|  |
| --- |
| STAGE 1: DESIGN DOCUMENT |
| FX Problem - Design Document |
| Innovation Hack! |
|  |
| **TEAM:** One man Army :P |
| **AMBAREESH REVANUR**  ambareesh.r@gmail.com  (+91) 9916-084-938  @revanurambareesh/forexInsights |
|  |

**Overview**

This project aims to support forex business using data available on public web domain using NLP and web-scraping. A set of most frequently appearing *keywords* are used as feature set. The dataset is generated for each of the entry given in excel sheet (problem statement) using web search APIs and each link is scraped for data within specified HTML tags. The presence of the keyword in the data, is considered as *y* = 1 or positive case, after processing for its presence in majority of the links crawled by search engine. Once the dataset is ready, Naïve Bayes modeling is performed. This model is used to predict the probability to determine a **potential FX customer**. An intuitive GUI is provided for this project. Complete project is developed in Python 2.7. This multi-threaded tool provides complete solution to the forex problem.

Design Document is uploaded in git repository/Stage 1 Documents/FX Problem.pdf

**Table of Contents**

[Introduction 3](#_Toc470809961)

[Tools and Technology 3](#_Toc470809962)

[Feature Selection 3](#_Toc470809963)

[My solution: Algorithm 3](#_Toc470809964)

[Code analysis 3](#_Toc470809965)

[Other features 3](#_Toc470809966)

[Results 4](#_Toc470809967)

[Future scope 4](#_Toc470809968)

[Publication references 4](#_Toc470809969)

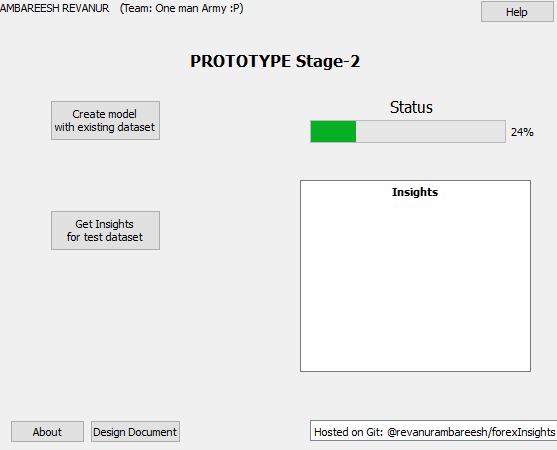


Figure : Stage 2 Prototype GUI

# Introduction

Machine Learning is science of predicting required parameter without have the computer program explicitly coded for so. This is possible by learning, what is called a ‘model’, for the collected data and, based on the nature and interrelationship of the data prediction can be made for an unknown parameter.

Machine Learning is gaining prominence due to large availability of data. Machine Learning’s philosophy centrally lies around the fact that more the genuine data available, better is the model learnt, and hence the accuracy of such a model is high.

Web technology has developed has seen boom in recent years with development of higher compute power and cheaper storage materials and tools available for the same. Lots of information is available on internet. According to recently released *voucherCloud* survey, 25,000GB of data is being uploaded per second to the internet. That is huge number. With this amount of data being available at finger-tips, it is possible to get useful information for just about anything.

In order to use the data already present on the web, two techniques are generally used, web-crawling and web-scraping. A web-crawler (sometimes referred to as web-spider) is a internet-bot that visits each allowed end-points systematically on the web for indexing the page. But web-scraper on the other hand, accesses data of the web for specific reasons including data mining, computer vision, machine learning and NLP.

In this project, data is scraped from across the web from various sources for forex and non-forex classes as per the given guidelines and is stored for learning a machine learning model. In this document following conventions are used frequently:

: Dataset created from web

: The label assigned with **1** as Forex-Class and **0** as Non-Forex Class

## Scope

This document gives a complete picture of the algorithm, code, folder structure and also some of the NLP research papers referred for this project. The software developed (herein referred to as “**forexTool**”) will scrape the data from internet, generate definitions, and creates a machine learning model. This model can be used to derive useful insights like:

* If the company is likely to opt for Forex or not
* How the model predicted

This tool is a prototype and offers solution to the stated problem statement.

## Overview

This document is organized into:

* Section 1: System overview
* Section 2: Software development tools
* Section 3: Machine learning feature selection
* Section 4: Solution approach
* Section 5: Relevant code snippets
  + Data generation
  + Insight generation
* Section 6: Tool features
* Section 7: Results & Future scope
* Section 8: References

# Chapter 1: System overview

There are 5 major components in the tool:

1. Web-Scraper
2. User-Interface (UI)
3. NLP based keyword analysis
4. Machine Learning model
5. Web-Search

Fig.2 shows how each of the above components interact:-

* Web scraping is performed based on the web-search results.
* The data obtained through web scraping is stored into file database (hardisk).
* Now to create a machine learning model,
  + Features: *keyword*s describing the word ‘Forex’ on few popular internet pages.
    - In this project about ~200 keywords have been extracted from internet pages. Each of these keywords a
  + Label :

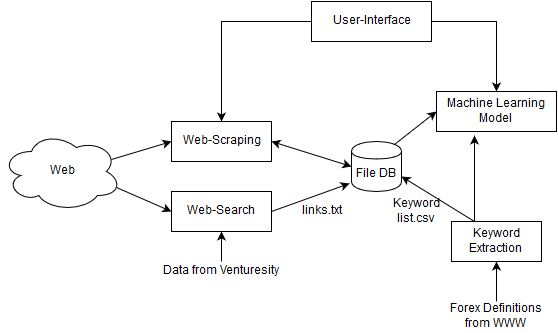


Figure : System architecture

## Keyword Extraction

# Tools and Technological Approach

## Python

### Google Custom Search API

Responds in JSON

### Scrapy

Limit is not imposed as DEPTH\_LIMIT = 0 by default

### RAKE (Rapid Automatic Keyword Extractor)

### Qt

### Scikit

### Multithreading

# Feature Selection

## Justification of feature selection

# My solution: Algorithm

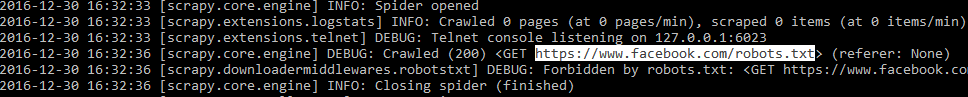
## Project architecture

# Code analysis

## Demonstration

# Other features

* Changing defining words may improve results.
* This project uses Scrapy 1.3.0 which respects *robots.txt*.
  + As defined in settings/settings.py => ROBOTSTXT\_OBEY = True



# Results

# Future scope

# Publication references

### About Me: AMBAREESH REVANUR

I am currently pursuing B.E. degree 3rd year, CSE at R. V. College of Engineering. I had obtained CET rank of 46 out of 0.1M competitors. I have closely worked with both industries and academia to solve real world, societal and business problems using technological solution.

At R&D labs of CSE Dept of RVCE, I developed a Textile image processing tool with C#.NET which is now used by about 100 weavers. Also, I have worked on challenging computer vision problems like Underwater Object Tracking using machine learning and published some of my work in conferences (using ML and CV).

At Siemens Healthcare Pvt Ltd, I have worked on an internal security tool called Security Vulnerability Monitoring (SVM) Tool. SVM automates vulnerability monitoring for many of the Syngo products (Core Siemens HC medical system) and provides a robust solution to protection of their product against vulnerabilities of the code.

In my free time, I compete and participate in Hackathons. I am an active member of IEEE Society. I have helped Team Chimera, RVCE, build web-interface and website.

Feel free to reach me at, ☺

 [@revanurambareesh](https://github.com/revanurambareesh)

 <https://in.linkedin.com/in/ambareeshr>