PROGRAMMING ASSIGNMENT 4

Subject : Data Visualization: Tweet Miner
Advisor : Res. Assist. (Selim Yilmaz, Levent Karacan, Burcak Asal, Feyza Nur Cubukcuoglu

Due Date : 10.12.2015

Introduction

Twitter is a popular social media platform where users can share short messages called tweets restricted with 140 characters. Users share thoughts, links and pictures on Twitter and comment on recent events, companies advertise their products and engage with customers. Everyday, millions of tweets are shared by the users and this provides us lots of data to be analyzed.

In this assignment, you will implement a simple Python program which analyze a number of tweets to obtain various type of information and visualize it.

Overview

The goal of this assignment is to extract useful information from Twitter data and visualize it. We provide you the Tweeter data that is in *JSON* format, so we also give you a starter code to read tweets and various information related with these tweets. In brief, you will carry out the following steps:

- Compute term frequencies and plot a histogram for terms in top 20,
- Plot a time-frequency graph for terms in top 5,
- Build a co-occurrence matrix to show which terms are tweeted together and visualize it for terms in top 10,

Details

1. **Computing term frequencies and plot a histogram for terms in top 20**

   After you read tweets line by line using the provided code, you will tokenise text into words and count different words along the all tweets. You can use dictionary structure for this purpose if each different term is thought as key and term frequency as value. Finally you will plot a histogram as in Figure 1. Some situations you will consider are listed below:

   - You should convert tweets into the lowercase characters to avoid case sensitivity. For example Python and python are same terms.

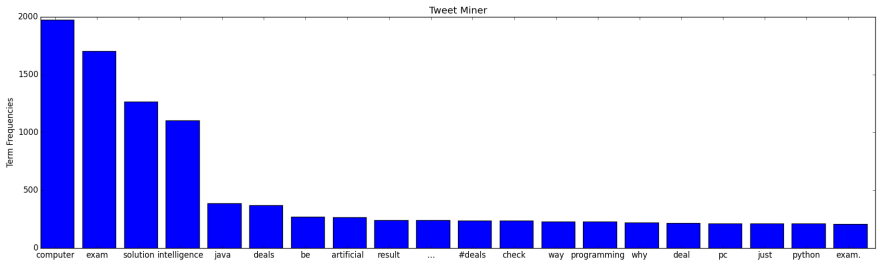

Figure 1: A sample histogram plot

- You do not have to deal with different forms of same word such as break-broken,#python-python.-python, get-got etc.
- You are given a Python list in the given starter code that include some English words which are meaningless when they are not used in combination with other words such as punctuation, articles(a, an, the), conjunctions (however, but, when, who, etc.), some adverbs. You should not include these words to general terms.
- You will print your result to a file named *term_frequencies.txt* as can be shown below.

  ('computer', 1975)
  ('exam', 1704)
  ('solution', 1266)
  ('intelligence', 1103)
  ('java', 386)
  ('deals', 369)
  ('be', 269)
  ('artificial', 267)

2. **Plotting a time-frequency plot for terms in top 5**

   After finding most frequently occurred terms, you will count the frequencies over time of most frequent 5 terms. The series will be re-sampled with intervals of 1 minute. Your figure should look like the one in Figure 2. In the given code, you are presented how date information is gathered and manipulated. Similarly you will print your result to a file named *term_frequencies_overtime.txt* as can be shown below. You will be also given example output files.

   Term Datetime Frequency
   .
   .
   .
   intelligence 2015-11-23 09:35:00 10
   intelligence 2015-11-23 09:36:00 12
   intelligence 2015-11-23 09:37:00 19
   intelligence 2015-11-23 09:38:00 16

intelligence 2015-11-23 09:39:00 15
intelligence 2015-11-23 09:40:00 10
intelligence 2015-11-23 09:41:00 19
intelligence 2015-11-23 09:42:00 32
java 2015-11-23 09:35:00 6
java 2015-11-23 09:36:00 7
java 2015-11-23 09:37:00 5
java 2015-11-23 09:38:00 2
java 2015-11-23 09:39:00 7
java 2015-11-23 09:40:00 7
java 2015-11-23 09:41:00 8
java 2015-11-23 09:42:00 5
.
.
.

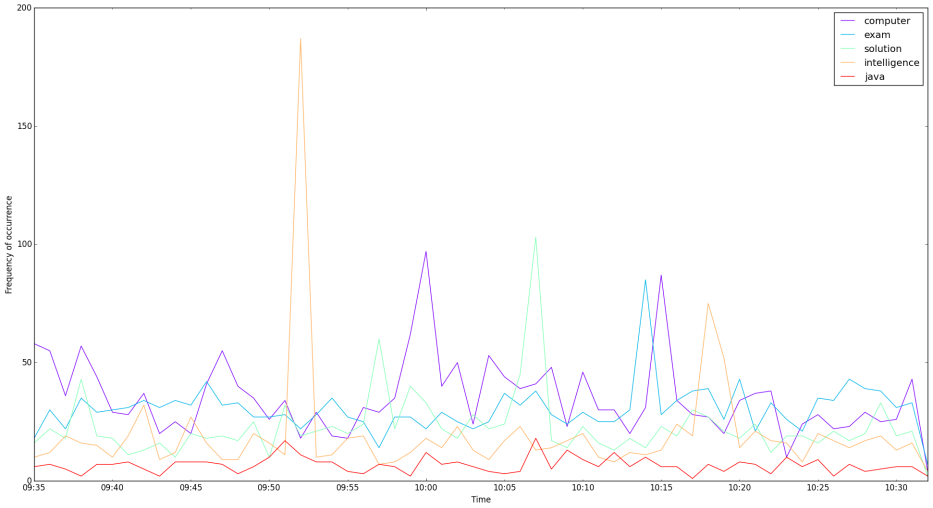

Figure 2: A sample plot showing term frequencies a period of time

3. **Building a co-occurrences matrix**

   Lastly, you will build a co-occurrences matrix between most frequent top 10 terms to show which ones are mostly occurred together in tweets. You will print result to *term_co-occurrences.txt* as in previous steps. You can see visualization of co-occurrences matrix in Figure 3.

   computer-exam 0
   computer-solution 1
   computer-intelligence 0
   computer-java 0

computer-deals 304
computer-be 44
computer-artificial 0
.
.
.

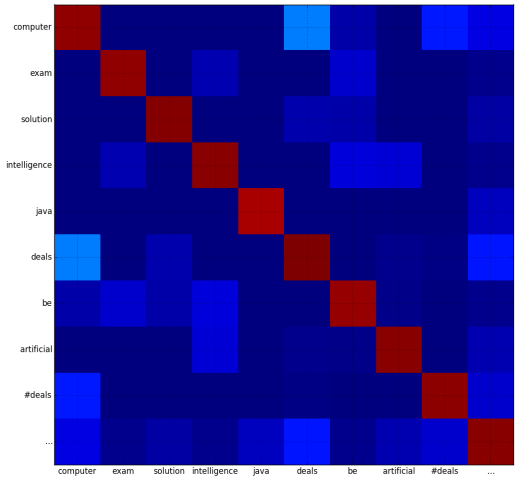

Figure 3: A sample co-occurrence matrix

Notes

- You will save plots as *.png* images. You are given a sample code to show how you can save plot figure as images.
- Do not miss the deadline.
- Save all your work until the assignment is graded.
- The assignment must be original, individual work. Duplicate or very similar assignments are both going to be considered as cheating.
- You can ask your questions via Piazza (`https://piazza.com/hacettepe.edu.tr/fall2015/bbm103`) and you are supposed to be aware of everything discussed in Piazza.
- The submissions whose upload score is 0 will not be considered for evaluation.
- You will submit your work from `https://submit.cs.hacettepe.edu.tr/index.php` with the file hierarchy as below:

This file hierarchy must be zipped before submitted (Not .rar, only .zip files are supported by the system)

→ <student id>
- → main.py
- → *.py
- → term_frequency.txt
- → term_frequency.png
- → term_frequency_overtime.txt
- → term_frequency_overtime.png
- → term_co-occurrences.txt
- → term_co-occurrences.png

This file hierarchy must be zipped before submitted (Not .rar , only .zip files are supported by the system).