IDS 521 Homework 2
Prof. Ali Tafti
Spring 2015
This version: March 3, 2015

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

In this assignment you will use SQL to do some variants of computations described in the attached research paper entitled "Real-time Diffusion of Information on Twitter and the Financial Markets." In particular, 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 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, as detailed in the steps below, measures changes in financial trading volume for stocks after earnings releases and also after peaks in Twitter activity.

Please submit the following to show your work for Parts 1 - 3:

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. The script needs to run without breaking from beginning to end in MySQL. Please build upon the SQL script provided to you for this exercise. Please indicate clearly the names of all project partners that worked with you, on top of the file.
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 ***small*** tables of final 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. ***Please do not paste large tables in your final report; simply refer to them in your database by name and describe how your script creates them.*** Please indicate clearly the names of all project partners that worked with you, on top of the file.

Part 1: Trading volume levels following Earnings Releases

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) **Write a script of SQL statements to** create and populate a table that shows changes in trading volume immediately following each earnings release. For each earnings release, show the 40-minute change in volume of shares traded immediately after the earnings release. 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).

IDS 521 Homework 2
Prof. Ali Tafti
Spring 2015
This version: March 3, 2015

Part 2: **Identifying Peaks in Twitter Activity**

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 or 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.

3) **Write a script of SQL statements to create** a table in which each row represents a unique combination of firm and date (daily), and lists a 99th percentile threshold for that firm-date combination. Define the 99th percentile threshold to be based upon all of the data points from the beginning of data collection up to the 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 the history of Tweets mentioning Adobe up to and including September 16, 2013. You may omit non-trading days, or days for which there is no Twitter data from the calculations.

4) **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. For each Twitter peak, show the following: 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). Thus, the resulting table should have one row for each identified Twitter peak.

Part 3: Comparing the Trading Volume for Twitter Peaks and Earnings Releases

5) **Create a script of SQL statement(s) that creates a table** with three baseline average forty-minute changes in trading volume for each firm, beginning at 10 am, 12 pm (noon) and 2 pm; **averaged only over days in which there was no Twitter peak and no earnings release in the same week.** Thus, the table should have a row for each firm, and each of the baseline trading volume change calculations can be shown in three separate columns for each firm in the dataset. In your report, please describe how the baseline levels of trading volume change compare to trading volume changes following a Twitter peak event, and following an earnings release event.