INFO-633: Information Visualizations
The VAST Mini-Challenge
Using information visualizations to predict the opening weekend take and user ratings of a blockbuster film.

Travis Eckenroth, Troy Finamore, and Gretel Miller
College of Information Science and Technology
Drexel University

Abstract— The goal of the 2013 edition of the VAST Mini-Challenge was to make a prediction regarding the success of the movie *Oz the Great and Powerful* utilizing skills in visual analytics. Movie success is determined by tickets sales as well as the movie's view ratings during opening weekend. This analysis discusses some challenges with visual analytics, the history of predictive techniques used to estimate movie success, the approach used for this specific prediction including closed-world data provided by the challenge organizers and how it was used to create predictive visualizations. The resulting conclusion was that of a successful opening weekend for Oz.

Index Terms— data, forecasting, information visualization, revenue, Twitter, IMDb, VAST Mini-Challenge, visual analytics

1 Introduction

Applying the use of information visualization analytics has greatly improved the readability and use of large sums of data. It has only been widely adopted as a form of scientific analytics within the past ten years and has provided a major advancement in the field. Visual analytics includes accessing data, transforming the data into visualizations and data mining and refining models, and creating knowledge based on the assessments. To apply a definition to this field of science, Kohlhammer et al., 2011, states, "Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets."[1]

The successful utilization of visual analytics depends on the relevancy of the data obtained to the subject of study. Obtaining irrelevant data as well as processing or presenting data in an inappropriate fashion has the potential to render the study ineffective. To minimize this chance, the field of visual analytics provides goals such as providing timely, defensible, and understandable assessments of data as well as communicating the assessments effectively.[2] As the research and analytics are processed for the needs of this study, only relevant data from defined sources will be obtained and utilized according to the guidelines presented by the VAST Mini Challenge 2013: IMDb and Twitter.[3]

A visual analytical study will be conducted on the movie *Oz the Great and Powerful*, with an opening box office date of 8 March, 2013, to provide an estimate of the opening weekend success of the film. The success of the film will be determined by two factors: the box office take and user ratings. The data sources provided will be used to predict these two determining factors through the use of open-source software. Considering the youthful characteristics of the field of visual analytics and the complexity of revenue forecasting in the film industry, some challenges are to be expected.

2 Challenges

The VAST Mini-Challenge will test our ability to manage large data and to make decisions in a short time frame. The goal is to analyze data through the use of visualization techniques and predict the expected revenue of the film. In order to comply with the challenge's requirements we are restricted to using closed world data as specified by the organizers of the challenge.[4]

A series of defined challenges have been published throughout the years pertaining to the analytics field. Some of the challenges presented in scientific publishing are relevant to the VAST Mini-Challenge study, whereas others are not so pertinent due to the focus on extreme scale analysis rather than a few hundred megabytes worth of data. However, despite the possible irrelevancy, it is important to briefly discuss the challenges associated with visual analytics followed by specific challenges that will be faced while performing a box office forecast of a Hollywood movie.

[1] Kohlhammer J., et al. (2011). Solving Problems with Visual Analytics. *Procedia Computer Science (7)*. Retrieved from http://bib.dbvis.de/uploadedFiles/373.pdf

[2] Keim, D. et al. (2008). Visual Analytics: Definition, Process, and Challenges. *Information* Visualization. p 154-175. Retrieved from http://hal-lirmm.ccsd.cnrs.fr/docs/00/27/27/79/PDF/VAChapter_final.pdf

[3] VAST Mini-Challenge 2013. Retrieved from http://www.boxofficevast.org/vast-welcome.html

[4] VAST Mini-Challenge 2013. Ibid.

2.1 Visual Analytics Science & Technology

As Wong et al, (2012) refers to in the IEEE publishing, *The Top 10 Challenges in Extreme-Scale Visual Analytics*, issues within the field were visible from the beginning of visual analytical research. As technology rapidly advanced, some of the challenges have been unsolved. To put this into perspective, in the early stages of use, scientists were analyzing terabytes (10^{12}) and petabytes (10^{15}) of data. Despite known challenges with this amount of data, technology had increased the use of data to exabytes (10^{18}).[5] This contributes to the list of challenges in visual analytics.

Wong et al, (2012) summarized ten challenges that were published during a Department of Energy (DOE) Office of Advanced Scientific Computing Research (ASCR) workshop. The proceeding list of challenges is based off a research focused on the visualization and analysis of computational science applications at an extreme scale, which are summarized below.[6,7]

- In Situ Analysis: a switch from reading data from a hard drive to reading data from memory which creates challenges with interactive analytics, algorithms, memory, I/O, workflow, and threading.
- Interaction and User Interfaces: the expanding ratio of technology development to human cognition.
- Large-Data Visualization: human's inability to clearly see through the large scale of visualizations.
- Databases and Storage: issues revolving around data residing in cloud storage.
- Algorithms: the need for adaptable learning algorithms to aid in data exploration for humans.
- Data Movement/Transport and Network Infrastructure: the need for algorithms to effectively use network resources
- Uncertainty Quantification: inability to exam entire data sources, thus relying on subsampling which increases the possibility of misleading results.
- Parallelism: maximizing the use of processor memory with newer analytical models
- Domain and Development Libraries, Frameworks, and Tools: the lack of visualization software for high performance computers
- Social, Community, and Government Engagements: the need for community and government to engage in future development of technologies.

In addition to the aforementioned challenges of the visual analytics, Keim et al., (2008) reference the inability to provide extensive analytics on data due to the lack of computing power and infrastructure availability. The overall consensus of Keim et al. is that technology needs to be applied in a more effective manner in order for analysts to more easily examine the data.[8] Examining thousands of pieces of data can be a strenuous process without adequate software and computer capabilities. In addition to the general challenges, some applicable fields have become greatly influenced by this technology.

Environment and climate change makes promising use of visual analytics through the examination of historical data. Such data can be used to create recognizable patterns in meteorological changes or even animal population. Also, the use of geospatial analysis can be used to visualize the relationships among a modern community in terms of political and economic decisions, as well cultural and demographic effects. Financial analysis is also a promising field which visual analytics could provide exceptional results based on the data.[9]

The increasing amounts of data are continuously proving itself as a challenge to visual analytics, regardless of which applicable field is under study. The challenges mentioned previously are solely based on the use of large amounts of data as such fields are deeply involved in mass amounts. However, in a small scale level, those which only involve a few hundred megabytes of data, there are other challenges that are applicable in the analytical process.

2.2 Box Office Forecasting

The success of a movie is critical to the stakeholders involved and the conglomerate of individuals that it takes to create a movie. With movies being a large part of the American economy, and with having such a global reach, there is a concentrated focus on the success of the overall movie production in the U.S. The Motion Picture Association of America (MPAA) is a large influence in the industry as it is associated with some of the top movie production companies. In 2011, the overall global revenue of American movies increased 3% over 2010.[10]

> *Global box office for all films released in each country around the world reached $32.6 billion in 2011, up 3% over 2010's total, due entirely to the increase in international box office ($22.4 billion). Each international region experienced growth in 2011. Chinese box office grew by 35% in 2011 to become the 2nd largest International market behind Japan, experiencing by far the largest growth in major markets.[11](MPAA)*

Due to the size of this industry, it creates some obstacles when attempting to predict the success of a film. Adding the means of distribution (movie theaters) to consideration also adds additional obstacles outside the control of movie producers and production companies. Movie theater managers constantly decide on which movie to show on which screen as well as which movie to stop showing. It's a constant challenge as especially during peak seasons there are more mov-

[5] Pak et al., (2012). The Top 10 Challenges in Extreme-Scale Visual Analytics. *IEEE*. Retrieved from http://ieeexplore.ieee.org.ezproxy2.library.drexel.edu/stamp/stamp.jsp?tp=&arnumber=6265057

[6] Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale. (October 2007). Retrieved from http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Doe_visualization_report_2007.pdf

[7] Pak et al., Ibid.

[8] Keim, D. et al., Ibid.

[9] Keim, D. et al., Ibid.

[10] Theatrical Market Statistics 2011. (2011). Retrieved from http://www.mpaa.org/Resources/5bec4ac9-a95e-443b-987b-bff6fb5455a9.pdf

[11] Theatrical Market Statistics 2011. Ibid.

ies than there are screens.[12]

Eliashberg et al., (2009) describe a couple challenges specifically related to movie theaters. The first challenge pertains to the physical constraints of theaters. "There are different numbers of screens in the different theaters, different screening room capacities, different consumer preferences and demand situations..." (Eliashberg et al., 2009) In addition to that, some larger theaters tend to have contracts with certain production companies, therefore giving priority over other films. Following suit with that, there is a certain 'special treatment' involved when it comes to kid's movies and matinees.[13]

The second issue that is pointed out pertains to the actual decision making process for which movie is to be shown on which screen. Theater managers tend to be reluctant in predicting actual box office forecast based off of similar movies. The reasoning behind this is that, as described by Eliashberg et al., managers can often be punished for picking an unsuccessful movie to show. Despite their knowledge and experience in the film industry, without formal means of forecasting, selecting films becomes a guessing game.[14] This can negatively impact a film's success if certain theaters or showing films in smaller rooms as opposed to larger ones, or not even showing them at all due to informal forecasting or contractual agreements.

Box office forecasting has many challenges that range from the analytics used to predict the outcomes as well as industry related challenges. Some of the challenges lie outside the control of movie producers or even movie studios. However, from a research perspective, there are additional problems associated with the VAST Mini-Challenge. The VAST Mini-Challenge has defined guidelines which restrict the methods the user can use as well as restrictions on data. This creates the closed world data as eluded to earlier.

2.3 VAST Mini-Challenge Issues

The purpose of the VAST Mini-Challenge is to predict both the ticket sales for the opening weekend and the resulting IMDb user ratings in the U.S. for a specific film. Per the VAST Mini-Challenge rules, the data that is to be used to make the prediction most come from the VAST Mini-Challenge organization. The data is described as being 'closed world' as it consists of only the predetermined data from IMDb, Twitter, and bitly.[15] After registering for the challenge, the links to the data were provided.

The closed world data will prove to be a significant challenge for predicting the opening weekend revenue. As previously stated, there are numerous variables that go into predicting the success of a movie throughout its lifetime or even just the opening weekend. The restriction on data prevents any user in the VAST Mini-Challenge from comparing data or even looking at the variables from a different source. The data that was supplied may not be the most appropriate data for an in-depth analysis of predicting box office revenue for the opening weekend. The data and issues that were associated with the VAST Mini-Challenge are described in a late section.

3 Similar Research

Producing box office forecasts is a generally common analysis of data. It is used by major movie production companies to understand how successful a particular movie will be in terms of movie revenue. As it is safe to assume that Universal Pictures and production companies alike have their own methods to predict success of a film, there are a variety of methods that had been created and utilized in the scientific and educational divisions. One notable paper, *A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures*, was published in 1996.

In a similar study by Sawhney and Eliashberg (1996), a parsimony method is used forecast the national box office revenue per screen. The method involves a two-step processes by analyzing the exposure of the information to an individual and the time it takes that individual to decide to see the movie. "Thus, we model the individual's (stochastic) time to adopt (see) a new movie as the sum of the (stochastic) time to decide to see the new movie; and the (stochastic) time to act on the decision." (Sawhney & Eliashberg, 1996) The model adopts of a series of calculations pertaining to the parameters of time-to-decide and time-to-adopt.[16]

The three weeks' worth of data that was used for the forecasting model was obtained from Variety, which at that time was a leader in trade publications. The data included 19 movies, both successful and unsuccessful, from 1990-1991 and based the box office popularity off of ticket sales rather than attendance. For comparison purposes, Sawhney and Eliashberg analyzed the data with their model (BOXMOD-I) as well as the Bass model (1969) and the Nonlinear Screens Model. Based on the forecasting results, the BOXMOD-I model outperformed both Bass model and Nonlinear Screens Model in predictive errors; 11.23%, 16.06%, and 22.76%, respectively. The data showed that according to the BOXMOD-I model, movies that were heavily advertised resulted in a shorter time for individuals to decide to see the movie as opposed to the movies that relied on word-of-mouth to advertise. The model was also adapted to a zero, one, and two weeks model of data.[17]

The overall findings of Sawhney and Eliashberg were that movies that had notable actors, awareness through sequels, and positive critic reviews increased the potential for good box office revenues. Contrary to that, movies that have an 'R' rating are more likely to have lower revenues and have a higher time-to-act value. Based on the three week data model of BOXMOD-I, the overall box office revenue was fairly accurate when forecasting but was not as accurate when forecasting the revenue throughout the length of time

[12] Eliashberg et al. (April 2009) Evolutionary Approach to the Development of Decision Support Systems in the Movie Industry. *Science Direct (47),* 1. P 1 – 12. Retrieved from http://www.sciencedirect.com.ezproxy2.library.drexel.edu/science/article/pii/S0167923609000025

[13] Eliashberg et al. Ibid.

[14] Eliashberg et al. Ibid.

[15] VAST Mini-Challenge 2013. Retrieved from http://www.boxofficevast.org/vast-welcome.html#

[16] Sawhnet, Mohanbir S. & Eliashberg, Jehoshua. (1996). Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science (15),* 2. P 113-131. Retrieved from http://www.jstor.org.ezproxy2.library.drexel.edu/stable/pdfplus/184189.pdf?acceptTC=true

[17] Mohanbur & Eliashberg, Ibid.

the movie was in theaters.[18] This model was accurate when tested in 1996. In addition to that, another model was used in 2011 to predict box office forecasts.

In a study by Kulkarni, Kannan, and Moe (2011), the use of search engines was used as a medium for predicting the future sales of a product, most notably movie sales. The basis for using search engines was the idea that users will utilize the web servers to seek knowledge on specific attributes of products as well as comparing products to others that are similar. Through the use of their conceptualized model, Kulkarni et al. seek to confirm two questions: "(1) Is the search term volume a good measure of consumer interest? and (2) Does this measure offer good forecasting power?" Because many searches are conducted prior to the release of a movie, the data is expected to predict the future sales of movies.[19]

The model used in this prediction illustrates pre-release and post-release search engine queries stating that both contribute to the interest in the product and the likeliness to purchase the product, respectively. The volume and pattern of the search activity was obtained through a service that compiles data from leading search engines such as Google. The data was then categorized by search data, movie-related variables, and advertising expenditures and then illustrated by using a Weibull hazard process which predicts the probability that a search will be conducted over a specified period of time. Furthermore, the data was used to create four distinguish models for forecasting the sales; "an independent model, a correlated model without movie-specific covariates in the search volume component (but included in the sales component), a correlated model with movie-specific covariates in both the sales and search volume components, and an advertising effects model." (Kulkarni et al., 2011) After analysis, the forecasting methods showed that the pre-launch search of movies within four weeks of its release was a good indicator of the future success in the box office.[20]

Over the years, many forms of research had been conducted to determine which prediction of box office revenue is most effective. Some other notable research in the field has come from Delen and Sharda with a 2006 paper *Predicting Box Office Success of Motion Pictures with Neural Networks*.[21] Two years later they follow suit with another paper titled *Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach* which focused on the use of, "artificial neural networks, decision trees and support vector machines along with information fusion based ensembles."[22] (Delen & Sharda, 2008) Box office forecasting is a unique field of study with various methods attributed to the research. A variation of data may be required to accurately predict revenue.

4 Approach

With the complexity of the movie industry, there is potential for great success as well as great failure. Predicting the potential success of a film includes multiple variables both dependent and independent of each other. Variables such as those mentioned previously (movie locations, contracts, etc.) are independent of the film itself. These types of variables can assist in the success but the gross revenue is not reliant on such factors. However, variables such as which actors are featured in the movie, who the director/writer is, and previous success of those individuals are all possible contributing factors to the success of the film and the revenue that it will create. Actors are the most notable and often the first influencing factor when deciding to see a particular film, thus having a strong influence on success.

Actors tend to have a large influence on the success of a film and the revenue that is expected. Elberse (2007) conducted research based off of the HSX online simulated movie market to and determined that, "The impact of a star on a film's box office revenues positively depends on (a) the star's economic reputation, reflected by his or her historical box office performance, and (b) the star's artistic reputation, reflected by his or her awards or award nomination." (Elberse, 2007). The findings of the research suggest that a movie star that has an average box office history of $100 million is likely to contribute approximately $4 million to the overall revenue of the movie. Additionally, an actor that had previously won an Oscar or Golden Globe award is likely to contribute an additional $400,000 to the gross revenue per award.[23] The success of an actor is strongly correlated with the revenue of a film and is thus important to select an appropriate cast if the film is expected to reach high levels of return.

In a more relative study, one which uses IMDb data such as is required for the VAST Mini-Challenge, Nelson and Glotfelty (2009) conducted a study over nine countries and concluded that a single "top star" can contribute over $5 million to revenue if the variables of production budget and the number of movie screens that it is played on is contributed in the model approach. Having multiple top actors has the potential to increase the revenue of $49-$79 million.[24] *Oz the Great and Powerful* has five popular actors starring in the film. The popular actors that are casted in this film are James Franco, Mila Kunis, Rachel Weisz, Michelle Williams, and Zach Braff.[25] With the VAST Mini-Challenge, we took many variables into consideration such as actors histor-

[18] Mohanbur & Eliashberg, Ibid.

[19] Kulkarni, G., et al., (February 2012). Using Online Search Data to Forecast New Product Sales. *Science Direct (52)*, 3. P 606-611. Retrieved from http://www.sciencedirect.com.ezproxy2.library.drexel.edu/science/article/pii/S0167923611001977#

[20] Kulkarni, G., et al., Ibid.

[21] Delen, Dursun & Sharda, Ramesh. (February, 2006). Predicting Box-office Success of Motion Pictures with Neural Networks. *Science Direct (30)*, 2. Retrieved from http://www.sciencedirect.com.ezproxy2.library.drexel.edu/science/article/pii/S0957417405001399

[22] Delen, Dursun & Sharda, Ramesh. (2008). Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach. Retrieved from http://www.mmo.org.tr/resimler/dosya_ekler/c5b45ddb3ff1f62_ek.pdf

[23] Elberse, Anita. (2007). The Power of Stars: Do Star Actors Drive the Success of Movies? *Journal of Marketing (71)*, 4. Retrieved from http://www.jstor.org.ezproxy2.library.drexel.edu/stable/30164000?seq=11

[24] Nelson, Randy A. & Glotfelty, Robert. (9 July 2009). Movie Stars and Box Office Revenues: An empirical analysis. *Journal of Cultural Economics (36)*. P 141-166. Retrieved from http://download.springer.com.ezproxy2.library.drexel.edu/static/pdf/478/art%253A10.1007%252Fs10824-012-9159-5.pdf?auth66=1363853731_80ecf93f267d5dbb7f213207ac7d6cf0&ext=.pdf5.pdf?auth66=1363853731_80ecf93f267d5dbb7f213207ac7d6cf0&ext=.pdf

[25] Oz the Great and Powerful. Retrieved from http://www.imdb.com/title/tt1623205/

ical success, directors/writers success, film popularity, user ratings, and award nominations/winnings.

Textual analysis of twitter activity relating to the movie also helped in determining general attitudes regarding its release. We also created visualizations of the number of tweets and re-tweets to look for any changes in popularity on the social network. The following section describes the data visualizations and further expands on the approach used for *Oz the Great and Powerful*.

5 DATA

The VAST Mini-Challenge required that only the data provided by the challenge organizers could be used to create interactive visualizations that will help us predict Oz the Great and Powerful movie's opening weekend (Friday through Sunday) box office take and viewer rating. The data provided was IMDb and Twitter data, where the IMDb data was a link to the IMDb database, and the Twitter data was an XML document with the IDs of several hundred recent relevant tweets. Bitly data was not forthcoming.

5.1 IMDb

The Internet Movie Database currently lists the budget for *Oz the Great and Powerful* (2013) at $200 Million, with a user rating of 7.2 from 1,483 votes. IMDb also offers MOVIEmeter rankings which provide a snapshot of a movie's popularity based on millions of IMDb users' searches. The MOVIEmeter lists it as the 14th most popular movie.

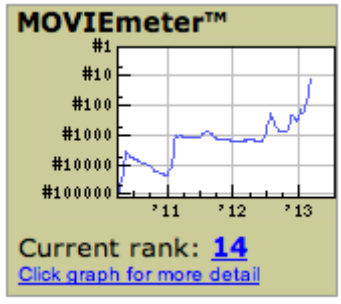

Figure 1: Oz the Great and Powerful MOVIEmeter ranking

Ratings Breakdown

Demographics		
• All votes	7.2	1,660
Males	6.9	641
Females	7.1	169
• Aged under 18	9.1	52
Males	8.5	33
Females	9.9	19
• Aged 18-29	7.3	456
Males	7.3	369
Females	8.2	87
• Aged 30-44	6.3	220
Males	6.5	181
Females	6.0	39
• Aged 45+	6.6	70
Males	6.3	50
Females	8.9	20
• Top 1000 voters	6.9	31
• US users	6.9	303
• Non-US users	6.9	507

All votes		
10	50.2%	833
9	8.4%	140
8	11.7%	195
7	9.9%	165
6	4.8%	79
5	4.6%	77
4	1.7%	28
3	1.8%	30
2	1.1%	18
1	5.7%	95

Arithmetic mean = 8.1
Median = 10

Average Rating[1]
7.2
Total Votes
1,660

Figure 2: Detailed view of Oz the Great and Poweful MOVIEmeter ranking

IMDb was also able to provide data on the key players in regards to awards and nominations. All except for one writer have a fair number of awards and nomination. Rachel Weisz is the only actor on the crew to have won an Oscar, yet Rachel's star rating is 104 while Franco and Kunis' ratings are below 50 (See Appendix 1: Oz the Great and Powerful Credits).

The only available data regarding IMDb user's ratings of the film is either broken down into demographics or total number/percentage of votes per score rating. It would have been helpful to see the number of votes and total score over time as is represented in the people's STARmeter ratings.

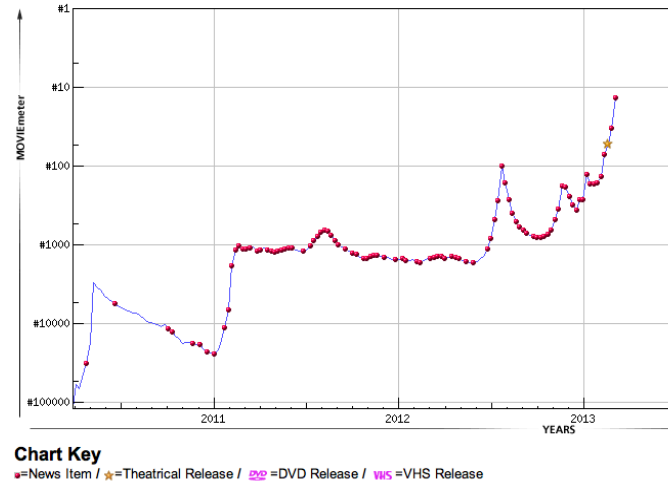

Figure 3: Oz the Great and Powerful User Ratings

We pulled the movies that the key players have been involved with along with the year, budget, opening weekend take, and user rating. Before this movie the Director had only worked with one key actor. Only two actors had previous experience working together. Interestingly enough this was Mila Kunis and James Franco who both starred in the movie *Ted* which is the only movie in our data pool where the opening take exceeded the budget. The writers had not previously worked together or with the director, nor any of the actors (See Appendix 2: Related Movie Data).

We were also able to collect data that shows the shift in popularity ratings of the people involved with the film over time as visualized in Figure 3.

5.2 Twitter

Supplementary data was provided via a list of specific tweet IDs ranging from February 22nd through March 6th. The content of approximately 1,800 unique tweet IDs was pulled using a Python script that called the Twitter API. Due to Twitter's limit of 150 tweets per hour, this script was scheduled to call the API hourly until it had looped through all the inputted IDs.

```
Python Script

'''
Created on Mar 4, 2013
```

```
@author: chase
'''
import requests
import xml.etree.ElementTree as ET
import csv
import glob
from time import sleep
from datetime import datetime

fileInputArray = glob.glob("data/*.xml")

idList = list()

sizeOfList = 0

for nextFile in fileInputArray:
    idsToParse = ET.parse(nextFile)
    idRoot = idsToParse.getroot()

    for nextId1 in idRoot.iter('ID'):
        idList.append(nextId1.text)
        sizeOfList = sizeOfList + 1

with open('output.csv', 'a') as f:
    counter = 0

    processedThusFar = 0

    for nextId in idList:

        if counter < 149:
            try:
                r           =           re-
quests.get('https://api.twitter.com/1/statu
ses/show.xml?id='      +      nextId      +
'&include_entities=true')
                rawMessage = r.content

                root                        =
ET.fromstring(rawMessage)

                created_at                  =
root.find('created_at').text
                retweet_count               =
root.find('retweet_count').text
                text                        =
root.find('text').text

                lineToAdd  =  [text,  creat-
ed_at, retweet_count, nextId]

                csvwriter   =   csv.writer(f,
delimiter=',')
                csvwrit-
er.writerow(lineToAdd)

                processedThusFar    =    pro-
cessedThusFar + 1
            except:
                print   "ecnountered   excep-
tion"

            counter = counter + 1
        else:
            print str(datetime.now()) + " -
sleeping for just over an hour. Processed "
+ str(processedThusFar) + " out of " +
str(sizeOfList) + " records."
            sleep(3900)
            print str(datetime.now()) + " -
resuming execution"

            counter = 0

    f.close()

'''
https://api.twitter.com/1/statuses/show.jso
n?id=112652479837110273&include_entities=tr
ue
'''
```

The Output file (a subset of which is displayed below) was uploaded to IBM's Many Eyes software to run text analysis against the content of the tweets[26] (See Appendix 3: Twitter Data).

Based on visual analysis, the content of the tweets were generally positive. A World Cloud that displays a popular set of words, See, Kunis, and Theodora, indicates that the public is most anticipating "see"-ing "Kunis' performance as Theodora. Kunis, with a high STARmeter ranking of 33, is certainly a cause to predict movie success. Franco's name appears boldly within the cloud. A Word Phrase Net visualization indicates that the public feels that *Oz the Great and Powerful* is… incredible, officially, showing, great, coming, killing. This leads to the conclusion that this movie is highly anticipated and can be expected to do well opening weekend.

6 VISUALIZATIONS

6.1 Actors' Success

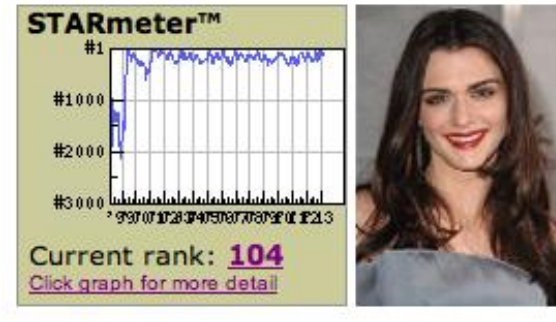

Figure 4: Oscar & Golden Globe winner Rachel Weisz's STARmeter

[26] "Many Eyes." IBM Research. Web. 13 Jan 2013. <http://www-958.ibm.com/software/analytics/manyeyes/>.

Figure 5: Golden Globe winner James Franco's STARmeter Rating

Figure 6: Golden Globe winner Michelle Williams' STARmeter Rating

Figure 7: Golden Globe nominee Mila Kunis' STARmeter Rating

Figure 8: Golden Globe nominee Zach Braff's STARmeter Rating

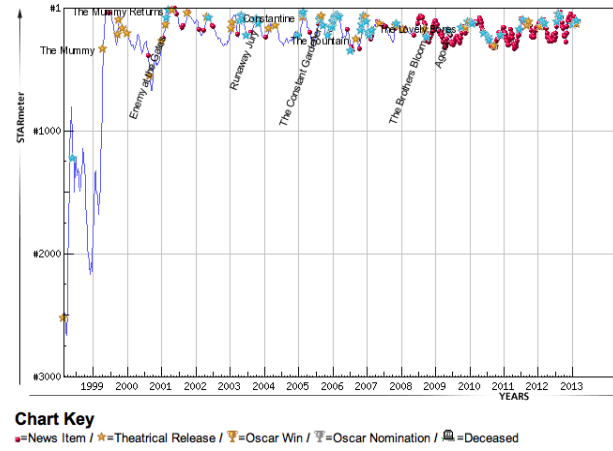

Figure 9: Detail of Rachel Weisz's STARmeter Rating

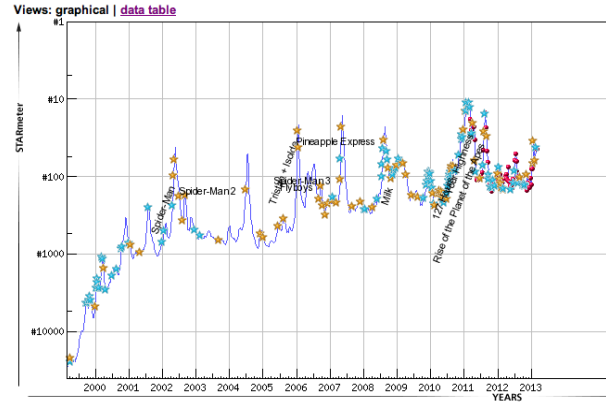

Figure 10: Detail of James Franco's STARmeter Rating

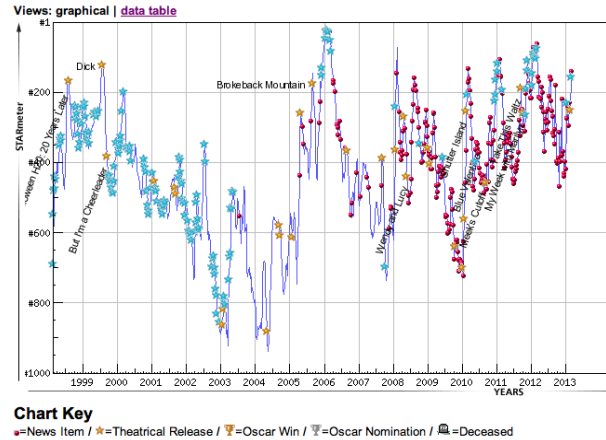

Figure 11: Detail of Michelle Williams' STARmeter Rating

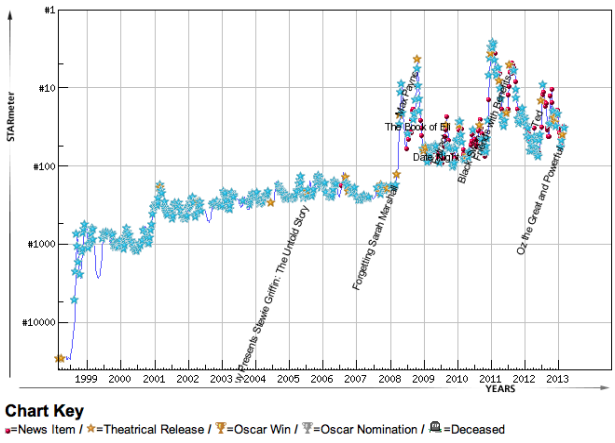

Figure 12 Detail of Mila Kunis' STARmeter Rating

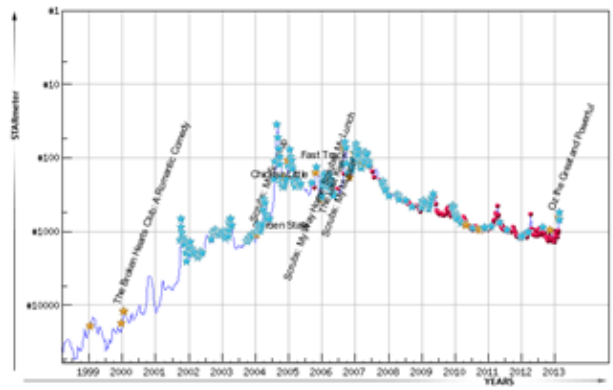

Figure 13: Detail of Zach Braff's STARmeter Rating

6.2 IMDb

Reference Appendix 4: IMDb Visualizations

6.3 Twitter

Reference Appendix 5: Twitter Visualizations

7 Conclusion

The idea of predicting a movie's success based on visualizations is an interesting idea. While the information provided by IMDb gave us good footing, the twitter data was lacking in depth. We question the reason the challenge organizers allow for full database access of IMDb but only a limited subset of twitter data? If we were able to use the entire twitter database, we could have looked for overall popularity and trending modes for each of the actors involved. The challenge website states that bit.ly data would be provided, but this was not the case. What would be the purpose of data provided by a URL shortening service? The number of times the movies home page was accessed via the service? A better partner would be Google, Facebook, Delicious, or Reddit. Facebook shows 762,419 likes and 216,900 talking about the movie. This data could have been compared to other movies with the actors and cross referenced with budgets and user ratings.

The average percentage of the budget that the films analyzed took in on opening weekend was 33.8%. The median percentage of box office take for all films analyzed was 26.6%. If we take these two numbers and multiply them against the budget for Oz, we would get a range of $53.3 - $67.6 million. With taking into consideration of previous research, using Elberse's (2007) approach would result in $45 million (5 awards * $4 million + 5 stars * $5 million).

Based off of out data and research, we predicted that the movie would be successful. After the opening weekend, IMDb released the data for *Oz the Great and Powerful*. The MOVIEmeter rank when to number one with a decrease in user rating from 7.5 to 7.1. Each actor, actress, and the director had an increase in their STARmeter (as of March 10, 2013) which ranged from 2-156 (Kunis, 2; Franco, 3; Weisz, 5; Williams, 23; Braff, 139; and Raimi, 156). However, despite our data, the movie actually took in $80.3 million on the opening weekend.

Research had shown the complexities of predicting a movie's success. It includes many variables, more than measure tweets and the IMDb database. With the data we were provided by the VAST Mini-Challenge, we were unable to accurately predict what the opening weekend revenue would be for *Oz the Great and Powerful*. While being able to conclude a successful opening weekend for *Oz* it is difficult to pinpoint a specific dollar amount with the limited data that was supplied

References

[1] Kohlhammer J., Keim, D., Pohl, M., Santucci, G., & Andrienko, G. (2011). Solving Problems with Visual Analytics. *Procedia Computer Science (7)*. Retrieved from http://bib.dbvis.de/uploadedFiles/373.pdf

[2] Keim, D., Andrienko, G., Fekete, J., Gorg, C., Kohlhammer, J., & Melancon, G.. (2008). Visual Analytics: Definition, Process, and Challenges. *Information* Visualization. p 154-175. Retrieved from http://hal-lirmm.ccsd.cnrs.fr/docs/00/27/27/79/PDF/VAChapter_final.pdf

[3] VAST Mini-Challenge 2013. Retrieved from http://www.boxofficevast.org/vast-welcome.html

[4] VAST Mini-Challenge 2013. Retrieved from http://www.boxofficevast.org/vast-welcome.html#

[5] Pak, C.W., Shen, H.W., Johnson, C.R., Chen, C., & Ross, R. (2012). The Top 10 Challenges in Extreme-Scale Visual Analytics. *IEEE*. Retrieved from thttp://ieeexplore.ieee.org.ezproxy2.library.drexel.edu/stamp/stamp.jsp?tp=&arnumber=6265057

[6] Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale. (October 2007). Retrieved from http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Doe_visualization_report_2007.pdf

[7] Theatrical Market Statistics 2011. (2011). Retrieved from http://www.mpaa.org/Resources/5bec4ac9-a95e-443b-987b-bff6fb5455a9.pdf

[8] Eliashberg, J., Swami, S., Weinberg, C.B., & Wierenga, B. (April 2009) Evolutionary Approach to the Development of Decision Support Systems in the Movie In-

dustry. *Science Direct (47)*, 1. P 1 – 12. Retrieved from http://www.sciencedirect.com.ezproxy2.library.drexel.edu/science/article/pii/S0167923609000025

[9] Sawhnet, Mohanbir S. & Eliashberg, Jehoshua. (1996). Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science (15)*, 2. P 113-131. Retrieved from http://www.jstor.org.ezproxy2.library.drexel.edu/stable/pdfplus/184189.pdf?acceptTC=true

[10] Kulkarni, G., Kannan, P.K., & Moe, W. (February 2012). Using Online Search Data to Forecast New Product Sales. *Science Direct (52)*, 3. P 606-611. Retrieved from http://www.sciencedirect.com.ezproxy2.library.drexel.edu/science/article/pii/S0167923611001977#

[11] Delen, Dursun & Sharda, Ramesh. (2008). Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach. Retrieved from http://www.mmo.org.tr/resimler/dosya_ekler/c5b45ddb3ff1f62_ek.pdf

[12] Delen, Dursun & Sharda, Ramesh. (February, 2006). Predicting Box-office Success of Motion Pictures with Neural Networks. *Science Direct (30)*, 2. Retrieved from http://www.sciencedirect.com.ezproxy2.library.drexel.edu/science/article/pii/S0957417405001399

[13] Elberse, Anita. (2007). The Power of Stars: Do Star Actors Drive the Success of Movies? *Journal of Marketing (71)*, 4. Retrieved from http://www.jstor.org.ezproxy2.library.drexel.edu/stable/30164000?seq=11

ACKNOWLEDGEMENTS

We certify that: To the best of our knowledge, this assignment is entirely work produced by muse. Any identification of our group work is accurate. We have not quoted the words of any other person from a printed source or a website without indicating what has been quoted and providing an appropriate citation. We have not submitted any of the material in this document to satisfy the requirements of any other course.

- Travis Eckenroth, Drexel University, tje38@drexel.edu
- Troy Finamore, Drexel University, twf23@drexel.edu
- Gretel Miller, Drexel University, ges43@drexel.edu

Appendix 1

OZ THE GREAT AND POWERFUL CREDITS

Name	Role	Awards	STARmeter Rating
Sam Raimi	Director	9 wins & 13 nominations	715
Mitchell Kapner	Writer	no listed nominations or awards	16,064
David Lindsay-Abaire	Writer	3 nominations	13,605
James Franco	Actor	**Won, Golden Globe**, *Best Performance by an Actor in a Mini-Series or a Motion Picture Made for Television* for James Dean (2001) Nominated, Oscar, *Best Performance by an Actor in a Leading Role* for 127 Hours (2010) Nominated, Golden Globe, *Best Performance by an Actor in a Motion Picture - Drama* for 127 Hours (2010) Nominated, Golden Globe, *Best Performance by an Actor in a Motion Picture - Comedy or Musical* for Pineapple Express (2008) Another 15 wins & 35 nominations	44
Mila Kunis	Actor	Nominated, Golden Globe. *Best Performance by an Actress in a Supporting Role in a Motion Picture* for Black Swan (2010) Another 6 wins & 31 nominations	33
Rachel Weisz	Actor	**Won, Oscar,** *Best Performance by an Actress in a Supporting Role* for The Constant Gardener (2006) **Won, Golden Globe,** *Best Performance by an Actress in a Supporting Role in a Motion Picture* for The Constant Gardener (2005) Nominated, Golden Globe, *Best Performance by an Actress in a Motion Picture - Drama* for The Deep Blue Sea (2011) Another 9 wins & 19 nominations	104
Michelle Williams	Actor	**Won, Golden Globe**, *Best Performance by an Actress in a Motion Picture - Comedy or Musical* for My Week with Marilyn (2011) Nominated, Oscar, *Best Performance by an Actress in a Leading Role* for My Week with Marilyn (2011) Nominated, Oscar, B*est Performance by an Actress in a Leading Role* for Blue Valentine (2010) Nominated, Oscar, *Best Performance by an Actress in a Supporting Role* for Brokeback Mountain (2005) Nominated, Golden Globe, *Best Performance by an Actress in a Motion Picture - Drama* for Blue Valentine Nominated, Golden Globe, *Best Performance by an Actress in a Supporting Role in a Motion Picture* for Brokeback Mountain Another 28 wins & 45 nominations	140
Zach Braff	Actor	Nominated, Golden Globe, *Best Performance by an Actor in a Television Series - Musical or Comedy* for Scrubs (2001) Nominated, Golden Globe, *Best Performance by an Actor in a Television Series - Musical or Comedy* for Scrubs (2001) Nominated, Golden Globe, *Best Performance by an Actor in a Television Series - Musical or*	542

Comedy for Scrubs (2001) Another 10 wins & 23 nominations

APPENDIX 2

RELATED MOVIE DATA

Movie	Year	Budget	Opening Weekend	User Rating	Director	Writer	Actor
Oz the Great and Powerful	2013	200		7.4	Sam Raimi	Mitchell Kapner David Lindsay-Abaire	James Franco Mila Kunis Rachel Weisz Michelle Williams Zach Braff
The Letter	2012	10		3.3			James Franco
The Bourne Legacy	2012	125	38.142	6.1			Rachel Weisz
Ted	2012	50	54.415	7.1			James Franco Mila Kunis
Rise of the Guardians	2012	145	23.773	7.4		David Lindsay-Abaire	
About Cherry	2012		0.002	4.6			James Franco
Your Highness	2011	49.9	9.36	5.5			James Franco
The Deep Blue Sea	2011		0.124	6.2			Rachel Weisz
Take This Waltz	2011		0.137	6.5			Michelle Williams
Rise of the Planet of the Apes	2011	93	54.806	7.6			James Franco
My Week with Marilyn	2011	9.671	1.75	7			Michelle Williams
Green Hornet	2011	120	33.5	6			James Franco
Friends with Benefits	2011	35	18.622	6.6			Mila Kunis
Dream House	2011	55	8.13	5.8			Rachel Weisz
360	2011		0.012	5.9			Rachel Weisz
The Whistleblower	2010		0.061	7.1			Rachel Weisz
The High Cost of Living	2010			6.3			Zach Braff
Shutter Island	2010	80	41.1	8			Michelle Williams
Rabbit Hole	2010	5	0.054	7		David Lindsay-Abaire	
Meek's Cutoff	2010		0.02	6.5			Michelle Williams
Howl	2010		0.051	6.6			James Franco
Eat Pray Love	2010	60	23.1	5.4			James Franco
Date Night	2010	55	25.2	6.3			James Franco Mila Kunis
Book of Eli	2010	80	38.4	6.8			
Blue Valentine	2010	1	0.194	7.5			Michelle Williams
Black Swan	2010	13	1.44	8.1			Mila Kunis
127 Hours	2010	18	0.265	7.7			James Franco
The Lovely Bones	2009	65	0.117	6.6			Rachel Weisz
Drag Me To Hell	2009	30	15.8	6.7	Sam Raimi		
Agora	2009	70	0.044	7.1			Rachel Weisz
Pineapple Express	2008	27	23.2	7			James Franco
Milk	2008	20	1.45	7.7			James Franco
Inkheart	2008	60	7.6	6		David Lindsay-Abaire	
Definitely, Maybe	2008		11.5	7.1			Rachel Weisz
Days of Wrath	2008	10		6		Mitchell Kapner	
Brothers Bloom	2008	20	0.09	6.8			Rachel Weisz
Spider-man 3	2007	258	151	6.3	Sam Raimi		James Franco
The Last Kiss	2006	20	4.63	6.5			Zach Braff
The Ex	2006		1.39	5.5			Zach Braff

Robots	2005	75	36	6.3		David Lindsay-Abaire	
Chicken Little	2005	150	40	5.8			Zach Braff
The Whole Ten Yards	2004	40	6.69	5.3		Mitchell Kapner	
Spider-man 2	2004	200	116	7.4	Sam Raimi		James Franco
Garden State	2004	2.5	0.201	7.6			Zach Braff
Spider-Man	2002	139	115	7.3	Sam Raimi		James Franco
The Whole Nine Yards	2000	24	15.9	6.7		Mitchell Kapner	
The Gift	2000	10	0.012	6.6	Sam Raimi		
Romeo Must Die	2000	25	18	5.9		Mitchell Kapner	

APPENDIX 3

TWITTER DATA

Status	Date	Re-tweets	ID
http://t.co/ncAeYWoj1F	Thu Feb 28 21:39:01 +0000 2013	0	307243536028418048
Mila Kunis as Theodora, one of the witches!!	Mon Mar 04 08:09:08 +0000 2013	0	308489270023319552
"Check out @MariahCarey's #AlmostHome here: https://t.co/JZuX8U9mZz http://t.co/YnaM7LYaPr" #idol #soundtrack #OzTheGreatAndPowerful	Thu Feb 21 01:30:04 +0000 2013	0	304402577125158912
"A miscast James Franco and a lack of charm and humor doom Sam Raimi's prequel to the 1939 Hollywood classic." #burn #ozthegreatandpowerful	Fri Mar 01 06:23:42 +0000 2013	0	307375576220573696
"A miscast James Franco and a lack of charm and humor doom Sam Raimi's prequel to the 1939 Hollywood classic." @THR #OzTheGreatAndPowerful	Fri Mar 01 05:40:53 +0000 2013	0	307364801074577409
"Are you the great man we've been waiting for?" | #Oz #OzTheGreatAndPowerful http://t.co/rxyIpd1CPF	Wed Feb 27 11:54:35 +0000 2013	0	306734069474025472
Ding dong, the witch is dead," you all sing to get ready for #OzTheGreatAndPowerful. It was only ever about getting her ruby slippers. #SMH	Wed Feb 27 21:15:15 +0000 2013	0	306875166879805440
"I wouldn't dare immunlate something that is beyond iconic" Mila Kunis #OzTheGreatAndPowerful	Fri Mar 01 14:47:59 +0000 2013	1	307502482714664960
"Its great working with great actors cause they know how to make a scene work" Sam Raimi #OzTheGreatAndPowerful	Fri Mar 01 19:48:57 +0000 2013	1	307578223577931777
"Oz the Great and Powerful" http://t.co/p7XQRSFsvZ	Wed Feb 27 10:41:40 +0000 2013	0	306715718777454592
#10WorstFeelings wen u wanted badly to be able to sing like @MariahCarey... Lol. #greatvocals #AlmostHome #OzTheGreatAndPowerful	Thu Feb 21 07:57:14 +0000 2013	0	304500011784368128
#4days #ozthegreatandpowerful #www #wickedwitchofthewest can't wait!! http://t.co/3jyvs4yVhS	Mon Mar 04 05:27:40 +0000 2013	0	308448636122066945
#almosthome #MariahCarey #disney #ozthegreatandpowerful #single #music #instamood http://t.co/Nif0qJMNkQ	Mon Mar 04 09:49:27 +0000 2013	0	308514515820085248

APPENDIX 4

IMDB VISUALIZATIONS

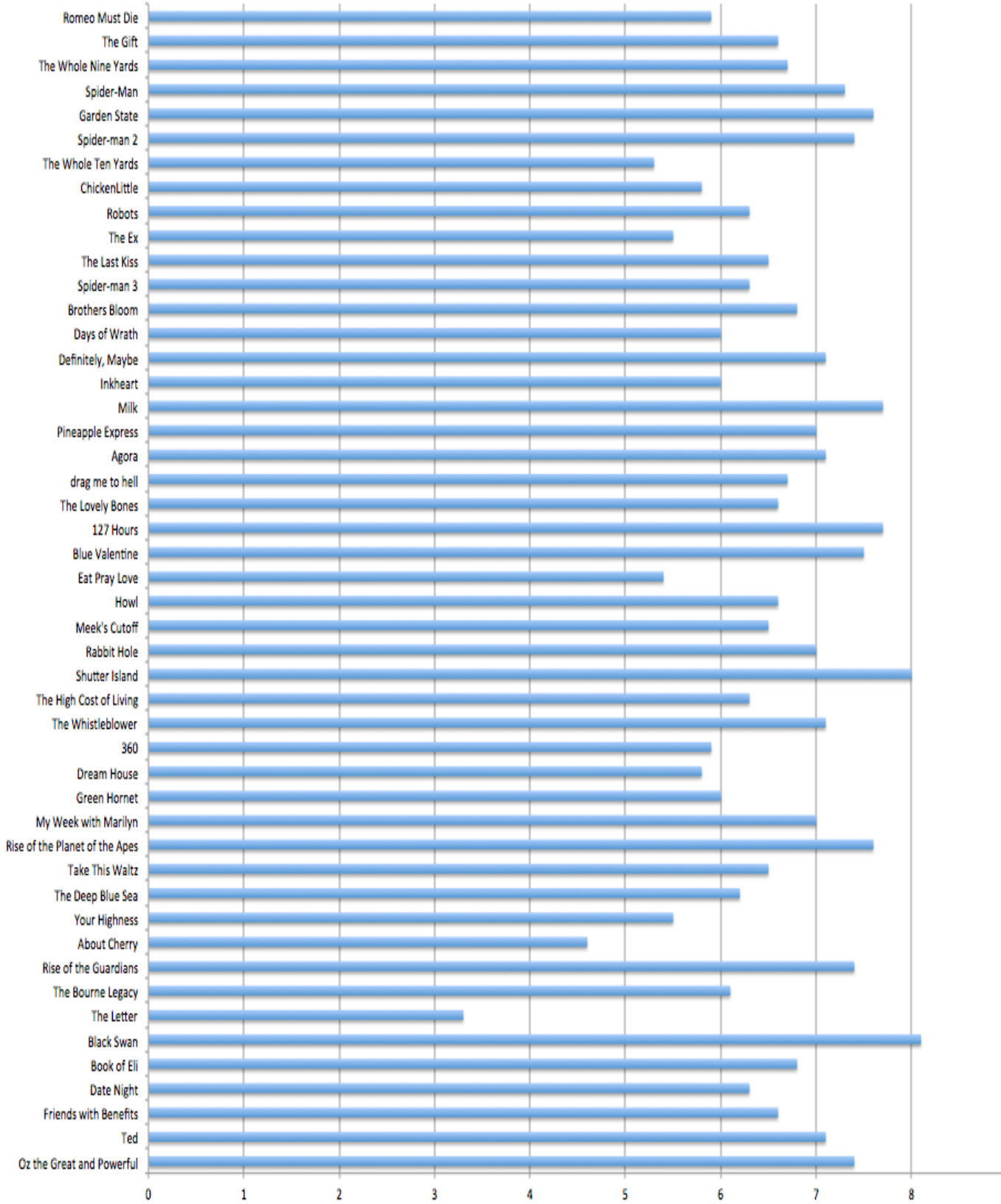

Figure Appendix 4.1: IMDb User Rating of Movies related to Actors, Writers, and Directors

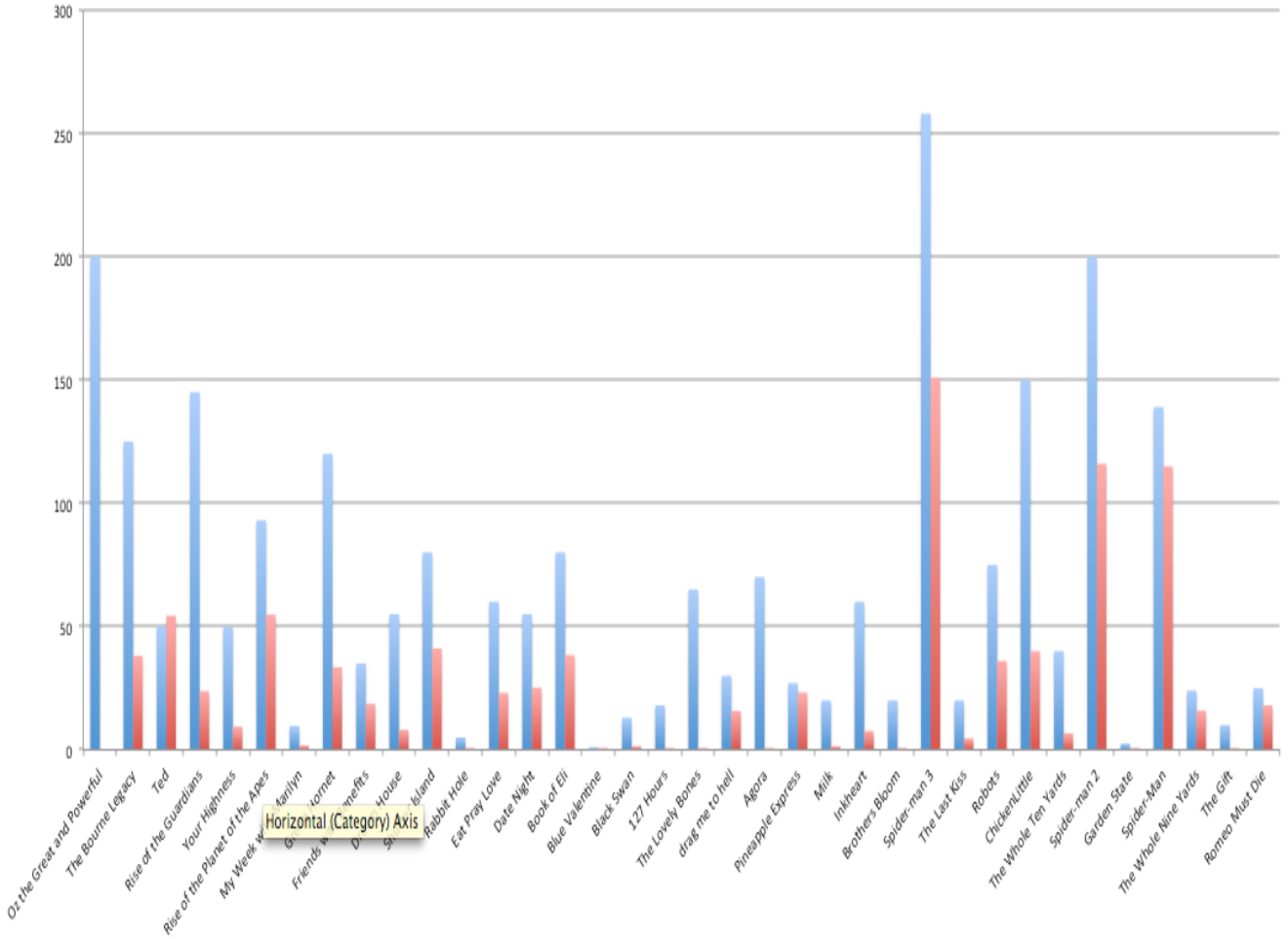

Figure Appendix 4.2: Budget v Opening Weekend for movies related to Actors, Writes, and Directors (Numbers in Millions)

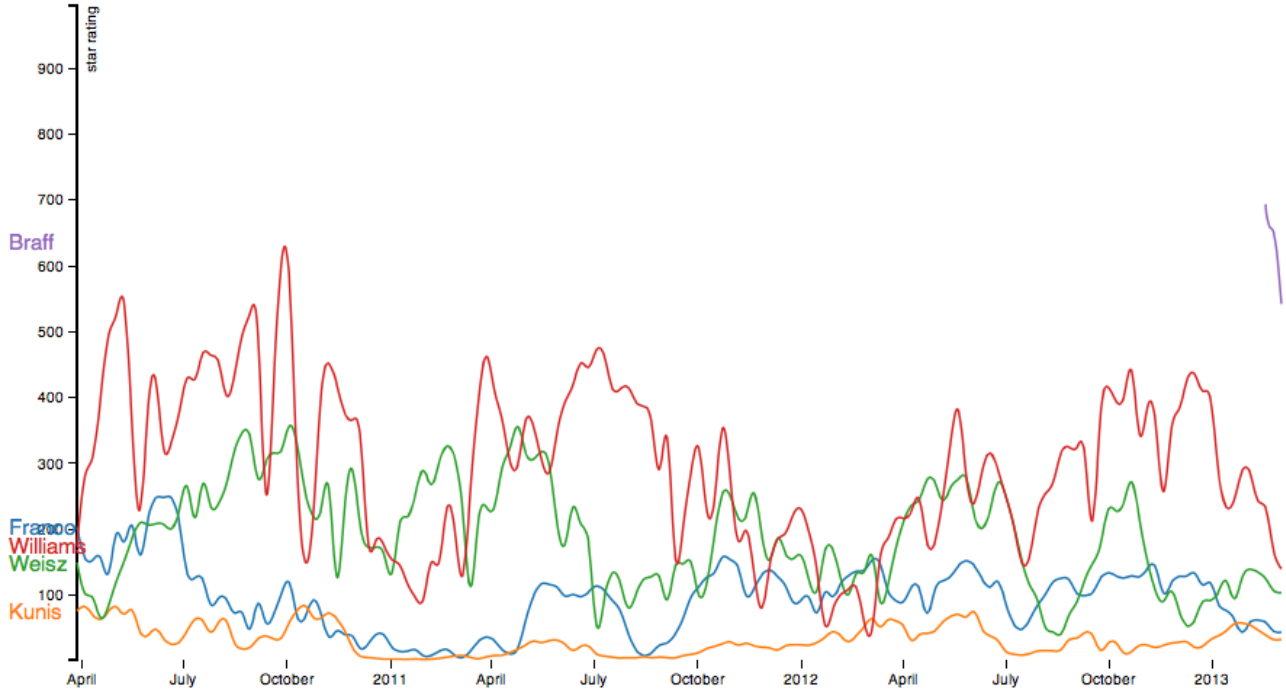

Figure Appendix 4.3: Popularity Rating of Lead Actors on IMDb

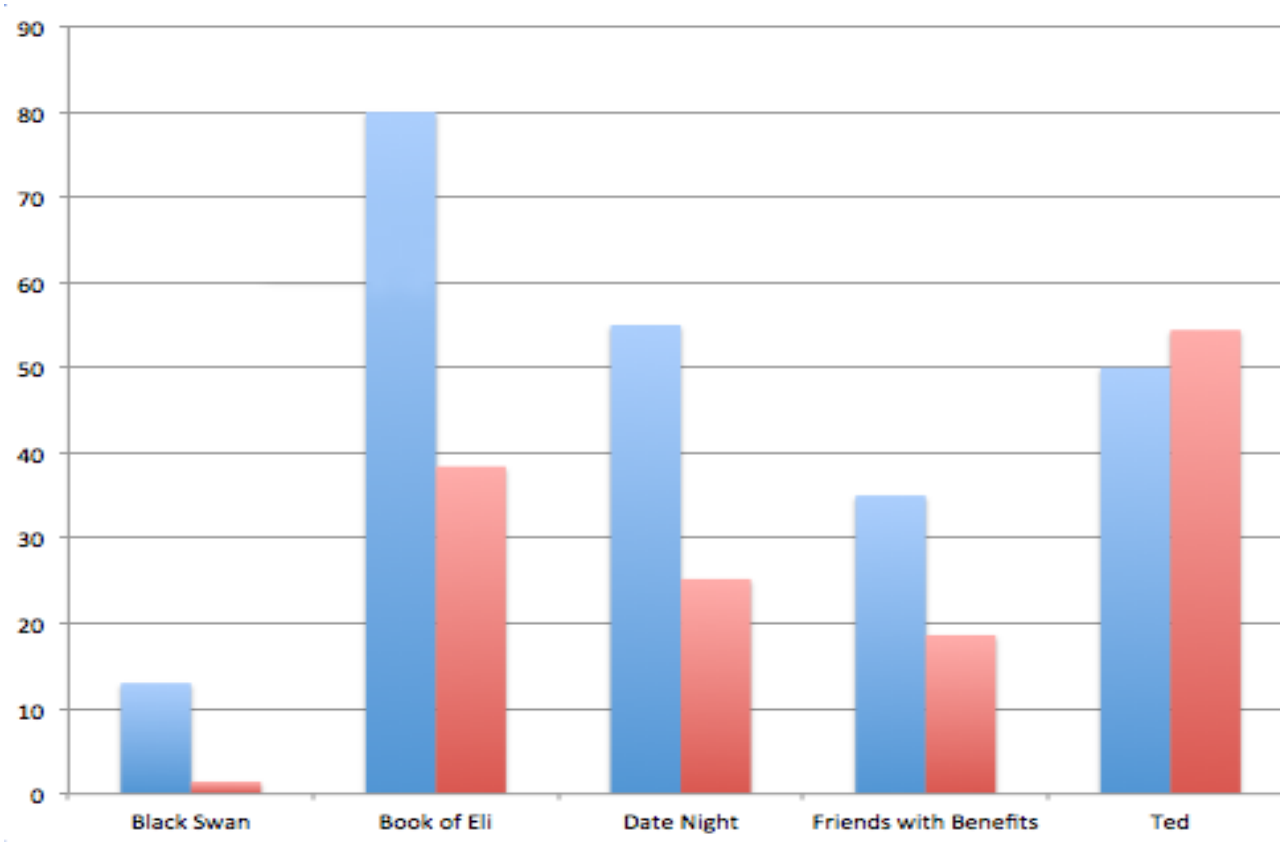

Figure Appendix 4.4: Budget vs. Opening Weekend for movies related to Mila Kunis (Number in Millions)

APPENDIX 5

TWITTER VISUALIZATIONS

Figure Appendix 5.1: Word Cloud based on Twitter data

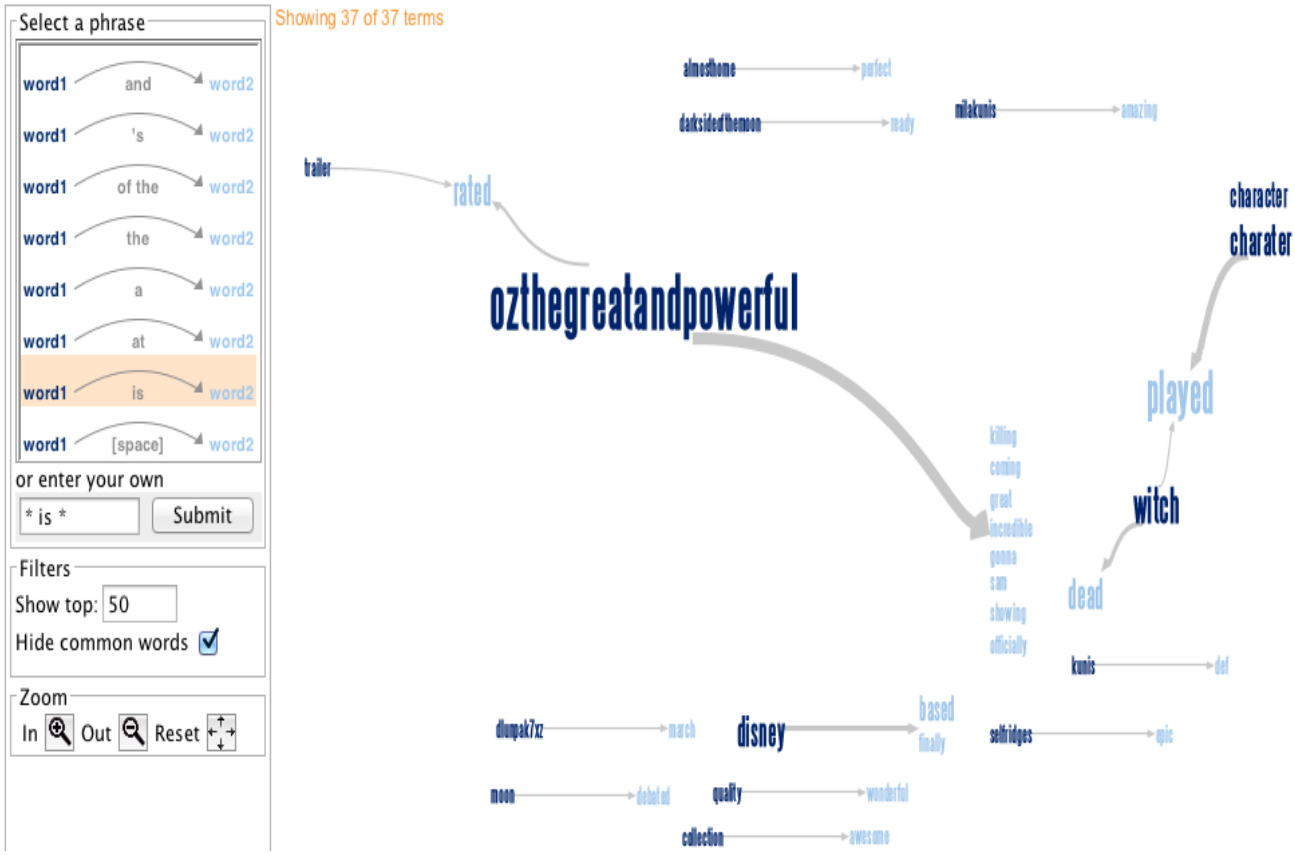

Figure Appendix 5.2: Word Phrase Net based on Twitter data

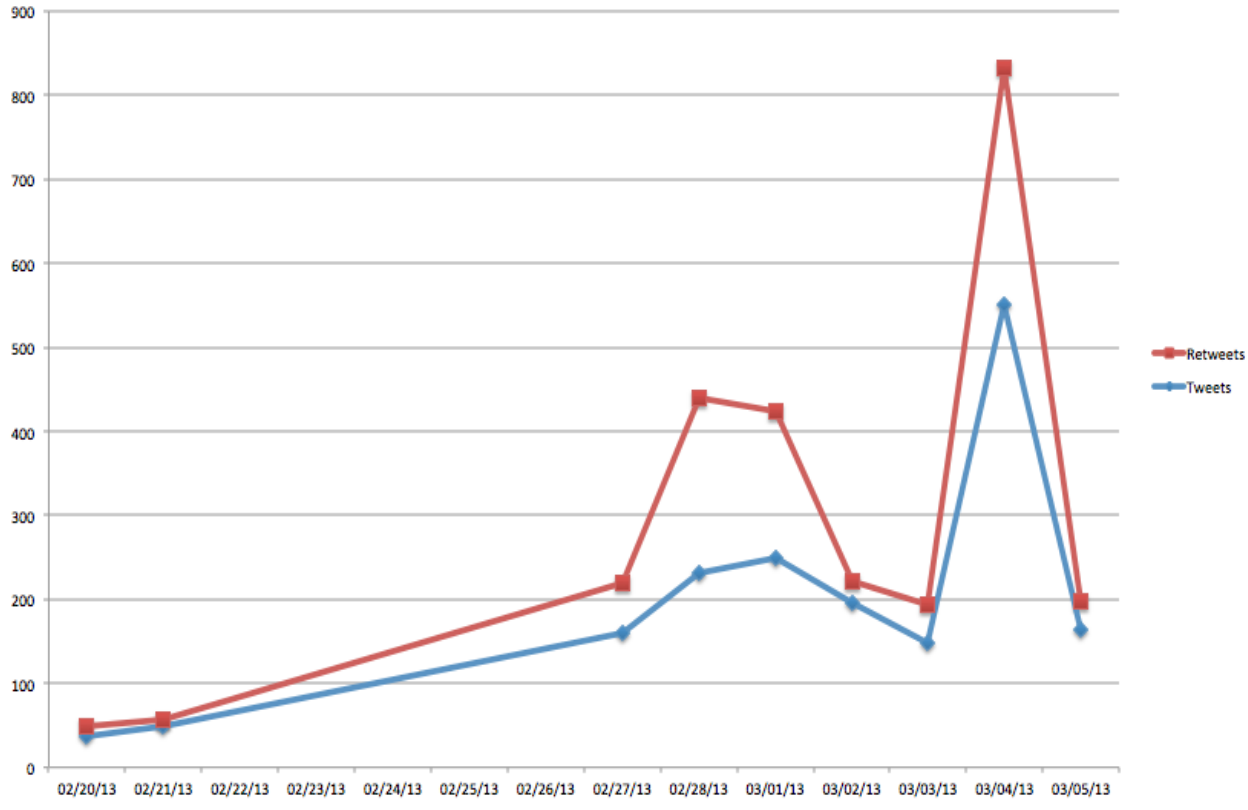

Figure Appendix 5.3: Number of Tweets and Re-Tweets