Chong Zhou (Andy)
Diana Batista
Marcus Moyses
Tabassum Kakar

20 September, 2014

Data Science: Case Study 1

Parts 1 and 2

Motivation

As potential professional data scientists in the near future, our group wanted to investigate a self-serving subject: data science. Not only do our interests lie in the opportunity to learn more about the subject while in an academic setting—where we could possibly acquire more knowledge than our peers and competitors—we are also interested in the potential and benefits this field of study will provide once we have left WPI.

Our research is driven by the following questions:

- Are the concepts we are learning in an academic setting reflective of current Twitter trends? I.e. are we seeing an equal distribution in mentions between computer science, mathematics, and business? If not, does the interest lie in the novelty of the field as opposed to its practical applications?
- Who is interested? What is their background?

While we are fully aware that accurately answering these questions within the scope of this assignment is unreasonable given the data's limitations and nature of being a snapshot in time, it does provide us with a starting point based on science as opposed to a gut feeling grounded in our individual biased timelines.

Data Collection

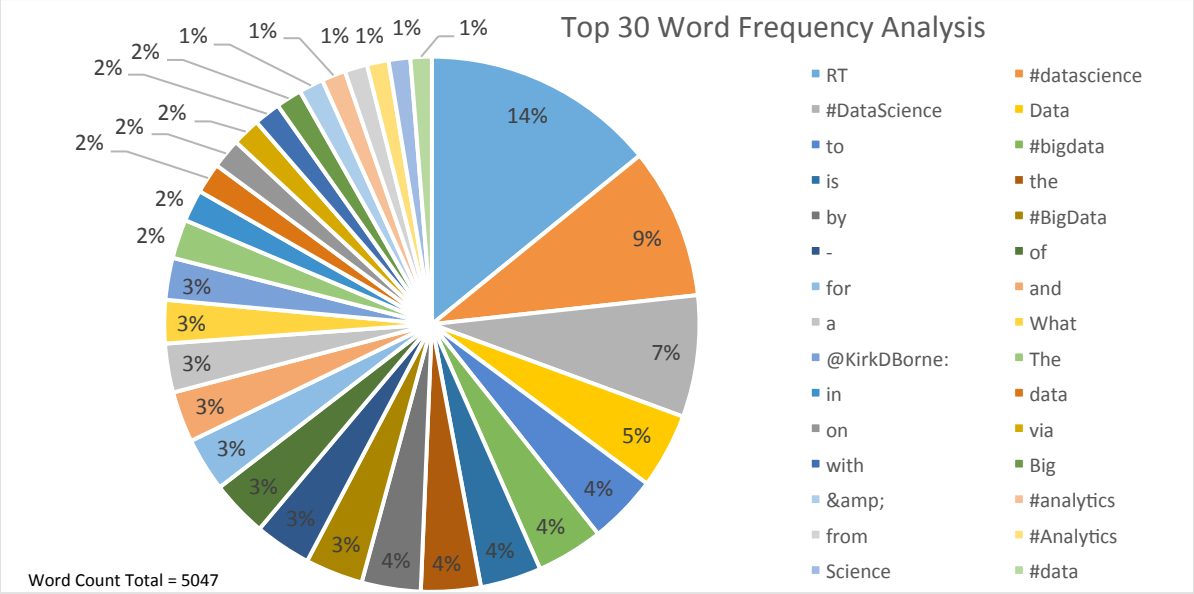
In order to conduct this analysis, real time tweets about Data Science were collected on Sunday, September 7th at 10:15 AM. Our output consisted of approximately 1200 tweets that were saved in a JSON file. Of these 1200, we analyzed the frequency of the top 30 words, the top 10 tweets with the highest retweet counts, and the most popular tweet entities split by hashtags and user mentions.

Method of Analysis and Results

Our method of analysis varied based on the results provided by the query. Where counts were provided, we conducted visualized frequency analysis and where actual tweets were provided, we dug deeper into the information available by researching accounts and links.

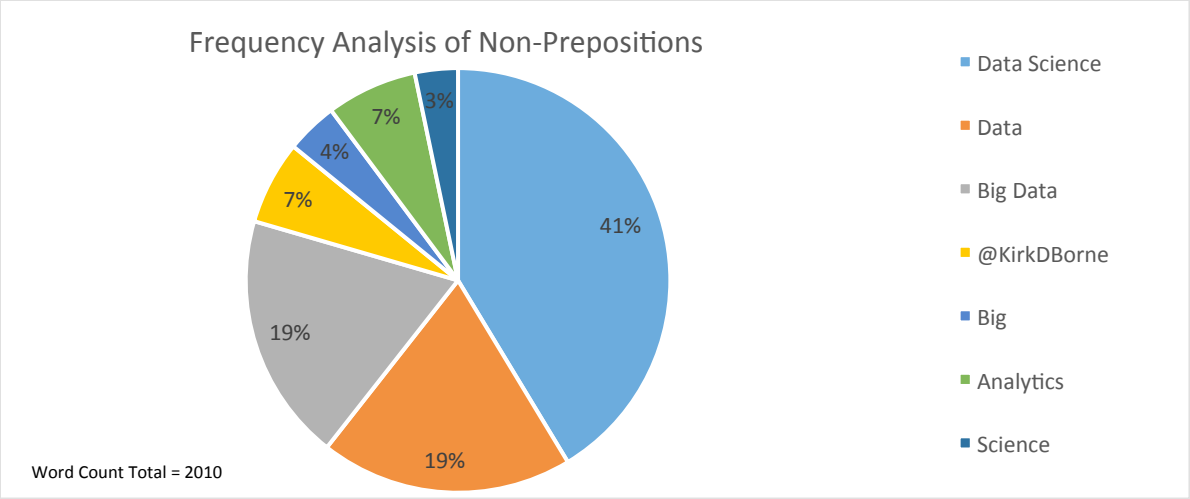
Top 30 Words

Below is a visualization of the results for the top 30 most frequently used words. Please see page 8 for the table output.



Unsurprisingly, the list of the top 30 words includes a significant amount of prepositions that do not provide any additional insight into our study. In order to focus on the frequency of words that would move our analysis forward, we analyzed the subset of non-preposition words and cleaned the results by combining counts that were split by case sensitivity. The subset is as follows:

Word Type	Count	Percentage of Total	Percentage of Subset
Prepositions	3037	60.2%	N/A
Non-Prepositions	2010	39.8%	100.0%
Data Science	831	41.3%	41.3%
Data	387	19.3%	19.3%
Big Data	380	18.9%	18.9%
@KirkDBorne	128	6.4%	6.4%
Big	79	3.9%	3.9%
Analytics	139	6.9%	6.9%
Science	66	3.3%	3.3%



While it is unfortunate that 60% of our results had to be eliminated, this process was necessary in order to render our output useful. Of the useful 40%, "Data Science" was the most common result with 41%, "Data" and "Big Data" were the second and third most common results with 19% each. Although a small percentage of this subset, Twitter user @KirkDBorne comprised 7% of the result. His twitter handle also appears in the list of most popular tweet entities. According to Kirk Borne's Twitter profile, he is a "PhD Data Scientist, Astrophysicist, top #BigData Influencer," with a passion for "Data Science, Data Mining, Astroinformatics, and Citizen Science." Following the link on his twitter profile to his LinkedIn account, we learned that Dr. Borne has been an Astrophysics and Computational Science Professor at George Mason University for the past 11 years. Dr. Borne is also the co-creator of the Data Science BS degree program. His deep passion and involvement in data science makes his twitter handle very likely to show in our research results multiple times.

Most Popular Tweets

Below are the results for the top 10 most popular tweets:

Count	Screen Name	Text
140	TEDTalks	RT @TEDTalks: "We want to tell the story of the 'eureka!' moment. But a lot of important ideas have very long incubation periods." http://t…
78	AnalyticsChap	RT @AnalyticsChap: Ahh, so Data Scientists look like Buddy Holly? But seriously, this is a good infographic #datascience #data http://t.co/…
33	ManeeshJuneja	RT @ManeeshJuneja: #DataScience skillset explained in an infographic http://t.co/M85bqOiJcZ via @MktngDistillery #bigdata http://t.co/UNW2e…
30	JoelGurin	RT @JoelGurin: What exactly is #BigData? See 42 definitions curated by @jndutc. Interesting reading! http://t.co/CKmerwjnVP
27	EricTopol	RT @EricTopol: Ask 40 thought leaders what #bigdata is? .....You get 40 different answers http://t.co/drbSqLeSlb @BerkeleyData http://t.co/…

18	kdnuggets	RT @kdnuggets: How Uber uses #DataScience to predict where its riders want to go #BigData http://t.co/TKRJjtFkK4
17	KirkDBorne	RT @KirkDBorne: The #MachineLearning Clickable Map = choosing the right algorithm for your #DataScience: http://t.co/UJ8blKcwBR http://t.co…
17	AngelaZutavern	RT @AngelaZutavern: What!? You mean #datascience can't predict the future? 3 mistaken assumptions about #bigdata http://t.co/Agt1PBHgJB h…
15	R_Programming	RT @R_Programming: A Full Fledged R Course on Youtube. Absolutely Free! How does that sound? http://t.co/bgJ6LGyrkY #rstats #datascience …
15	KirkDBorne	RT @KirkDBorne: Science's #BigData Problem and The Science of Data [#DataScience] http://t.co/HhrkWPQ2mQ | HT @andrewfogg

We initially questioned the validity of our query based on the top tweet since it has no mention of data science. After searching the results however, we noticed this tweet appeared in our output because the retweeter's twitter handle is @DataScience_SPM. The list of most popular tweets contains those that have the highest retweet counts and do not necessarily appear multiple times in the data retrieved using the API.

The second and third top tweets on the list both reference the same visual titled the "Modern Data Scientist." This infographic addresses the first question driving our motivation to research data science. MarketingDistillery.com, a group of practitioners in the area of e-commerce marketing, states the following: "Data Scientist, the sexiest job of the 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard." The visual can be found at http://t.co/M85bqOiJcZ. This infographic supports the concept we have been taught in DS501, where the ideal data scientist has knowledge of computer science, mathematics, and business.

The fourth and fifth most popular tweets lead to the same link on the Berkeley School of Information website. The site can be found at http://datascience.berkeley.edu/what-is-big-data/. This datascience@Berkeley blog attempts to define big data by asking 40 leaders their opinions on the definition of big data. This further supports the idea that data science is still a developing field of study, so much so that well-established professionals have different opinions of what it means.

The last 5 most popular tweets covers to the following topics:

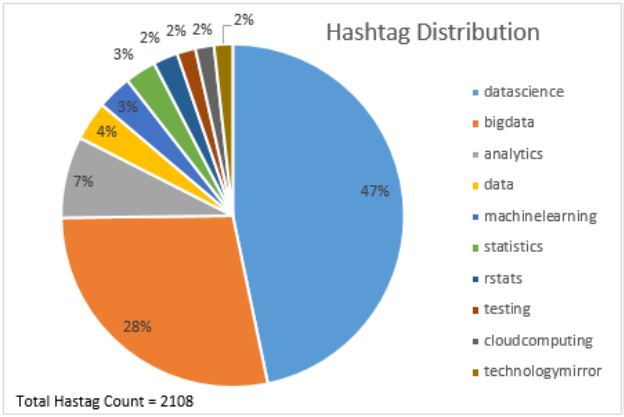
- Uber uses Data Science to predict where riders want to go
- Choosing the right estimator
- 3 mistaken assumptions about what big data can do for you
- 3 way to debug R code
- Data Science's big problem

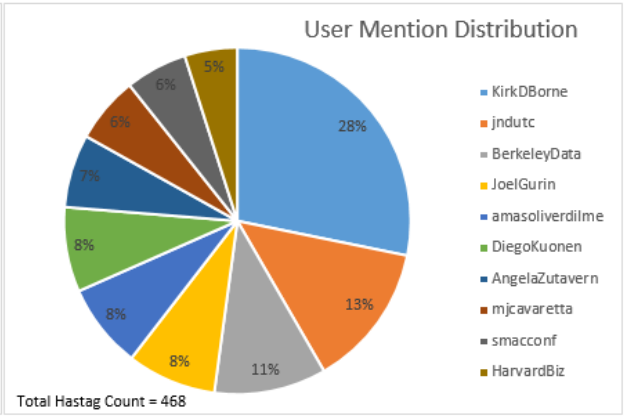
Most Popular Tweet Entities

Below are the results for the top 10 hashtags and user mentions:

Hashtag	Count	Percentage
datascience	985	46.7%
bigdata	592	28.1%
analytics	162	7.7%
data	78	3.7%
machinelearning	69	3.3%
statistics	62	2.9%
rstats	49	2.3%
testing	37	1.8%
cloudcomputing	37	1.8%
technologymirror	37	1.8%
Total	2108	100%

User Mention	Count	Percentage
KirkDBorne	132	28.21%
jndutc	63	13.46%
BerkeleyData	49	10.47%
JoelGurin	39	8.33%
amasoliverdilme	37	7.91%
DiegoKuonen	37	7.91%
AngelaZutavern	32	6.84%
mjcavaretta	29	6.20%
smacconf	27	5.77%
HarvardBiz	23	4.91%
Total	468	100%





The top 10 hashtags and user mentions yield results we have seen before. “Data Science”, “big data”, “analytics”, and “data” compose 86.2% of our top 10 hashtags. The same four words represent 34.2% of the 30 top mentioned words and 86.4% of the subset of non-preposition words in the first frequency analysis. The complement of the hashtag dataset contains topics also relevant to data science such as machine learning, statistics, and cloud computing. These are all subjects that have been addressed in class further validating that Twitter trends about data science are in line with what is being taught at WPI.

Expectedly, Dr. Borne who was mentioned in the first frequency analysis, is the most frequently mentioned user. The second most mentioned user is @jndutc, a social brand manager at Berkeley Data—our third most mentioned user. Together, these two accounts compose almost 25% of the most mentioned users. The fourth on our list,

@JoelGurin, was the chair of the White House task force on consumer data. It is no surprise that these accounts are so highly associated with data science trends on Twitter.

Part 3

Randy Couture is a UFC champion born on June 22, 1963 in Everett, Washington. In addition to being the only person over the age of 40 to have won a UFC championship fight, he also expanded his career into Hollywood. He has acted in The Expendables, Hijacked, and Stretch, among others. Our interest in researching Randy is in part to honor one of our data science professors and partly because of Randy Couture's varied career choices.

The data acquired for this study is of Randy Couture's Twitter friends and followers, with a count of 268 and 360K as of September 20th, 2014.

Below is a list of 20 of Randy's friends and followers:

Friends	
User_id	**Screen_name**
120943272	JimCarrey
186344612	JRsBBQ
34212917	TheLinaCarollo
1165450428	RealDDP
887496402	ALaForce
86868062	ErinAndrews
89556802	TheRue
76856486	cabrasted
1158861247	Followtheblonde
505219769	privatetraining
20019468	jeffshearer3
31293110	dolvett
29255539	Karina_Smirnoff
150151257	CHRIS_Daughtry
111409543	IanZiering
41630638	TheDeliverer_32
237611757	chillzone95
41838544	Uldouz
52551600	HISNHERS_TV

Followers	
User_id	**Screen_name**
2778081592	alden_marano
2603767360	AJCrossman6692
2817756103	airc1225
2780760186	j02010333
2601678608	pota_esxt
1969567669	scottcampione
1244611405	05c799c5d41348f
2755327413	_mo_robbins_
210839313	1MMANEWS
2820488630	JaynaraNeves
2355240157	grahambell91
2822246504	madisontweaverr
2474884813	vegasfantasyvip
494215479	fltruckguy77
267326533	BarcelosCarlito
2822160120	rightcowboy
1849608163	damiancaceres30
1942570939	EarleyGarrett
2821733490	Caggiexforxlife

An intersection of the two datasets yields the following:

Friends and Followers	
User_id	Screen_name
120943272	ErinAndrews
34212917	Karina_Smirnoff

Given Randy's public image, it was surprising that the intersection of the two datasets only yielded two results. Conversely, these results are likely inaccurate because of the Twitter API data limitations. We are unable to acquire all 360K followers to provide an accurate intersection of both friends and followers.

The two results yielded are of Erin Andrews and Karina Smirnoff, fairly well-known television personalities. Deeper research into Randy's most recent career move reveals he is currently cast in Dancing with the Stars. Karina Smirnoff is Randy's dance partner on the show. Erin Andrews is a Fox Sports Broadcaster and Dancing with the Stars co-host. Given this information, it is reasonable that Randy would consider both women friends and they would reciprocate the friendship.

Conclusion

Although the capacity of this study is limited, we were able to gather information that would possibly allow us to answer the questions that drove our interest. The data we acquired at a specific point in time does support the idea that a data scientist is a professional with a computer science, mathematics, and business background. Our data also suggests that data science is a relatively new field where there are multiple views on its definition and possible misunderstanding of its purpose. Predictably, the users who appeared most often in our frequency analysis are those with strong data science backgrounds who may have a vested interested in promoting the field through social network.

Part 1 Frequency Table

Below is the frequency table for the Top 30 word count in our tweets.

Word	Count	Percent
RT	714	14.15%
#datascience	460	9.11%
#DataScience	371	7.35%
Data	229	4.54%
to	211	4.18%
#bigdata	204	4.04%
is	186	3.69%
the	182	3.61%
by	180	3.57%
#BigData	176	3.49%
-	173	3.43%
of	173	3.43%
for	164	3.25%
and	155	3.07%
a	149	2.95%
What	132	2.62%
@KirkDBorne:	128	2.54%
The	119	2.36%
in	97	1.92%
data	93	1.84%
on	91	1.80%
via	86	1.70%
with	81	1.60%
Big	79	1.57%
&	75	1.49%
#analytics	73	1.45%
from	69	1.37%
#Analytics	66	1.31%
Science	66	1.31%
#data	65	1.29%
Total	5047	100.00%