IDS 521 Homework 2
Prof. Ali Tafti
Fall 2014
This version: November 6, 2014

Real-time Diffusion of Information on Twitter and the Financial Markets

Please read the attached working paper entitled "Real-time Diffusion of Information on Twitter and the Financial Markets," by Ali Tafti, Ryan Zotti, and Wolfgang Jank. One of the challenges of this research project is to compute 99th percentile threshold values of Twitter mentions or trading volume for each firm that change over time. This may be done with relative elegance and efficiency using SQL as a declarative language; and by taking advantage of the relational database's capabilities to perform joins or other computations efficiently.

In this homework assignment, you will extend the research described in the attached paper by considering the impact of earnings releases. You are given a SQL script that does the following:

1) Loads the main dataset in which each tuple is identified by a combination of the firm's symbol and ten-minute time increment during data collection hours. This set contains a total of 1,223,906 tuples, from May 21, 2012 through September 18, 2013.
2) Loads a file containing a comprehensive list of earnings releases; and then creates a new table in which earnings releases are matched to the appropriate firm and period-number combination of the main data set. Matches are based on the firm, date, and ten-minute time increment.

Building on the provided SQL script, you will write a complete SQL script that does the following (detailed in the steps below): A) Show the association between earnings releases and oncoming Tweet per-minute levels for the firm, in comparison with a baseline level computed based on the prior history of tweets for the same combination of firm, day of week, and half-hour of day (detailed in step 2 below). B) Compute the same types of calculations for trading volume (detailed in step 3 below). C) Identify Twitter peak events based on the prior history of Twitter messages mentioning the firm (detailed in steps 4 and 5 below).

IDS 521 Homework 2
Prof. Ali Tafti
Fall 2014
This version: November 6, 2014

Please submit the following:

1) A SQL script (a .sql file) that will load the source .csv files into tables and then do all the processing to create the final tables of results.
2) Any comma-separated value files used in your SQL scripts besides the files provided to you. For example, if you decide to create a Date-Hour database table, provide the .csv used to populate it and make sure its name matches the name used in your Load Data command.
3) An MS Word or PDF file in which you describe your final results. Please present any any small tables of results here. Clearly list the name of each database table which contains your results for each intermediate step. Briefly describe each of the SQL statements used in the process; and refer to specific SQL statements in your SQL script file.

Steps:

1) The following query shows the number of unique firms present in the main dataset (it should be 100). Please modify this query (or write a new query) to show the number of firms that are represented in at least 30 days. Based on your query, how many different firms are represented in at least 30 days in the main dataset?

```
Select count(*) from  (
        Select null from Tweets  group by symbol) t;
```

2) Create and populate a table which shows average Twitter levels immediately following each earnings release, and a comparative baseline average for the firm and specific time period. The following columns should be included:
    a. Average tweet levels (tweets per minute) in **the 40 minutes** following earnings release.
    b. Average tweets per minute for the firm in **the 2 hours** following earnings release.
    c. Average tweets per minute for the firm in **the 24 hours** following earnings release.
    d. Average tweets per minute for the firm **in the one week** following earnings release
    e. **Baseline average:** Average tweets per minute from the beginning of data collection (approx. May - June 2012, depending on the firm) to one week before the earnings release date; *matching on firm, day of week, and half-hour of day.* Hint: You may create a Date table to do the matches, though this is not required.

i. For example, suppose (hypothetically) that a Twitter peak is identified for the firm Adobe (ADBE) on February 14, 2013 between 10:59 am and 11:10 am. Note that this is a Thursday morning. The baseline average corresponding to this peak would be the average tweets per minute mentioning the firm ADBE between 11 am and 11:30 am on all Thursday mornings prior to (and not including) Feb. 14, 2013. (You may use the end-time of the increments to do the time of day matching. In the example here, 11:10 am was used rather than 10:59 am. Hence the matching half-hour of day is between 11 am and 11:30 am.)

f. Please describe briefly: Compare the average Tweet levels in a through d with the baseline average described in e.

3) Apply the calculation methods in step 2 to the trading volume measure (volumeend). In particular, include the following columns in your table of results for each earninges release, and show all SQL scripts that populate the table. You may present these within the same or in different SQL scripts as in step 1
   a. Average trading volume levels for the firm in **the 40 minutes following** earnings release.
   b. Average trading volume levels for the firm in the **2 hours following** earnings release.
   c. Average trading volume levels for the firm in the **24 hours following** earnings release.
   d. Average trading volume levels for the firm in **the one week following** earnings release.
   e. Baseline average trading volume: Average trading volume levels for the firm from the beginning of the sample period (approx. May - June 2012, depending on the firm) to one week before the earnings release date; *matching on firm, day of week, and half-hour of day.*
   f. For discussion: Compare the average trading volume levels in a through d with the baseline average described in e.

IDS 521 Homework 2
Prof. Ali Tafti
Fall 2014
This version: November 6, 2014

Parts 4 and 5 are optional, but may be done for the final course project, in lieu of the Dimensional Model Design. If this option is chosen for the final project, the final project deliverable should include a well-written analysis based on the results from the SQL exercises below.

4) **(Optional – Consider for final project)** The attached research paper (Tafti, Zotti and Jank 2014) defines peaks in Twitter activity as occurring whenever the number of tweets-per-minute that mention the firm in a 10-minute period exceed the established 99th percentile threshold for that firm. The 99th percentile threshold is a changing quantity, based on the past history of data from the beginning of data collection (approx. May - June 2012) for each firm up to the end of the prior week. For example, see the graph in Figure 5 of the paper that shows how the 99th percentile threshold changes for Adobe over time.

   You will create a table in which each row contains a unique combination of firm and date (daily). Define the 99th percentile threshold to be based upon all of the data points from the beginning of data collection to prior day (this is slightly different from the research paper). For example, for Adobe on September 17, 2013, the 99th percentile level would be calculated based upon this history of Tweets mentioning Adobe up to an including September 16, 2013. You may omit non-trading days, or days for which there is no Twitter data from the calculations.

5) **(Optional – Consider for final project) Create a script of SQL statements that** populate a new table of Twitter peaks, defined as every instance when the number of Tweets for the firm exceeds the 99th percentile threshold established for that firm-day combination.
   a. For each Twitter peak, show the following: Average 40-minute change in volume of shares traded starting from the end of the period of the Twitter peak. Change is calculated as ($volumeend_{t+4}$ - $volumeend_t$) where the $volumeend_t$ represents the firm's trading volume at the end of the ten-minute increment t. Likewise, $volumeend_{t+4}$ represents trading volume for the same firm after forty minutes (or four periods into the future).
   b. Average 40-minute change in volume of shares traded, for the same combination of firm, day of week, and half-hour combination
   c. Calculate the mean and standard deviation for the quantities in a and b. How do they compare? Is there a significant difference between a and b?