Board 34: Use of Big Data Analytics in a First Year Engineering Project

Dr. Kevin D. Dahm, Rowan University

Kevin Dahm is a Professor of Chemical Engineering at Rowan University. He earned his BS from Worcester Polytechnic Institute (92) and his PhD from Massachusetts Institute of Technology (98). He has published two books, "Fundamentals of Chemical Engineering Thermodynamics" and "Interpreting Diffuse Reflectance and Transmittance." He has also published papers on effective use of simulation in engineering, teaching design and engineering economics, and assessment of student learning.

Nidhal Carla Bouaynaya, Rowan University

Nidhal Bouaynaya received the B.S. degree in Electrical Engineering and Computer Science from the Ecole Nationale Superieure de L'Electronique et de ses Applications (ENSEA), France, in 2002, the MS degree in Electrical and Computer Engineering from the Illinois Institute of Technology, Chicago, in 2002, the Diplome d'Etudes Approfondies in Signal and Image processing from ENSEA, France, in 2003, the M.S. degree in Mathematics and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Chicago, in 2007. From 2007-2013, she was an Assistant then Associate Professor with the Department of Systems Engineering at the University of Arkansas at Little Rock. Since 2013, she joined Rowan University, where she is currently an Associate Professor with the Department of Electrical and Computer Engineering. Dr. Bouaynaya won the Best Student Paper Award in Visual Communication and Image Processing 2006, the Best Paper Award at the IEEE International Workshop on Genomic Signal Processing and Statistics 2013 and the runner-up Best Paper Award at the IEEE International Conference on Bioinformatics and Biomedicine 2015. She is also one of the winners of the Brain Tumor Image Segmentation (BRATS) Challenge 2016. Her current research interests are in medical imaging, machine learning, mathematical biology and dynamical systems.

Dr. Ravi P. Ramachandran, Rowan University

Ravi P. Ramachandran received the B. Eng degree (with great distinction) from Concordia University in 1984, the M. Eng degree from McGill University in 1986 and the Ph.D. degree from McGill University in 1990. From October 1990 to December 1992, he worked at the Speech Research Department at AT&T Bell Laboratories. From January 1993 to August 1997, he was a Research Assistant Professor at Rutgers University. He was also a Senior Speech Scientist at T-Netix from July 1996 to August 1997. Since September 1997, he is with the Department of Electrical and Computer Engineering at Rowan University where he has been a Professor since September 2006. He has served as a consultant to T-Netix, Avenir Inc., Motorola and Focalcool. From September 2002 to September 2005, he was an Associate Editor for the IEEE Transactions on Speech and Audio Processing and was on the Speech Technical Committee for the IEEE Signal Processing society. Since September 2000, he has been on the Editorial Board of the IEEE Circuits and Systems Magazine. Since May 2002, he has been on the Digital Signal Processing Technical Committee for the IEEE Circuits and Systems society. His research interests are in digital signal processing, speech processing, biometrics, pattern recognition and filter design.

Use of Big Data Analytics in a First Year Engineering Project

This paper describes a module on big data analytics that was developed and introduced into Freshman Engineering Clinic, which is an introductory course for students in all engineering disciplines at Rowan University. Learning objectives for the Freshman Engineering Clinic include developing skills in data collection, analyzing data to draw sound conclusions, and writing reports, with visual/graphical representation of information recognized as one critical component of effective technical writing. The NSF has awarded a grant to Rowan University to support vertical integration of big data analytics throughout the engineering curriculum. This paper focuses on the Freshman Clinic big data project, the intent of which was to introduce students to big data analytics while also furthering the general instructional objectives of the freshman course.

The project was titled "Introduction to Big Data Analytics: Analyzing Tweets with Matlab". The instructor provided the students with a Matlab code that was designed to facilitate applying Sentiment Analysis to tweets. For example, the code can be used to (1) identify tweets that contain one or more specific keywords and (2) create a histogram of words used in these tweets, in order to identify recurring themes in tweets that mention the keyword(s). The final deliverable for the project was a report in which students detailed how they used the Matlab code to answer a number of open-ended questions, as well as an introductory section in which students discussed the importance and applications of big data analytics in general. This paper describes the project and the expectations for the write-up that was submitted by the students, and also presents the results obtained from a 10-question concept inventory that was used for preliminary assessment.

Introduction

Data is more thoroughly integrated into our society and our lives than ever before. Ever-expanding amounts of heterogeneous data have become available across all disciplines. Examples include humanities and social sciences [1], political sciences [2], weather monitoring and forecasting [3], web intelligence [4], healthcare and pharmaceuticals [5], and linguistic and psychological studies [6]. In 2013, estimates reached 4 zettabytes of data generated worldwide [7]. Technological advances have driven down the cost of creating, capturing, managing, and storing information. The National Science Foundation has defined big data as datasets "large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future" [8].

Making effective use of big data can boost economic productivity, drive improvements to consumer service, thwart terrorists, and save lives. As one example, the healthcare industry witnessed a 20% decrease in patient mortality by analyzing streaming patient data [9]. Big data and the growing "Internet of Things" have also made it possible to merge the industrial and

information economies. As one example, jet engines and delivery trucks can now be outfitted with sensors that monitor hundreds of data points and send automatic alerts when maintenance is needed [10].

However, the enormity and complexity of the data present great challenges in analyses and subsequent applications. A 2011 McKinsey report on big data for competition and productivity predicted that the United States would experience a shortage of the analytical and managerial talent necessary to make the most of big data [11]. Recognizing the need for increased expertise in big data technology, some educational institutions have taken steps to address this need. For example, Ohio State University (OSU) has started to offer a data analytics undergraduate major [12]. There are also a few graduate programs in big data analytics, such as those at Harvard University (MS or ME in Computational Science and Engineering [13]), North Carolina State University (MS in Analytics [14]), Northwestern University (MS in Analytics [15]), Stanford University (MS in Computer Science, Specialization in Information, Management, and Analytics [16]) and Southern New Hampshire University (online MS in Data Analytics [17]). Thus, most of the programs identified by the authors are at the level of graduate education.

The literature does contain examples of undergraduate experiences in big data that are more modest in scope than a full degree program. As a recent example, Pettis and co-authors [18] describe the integration of big data throughout an undergraduate computer science program, specifically implementing new one-week big data modules into several required mathematics courses. The current paper is part of an NSF-sponsored project that has very similar goals. The overall aim of the NSF grant proposal is to implement a vertical integration of big data applications and examples throughout the undergraduate curriculum in several different engineering disciplines. This paper focuses on one module that was incorporated into Freshman Engineering Clinic, an introductory course that is required for all engineering majors at Rowan University. The Rowan Clinic program is described in more detail by Jansson and co-authors [19]. Another recent example of big data in the undergraduate curriculum is given by Omar [20], who describes the development of a big data lab illustrating management of broadband wireless networks. This lab is nearing completion but has not yet been used in undergraduate courses. The motivation for this lab mirrors the previous discussion, but the paper details extensive technical challenges involved in design of the lab. The module described in this paper has very modest resource requirements, as it uses only MATLAB and readily available internet tools and data.

Freshman Clinic Module on Big Data Analytics

The big data module can be broadly divided into two parts, both of which examined data from Twitter. The first part of the big data module was an exercise in “sentiment analysis” as revealed by the content of tweets. The second part examined features of the social network within Twitter, such as the phenomenon of influential users, by examining numbers of followers and numbers of retweets.

The technical basis for the module was a series of MATLAB codes that can be used to mine and process data in various forms from Twitter. The first-semester freshman had no prior experience

with MATLAB, so the complete MATLAB codes were presented to the students, and students learned enough about MATLAB to run the programs and interpret the results. There are subsequent courses in the Engineering curriculum that require students to construct their own programs in MATLAB. One example is a sophomore wind turbine design project in which students built models of wind turbines in MATLAB and used these models for iterative optimization [21].

The first MATLAB code is used to find and store 100 tweets that each mention specific keywords. In this module, students used the code to find tweets that mentioned both "Amazon" and "Hachette." The relationship between Amazon and Hachette was considered a good example for sentiment analysis because many people were affected by their highly publicized dispute over e-book prices, as well as its eventual resolution [22]. The code, however, is readily modified to search for different keywords.

The second MATLAB code is used to identify all other words used in the 100 tweets stored by the first search, and store word counts in arrays. A list of "stop words" like "a" and "the" were excluded, as these are words that are very common but provide no insight into sentiment analysis. A third code is used to produce visual displays of the information from the arrays, in the form of a "word histogram." Students ran these programs and used the results to identify dominant themes in the stored tweets.

A fourth MATLAB code was used to identify the users within a data set who had the most followers- in this case, the set of people who had tweeted about Amazon and Hachette. A fifth MATLAB code was used to tabulate how frequently specific users tweeted and how frequently their tweets were re-tweeted. This program was applied to a larger data set (1000 tweets) on four topics that were trending at the time, as identified by their use of a common key phrase such as "I can't sing," or by their use of a common hashtag such as #InABlackHousehold. By using these programs, students were able to look for correlations between followers and retweets. Analysis of the results included students' thoughts on "followers" and "retweets" as metrics for how "influential" a Twitter user is. As a final step, students used the social network analysis tool Gephi (https://gephi.org/) to produce graphical representations of data sets, with an example from a student report shown in Figure 1. The sizes of the nodes in Figure 1 are weighted by the number of followers the user has, and the lines represent re-tweets.

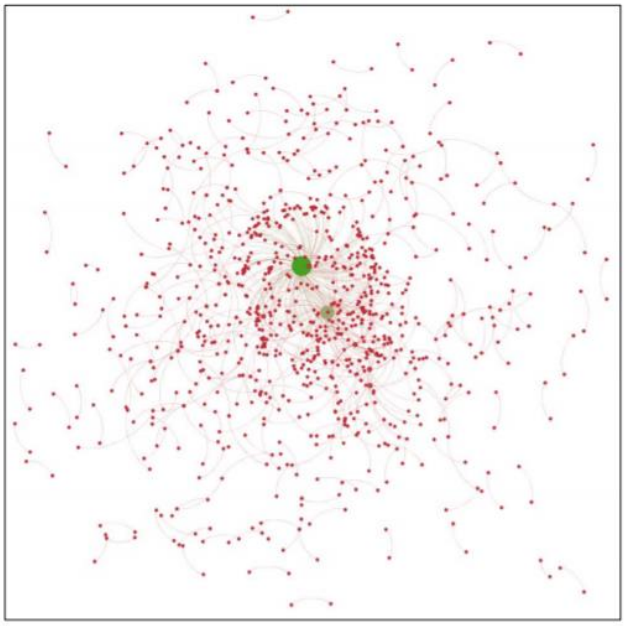

Figure 1: A graphical depiction of the social network defined by a set of tweets on a trending topic.

The complete handout (including MATLAB codes) that was used to communicate the goals and expectations for the module to the students is included in this paper as an Appendix. The deliverable for the module was a report that consisted of:

- An introductory section (20 points) in which students discussed the applications and importance of big data analytics in general.
- The main body of the report (70 points), in which students answered specific open-ended questions related to the results they obtained from the aforementioned activities.
- Conclusions (10 points)

Preliminary Assessment

A goal of the project was introducing students to big data and the prevalence of big data applications in our society. This was assessed by administering a 10-question concept inventory [23] to freshmen at the beginning of the semester and again after completion of the Freshman Engineering Clinic project related to big data. The concept inventory consisted of multiple choice questions as well as one free response question in which students were asked to name "three companies or industries that are involved in big data." Table 1 shows the results of the concept inventory. Student performance improved on the post-test relative to the pre-test from 5.63/10 to 6.26/10. While this improvement does not appear dramatic, it is statistically significant to 95% confidence ($p=0.023$).

Table 1: Results for a 10-question concept inventory that was administered to Freshman Engineering Clinic I students both before and after completing the Freshman Engineering Clinic project.

	Pre-Project	Post-Project
Number of respondents	30	34
Mean Score	5.63	6.26
Standard Deviation	1.19	0.93

In addition, the same concept inventory was administered to three other sections of Freshman Engineering Clinic at the end of the semester. These three sections did different projects that were unrelated to big data, and serve as a control group. The mean scores of the three control group sections were 4.39, 4.50 and 4.53, with a super average of 4.47 out of 10. Consequently, the control groups had a lower performance at the *end* of the semester than the experimental group at the *beginning* of the semester. This suggests that the improvement shown in Table 1 was attributable to the big data project specifically, rather than to the Freshman Engineering Clinic course outcomes in general.

Summary

This paper describes a project module on big data analytics that was incorporated into a multidisciplinary first year engineering course. The module was well suited for the instructional objectives of the course, which include collection, interpretation, and presentation of data, and writing effective reports. It can readily be adapted for introductory engineering courses at other universities. Preliminary assessment data shows that the project experience led to a modest but measurable increase in student awareness of big data concepts.

Acknowledgement

The authors gratefully acknowledge the support of NSF grant #1610911, entitled “ENGAGING IN STEM EDUCATION WITH BIG DATA ANALYTICS AND TECHNOLOGIES: A ROWAN-COVE INITIATIVE.”

Literature Cited

1. K.H. Leetaru, "A Big Data Approach to the Humanities, Arts, and Social Sciences: Wikipedia's view of the World through Supercomputing", Research Trends, vol. 30, September 2012.
2. S. Ansolabehere and E. Hersh, "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." Political Analysis, vol. 20, 2012, pp. 437-59.
3. NOAA Big Data project: https://data-alliance.noaa.gov/, accessed Dec. 12, 2018.
4. G. Shroff, The Intelligent Web: Search, Smart Algorithms and Big Data, Oxford University Press, UK, November 2013.
5. W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2, no. 3, 2014.
6. C.N. DeWall, R.S. Jr. Pond, W.K. Campbell, J.M. Twenge, "Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics", Psychology of Aesthetics, Creativity, and the Arts, vol. 5, no. 3, Aug. 2011, pp. 200-207.
7. M. and L. Yu, KPBC Internet Trends conference 2013, http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013, accessed Dec. 12, 2018.
8. National Science Foundation, Solicitation 12-499: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), 2012, http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf, accessed Dec. 12, 2018.
9. IBM webcast: Big Data at the Speed of Business, 2013.
10. Salesforce.com, "Collaboration helps GE Aviation bring its best inventions to life," https://www.salesforce.com/ap/customer-success-stories/ge/, accessed Dec. 12, 2018.
11. The McKinsey Global Institute, "Big data: The next frontier for innovation, competition, and productivity," June 2011.
12. https://data-analytics.osu.edu/future-students/future-freshmen, accessed Dec. 14, 2018.
13. https://gsas.harvard.edu/programs-of-study/all/computational-science-engineering, accessed Dec. 14, 2018.
14. https://analytics.ncsu.edu/, accessed Dec. 14, 2018.
15. https://www.mccormick.northwestern.edu/analytics/, accessed Dec. 14, 2018.
16. https://cs.stanford.edu/degrees/mscs/specializations/, accessed Dec. 14, 2018.
17. https://degrees.snhu.edu/programs/ms-in-data-analytics, accessed Dec. 14, 2018.
18. C. Pettis, R. Swamidurai, A. Abebe, and D. Shannon, "Infusion of Big Data Concepts Across the Undergraduate Computer Science Mathematics and Statistics Curriculum," Proceedings of the ASEE Conference and Exposition, June 2018.
19. Jansson, P., Ramachandran, R., Schmalzel, J. and Mandayam, S., "Creating an Agile ECE Learning Environment Through Engineering Clinics," IEEE Transactions on Education, 53, 3, August 2010.

20. T. Omar, “Implementation of Big Data Lab for Broadband Wireless Networks Intelligent Traffic Management System: Evaluation and Challenges”, Proceedings of the ASEE Conference and Exposition, June 2018.
21. S. Bakrania, W. Riddell, K. Dahm and L. Weiss, “Wind Turbines for Teaching Parametric Design,” ASEE Annual Conference and Exposition, June 2009, Austin, TX.
22. B. Stelter, “Amazon and Hachette settle bitter fight over e-book pricing,” CNN Business, available at https://money.cnn.com/2014/11/13/media/amazon-hachette-reach-deal/index.html, accessed Dec. 14, 2018
23. T. Ogunfunmi, G. Herman, and M. Rahman, “On the Use of Concept Inventories for Circuits and Systems Courses,” IEEE Circuits and Systems Magazine, Third Quarter 2014.

Appendix: Student Handout

This is the student handout that was used to introduce the Big Data Analytics project to the students. (Aside from the suppression of the instructor's name, it is the unedited handout as it was presented to the students, and therefore may not align with ASEE formatting guidelines.)

Introduction to Big Data Analytics: Analyzing Tweets with Matlab

Freshmen Engineering Clinic

Fall 2017

Instructor: XXXXX

Whatever your opinion of social media these days, there is no denying it is now an integral part of our digital life. So much so, that social media metrics are now considered part of *altmetrics*, an alternative to the established metrics, such as citations, to measure the impact of scientific papers.

Why Twitter?

Twitter is a good starting point for social media analysis because people openly share their opinions to the general public. This is very different from Facebook, where social interactions are often private. In this project, you will perform basic sentiment analysis and social graph visualization using Twitter's Search and Streaming Application Programming Interface (API).

Sentiment Analysis

One of the very common analyses you can perform on a large number of tweets is sentiment analysis. Sentiment is scored based on the words contained in a tweet. If you manage a brand or political campaign, for example, it may be important to keep track of your popularity, and sentiment analysis provides a convenient way to take the pulse of the tweeting public.

Final Project Format

I. Introduction to Big Data Analytics in general. Discuss importance, applications (e.g., politics, health, industry, etc.) and perspectives.

   In a second paragraph, discuss importance of sentiment analysis from Twitter data as a special application of big data. Check out some examples from this search result from PLOS ONE that list various papers that used Twitter for their study:

   http://journals.plos.org/plosone/search?from=globalSimpleSearch&filterJournals=PLoSONE&q=twitter&x=0&y=0

   [20 pts]

II. Use the below Matlab code and answer the questions to analyze Twitter with Matlab. Include the below Matlab code in your final report. [70 pts]

III. Conclusion [10 pts]

Check out this final project on sentiment analysis using Twitter to capture the mood of the people after President Trump's election:
http://rstudio-pubs-static.s3.amazonaws.com/237114_e4b645db7c8c449b882e68e7ed32bec9.html

I. Getting started with Twitter using Twitty

To get started with Twitter, you need to get your developer credentials. You also need Twitty by Vladimir Bondarenko. It is simple to use and comes with excellent documentation.

1. Create a Twitter account if you do not already have one:
   https://twitter.com/
2. Create a Twitter app to obtain developer credentials:
   https://apps.twitter.com/
3. Download and install Twitty from the FileExchange, along with the JSON Parser and optionally JSONLab:
   Twitty: https://www.mathworks.com/matlabcentral/fileexchange/34837-twitty
   JSON Parser: https://www.mathworks.com/matlabcentral/fileexchange/20565-json-parser
   JSONLab: https://www.mathworks.com/matlabcentral/fileexchange/33381-jsonlab--a-toolbox-to-encode-decode-json-files
4. Create a structure array to store your credentials for Twitty

The following Matlab code searches for tweets that mention 'amazon' and 'hachette'.

```
%% a sample structure array to store the credentials
creds = struct('ConsumerKey','3Q7ObFutZyYZtkYLoyF721cWk',...
  'ConsumerSecret','gWclYtHRdist7u9H2FdRriJZG1ad0wGNkJjfGyjV6crAwrzJsS',...
  'AccessToken','932272565581680640-eqUF30wkJ9VbTHTUB6p1NQVK0Az4fiZ',...
  'AccessTokenSecret','EKYkh99ujjmIdXZOAhbZYurrYDysa4Y60FOPJgXBnmTbZ');

% set up a Twitty object
addpath Twitty; % Twitty
addpath parse_json; % Twitty's default json parser
addpath jsonlab; % I prefer JSONlab, however.
tw = twitty(creds); % instantiate a Twitty object
tw.jsonParser = @loadjson; % specify JSONlab as json parser

% search for English tweets that mention 'amazon' and 'hachette'
amazon = tw.search('amazon','count',100,'include_entities','true','lang','en');
hachette = tw.search('hachette','count',100,'include_entities','true','lang','en');
both = tw.search('amazon hachette','count',100,'include_entities','true','lang','en');
```

II. Processing Tweets and Scoring Sentiments

Twitty stores tweets in structure array created from the API response in JSON format. It may be preferable to use a table when it comes to working with heterogeneous data containing a mix of numbers and text. The Matlab function processTweets.m, converts structure arrays into tables and computes sentiment scores. Make sure you add the Amazon-Hachette data file to the same folder.
For sentiment analysis, we will use AFINN (http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010), along with a list of English stop words (https://www.textfixer.com/tutorials/common-english-words.txt) so that we don't count frequent common words like "a" or "the".

```
%% load supporting data for text processing
scoreFile = 'AFINN/AFINN-111.txt';
stopwordsURL ='http://www.textfixer.com/resources/common-english-words.txt';
% load previously saved data
load amazonHachette.mat

% process the structure array with a utility method |extract|
[amazonUsers,amazonTweets] = processTweets.extract(amazon);
% compute the sentiment scores with |scoreSentiment|
amazonTweets.Sentiment = processTweets.scoreSentiment(amazonTweets, ...
   scoreFile,stopwordsURL);

% repeat the process for hachette
[hachetteUsers,hachetteTweets] = processTweets.extract(hachette);
hachetteTweets.Sentiment = processTweets.scoreSentiment(hachetteTweets, ...
   scoreFile,stopwordsURL);

% repeat the process for tweets containing both
[bothUsers,bothTweets] = processTweets.extract(both);
bothTweets.Sentiment = processTweets.scoreSentiment(bothTweets, ...
   scoreFile,stopwordsURL);

% calculate and print NSRs
amazonNSR = (sum(amazonTweets.Sentiment>=0) ...
   -sum(amazonTweets.Sentiment<0)) ...
   /height(amazonTweets);
hachetteNSR = (sum(hachetteTweets.Sentiment>=0) ...
   -sum(hachetteTweets.Sentiment<0)) ...
   /height(hachetteTweets);
bothNSR = (sum(bothTweets.Sentiment>=0) ...
   -sum(bothTweets.Sentiment<0)) ...
   /height(bothTweets);
fprintf('Amazon NSR  :  %.2f\n',amazonNSR)
fprintf('Hachette NSR:  %.2f\n',hachetteNSR)
fprintf('Both NSR    : %.2f\n\n',bothNSR)

% plot the sentiment histogram of two brands
binranges = min([amazonTweets.Sentiment; ...
   hachetteTweets.Sentiment; ...
   bothTweets.Sentiment]): ...
   max([amazonTweets.Sentiment; ...
```

```
   hachetteTweets.Sentiment; ...
   bothTweets.Sentiment]);
bincounts = [histc(amazonTweets.Sentiment,binranges)...
   histc(hachetteTweets.Sentiment,binranges)...
   histc(bothTweets.Sentiment,binranges)];
figure
bar(binranges,bincounts,'hist')
legend('Amazon','Hachette','Both','Location','Best')
title('Sentiment Distribution of 100 Tweets')
xlabel('Sentiment Score')
ylabel('# Tweets')
```

We will use the following metric for comparison: *Net Sentiment Rate (NSR)*

```
NSR = (Positive Tweets-Negative Tweets)/Total
```

1. Run the above code in a Matlab script and display the NSR results. [5 pts]
2. You could keep taking this measurement periodically for ongoing sentiment monitoring, if interested. Perhaps you may discover that NSR is correlated to their stock prices! [Bonus: 20 pts]
3. Plot the sentiment distribution of 100 tweets [5 pts]

III. Processing Tweets for Content Visualization

The function processTweets also has a function tokenize that parses tweets to calculate the word count.

```
%% tokenize tweets with |tokenize| method of |processTweets|
[words, dict] = processTweets.tokenize(bothTweets,stopwordsURL);
% create a dictionary of unique words
dict = unique(dict);
% create a word count matrix
[~,tdf] = processTweets.getTFIDF(words,dict);

% plot the word count
figure
plot(1:length(dict),sum(tdf),'b.')
xlabel('Word Indices')
ylabel('Word Count')
title('Words contained in the tweets')
% annotate high frequency words
annotated = find(sum(tdf)>= 10);
jitter = 6*rand(1,length(annotated))-3;
for i = 1:length(annotated)
   text(annotated(i)+3, ...
      sum(tdf(:,annotated(i)))+jitter(i),dict{annotated(i)})
end
```

4. Run the above code and plot the words contained in the tweets. [5 pts]

5. What were the main themes they tweeted about when those users mentioned both Amazon and Hachette? [5 pts]

Who Tweeted the News?

The 100 tweets collected came from 86 users. So on average each user only tweeted 1.16 times. Instead of frequency, let's find out who has a large number of followers (an indicator that they may be influential) and check their profile.

IV. Get the Profile of Top 5 Users

Twitty also supports the 'users/show' API to retrieve user profile information. Let's get the profile of the top 5 users based on the follower count.

```
%% sort the user table by follower count in descending order
[~,order] = sortrows(bothUsers,'Followers','descend');
% select top 5 users
top5users = bothUsers(order(1:5),[3,1,5]);
% add a column to store the profile
top5users.Description = repmat({''},height(top5users),1);
% retrieve user profile for each user
for i = 1:4
   userInfo = tw.usersShow('user_id', top5users.Id(i));
   K = strfind(userInfo, 'description');
   userInfo_description_temp = userInfo(K(1)+13:end);
   I = strfind(userInfo_description_temp,"");
   description = userInfo_description_temp(1:I(2));
   if ~isempty(description)
      top5users.Description{i} = description;
   end
end
% print the result
disp(top5users(:,2:end))
```

6. Display the name, number of followers and descriptions of the top 5 users. Comment on the user profiles and the number of their followers. [5 pts]

V. Streaming API for High Volume Real Time Tweets

In the previous section, we checked out the top 5 users based on their follower count. The assumption was that, if you have a large number of followers, you are considered more influential because more people may see your tweets. Now let's test this assumption. For that we need more than 100 tweets.

If you need more than 100 tweets to work with, then your only option is to use the Streaming API, which provides access to the sampled Twitter fire hose in real time. That also means you need to access the tweets that are currently active. You typically start with a trending topic from a specific

location. You get local trends by specifying the geography with WOEID (Where On Earth ID), available at WOEID Lookup: http://woeid.rosselliot.co.nz/

We collected a new batch of data - 1000 tweets from 4 trending topics from the UK in the Matlab data file "uk_data.mat".

```
% Load 'uk_data.mat' for 4 trending topics in the UK
load('uk_data.mat')

% plot
figure
for i = 1:4
   % process tweets
   [users,tweets] = processTweets.extract(uk_data(i).statuses);

   % get who are mentioned in retweets
   retweeted = tweets.Mentions(tweets.isRT);
   retweeted = retweeted(~cellfun('isempty',retweeted));
   [screen_names,~,idx] = unique(retweeted);
   count = accumarray(idx,1);
   retweeted = table(screen_names,count,'VariableNames',{'Screen_Name','Count'});

   % get the users who were mentioned in retweets
   match = ismember(users.Screen_Name,retweeted.Screen_Name);
   retweetedUsers = sortrows(users(match,:),'Screen_Name');
   match = ismember(retweeted.Screen_Name,retweetedUsers.Screen_Name);
   retweetedUsers.Retweeted_Count = retweeted.Count(match);
   [~,order] = sortrows(retweetedUsers,'Retweeted_Count','descend');

   % plot each topic
   subplot(2,2,i)
   scatter(retweetedUsers.Followers(order),...
      retweetedUsers.Retweeted_Count(order),retweetedUsers.Freq(order)*50,...
      retweetedUsers.Freq(order),'fill')

   if ismember(i, [1,2])
      ylim([-20,90]); xpos = 2; ypos1 = 50; ypos2 = 40;
   elseif i == 3
      ylim([-1,7])
      xlabel('Follower Count (Log Scale)')
      xpos = 1010; ypos1 = 0; ypos2 = -1;
   else
      ylim([-5,23])
      xlabel('Follower Count (Log Scale)')
      xpos = 110; ypos1 = 20; ypos2 = 17;
   end

   % set x axis to log scale
   set(gca, 'XScale', 'log')

   if ismember(i, [1,3])
```

```
    ylabel('Retweeted Count')
  end
  title(sprintf('UK Tweets for: "%s"',uk_data(i).query.name))
end
```

7. Run the above code to plot the users based on their follower counts vs. how often their tweets got retweeted. The size (and the color) of the bubbles show how often those users tweeted. [5 pts]

8. Comment on the importance of the number of followers and the frequency of getting retweeted. What is the difference between the top two charts and the bottom two charts? Analyze how a tweet goes viral. [10 pts]

VI. Social Graph Visualization

Retweeting of one user's tweet by others creates a network of relationships that can be represented as a social graph. We can visualize such relationship with a popular social networking analysis tool Gephi: https://gephi.org/

Gephi imports an edge list in CSV format. The following code saves the screen names of the users as source and the hashtags and screen names they mention in their tweets as target in a Gephi-ready CSV file.

```
processTweets.saveEdgeList(uk_data(1).statuses,'edgeList.csv');
```

File "edgeList.csv" was successfully saved.

9. Use Gephi to plot the “"I Can't Sing" Social Graph” [10 pts]

10. Use Gephi to plot the "#InABlackHousehold" Social Graph [10 pts]

11. Analyze the differences between the two social graphs. [10 pts]

We only scratched the surface in data analytics as applied to Twitter data, and hopefully you got the taste of what kind of analyses are possible, and their importance in understanding and influencing social behavior.