Twitterati Common Area of Interest Analysis

26.12.2018

R.Varadharajan
2016103604

S.Guru Prasath
2016103025

Overview

Phase I

This part of the project is aimed at identifying twitter public accounts by real time twitter API called “tweepy” in Python by randomly searching a topic and grouping the number of users tweeted on that particular topic and scraping all their tweets from scratch.This is the base idea for Scraping Real time Twitterati data from the twitter.

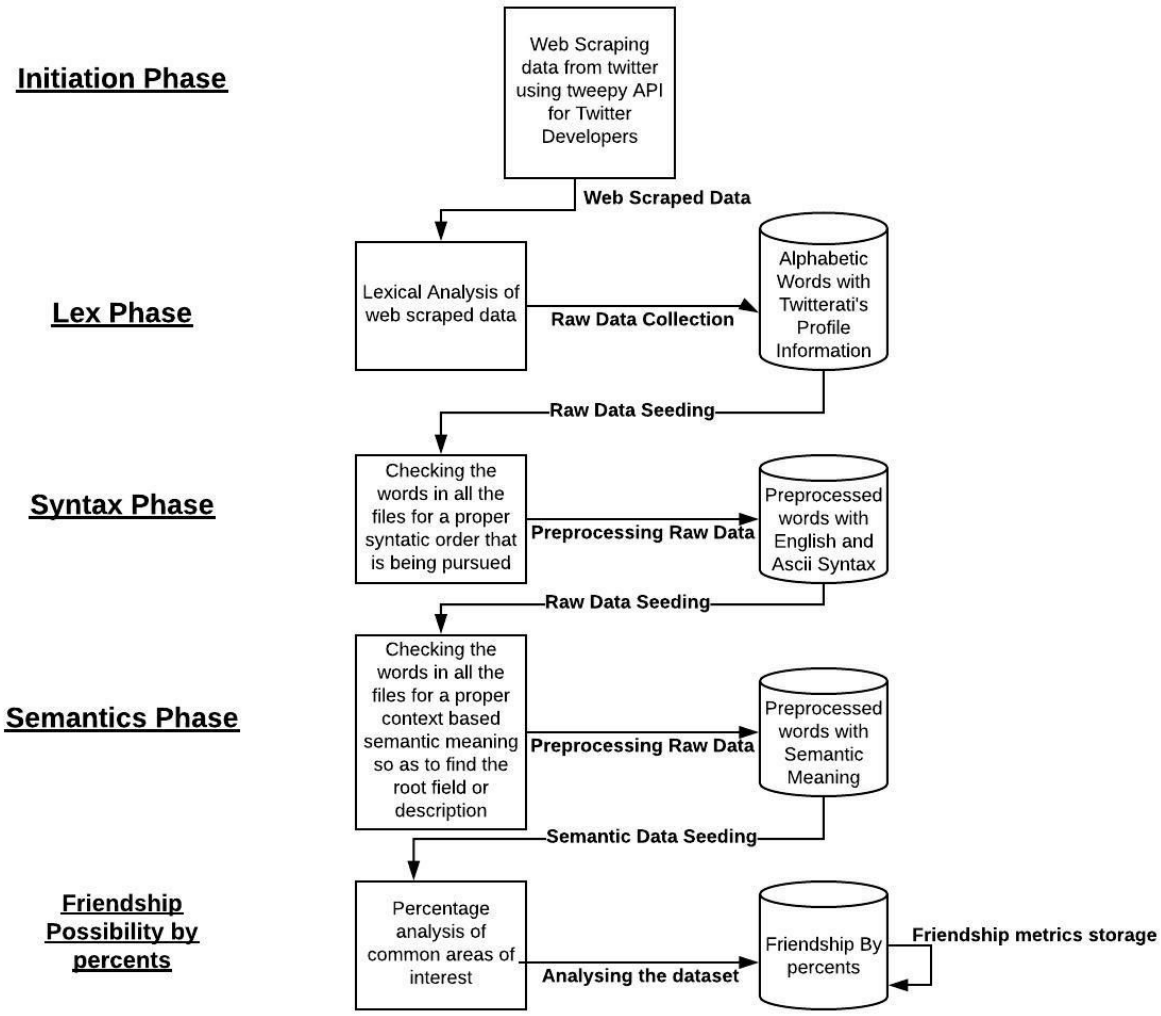
Phase II

The Second phase is tedious where in three major things are focussed.The three major concerns here are preprocessing the words in lexical,syntax and semantic analyses such that the identified words with proper satisfying conditions are then taken into concern for finding the common areas of interest for a particular user and then map it with all other users such that having common areas of interest could possibly strengthen the relationship between them and thereby helps identifying the suggestions list for a particular user by mapping the areas of interest patterns.

Goals

1. The scraped data is checked that whether it is an alphabet and all the other alphanumerals or non-ascii characters are removed as a part of pre processing.**(Lexical Analysis)**
2. The lexically preprocessed data are then checked whether they have a English Dictionary prescribed syntax i.e. finding whether the word is an english word or not.**(Syntax Analysis)**
3. The syntax checked english words are then identified their semantic meaning or context so that the area of interest of a user could be identified.All other waste words are removed while processing.**(Semantic Analysis)**

Diagrammatic Explanation :



Specifications

Input : Series of web scraped text files with Twitter usernames as filenames.

Processing: Lexical Analysis for filtering non-ascii characters and non alphabetic alphanumerals,Syntax Analysis for removing words that are non English words and finally Semantic Analysis to find the context and trace back the description to find the semantically correct area of interest.

Output: Suggesting a percentage of match between any two users by analysing their common area of interest and mapping it with each others.

Milestones

I. Efficiency

Efficiency is that what sits as a milestone in achieving such a idea but on improvising the idea with adaptive Machine Learning Techniques could find a better solution to strengthen the efficiency.

II. Social impact

But this idea as a base could possibly result in an unfavorability that may be identified as a result of analysis of data metrics who could possibly be any good friends.These ideas are just pacing off to ensure data metric analysis at better rates and such idea may also result in a bad manner.