Long Short Term Memory Recurrent Neural Networks** would be implemented. Currently working on the same.

COMPETITION/CONFERENCE

Secured Silver Medal in Data Analytics, Technology General Championships 2017

Competitive Strength Prediction of ATM Vendors in California:
•Part of a team of 15 members to analyse the competitive strength, of 3 major ATM vendors, from the demographic data of California, US.
•Personally was responsible for scraping data from http://www.unitedstateszipcodes.org/ for the demographic data on PIN Codes.
•**Visualised feature** importance using **Tableau** and clustered the ATM locations by utilising **k-means** approach.
•Combined per county demographic model with the whole state demographic model to calculate the final annual revenue generation of each ATM location.

WORK EXPERIENCES

Contributions to Open Source

• Helped improve the Bengali corpus under the **Classical Language Toolkit**. Built scrapers to scrape data off **Bengali WikiSource**.
• Created a function for **Lexical Dispersion plot** for **Classical Language Toolkit** and also updated their Documentation.
• Improved the **Indian Language Corpora** under **NLTK**.
• Currently working under **PSU** for **GSoC**, for the development of a **Natural Language Toolkit for Indian Languages**.
• Created a package that retrieves from **GBIF** 5000 **georeferenced records** of **Australian mammals**, and then sends all of them successfully to the **Geospatial Quality API.** This was part of a series of tests for **ropensci** under the **R Project.**

SKILLS AND EXPERTISE

FIELDS OF INTERESTS:
• Machine Learning • Data Analysis • Natural Language Processing
• Recommender Systems • Web Scraping • Data Mining
• Web Development • Database Management

LANGUAGES:
• **Python**: Worked for more than one year. Proficient with popular libraries like NumPy, SciPy, Pandas, Scikit-Learn, Matplotlib, Seaborn and NLTK.
• **R**: Beginner. Worked with R Project for a short time and ceated a package.
• C, C++ • HTML/CSS/Javascript

SOFTWARES, LIBRARIES AND IDEs:
• Jupiyter • Visual Studio • MATLAB

SYSTEMS:
• Linux/Unix, Windows

AWARDS AND ACHIEVEMENTS

JEE Advanced 2015

Ranked 2734 in JEE Advanced 2015

COURSEWORK INFORMATION

• Machine Learning, Andrew NG, Stanford, Coursera
• Machine Learning, Yaser Abu Mostafa, Caltech, EdX
• Neural Networks Class, Hugo Larochelle, Université de Sherbrooke
•Probability and Statistics •Programming and Data Structures

EXTRA CURRICULAR ACTIVITIES

I am an active writer on **Quora,** a knowledge sharing platform with **1.7 million views** and **49 thousand upvotes** on my content.