CrunchBase Exploration

Group Members:
Yi Li
Chang Sun
Jingshi Sun

good explorative of database.

no tested hypothesis

well written report — could use minor grammar improv

good graphics

A

Abstract

In this paper, we explore the CrunchBase, which is an online database including information about innovative companies. We try to find out the characteristics that can influence the total funding that a company can get. We also do some explorations using MySQL, Python, Spark and various data visualization tools to further analyze the distributions and the trends of funding, the people, and the overview, which is the description of the companies. We find out that the total funding has more relationships with industries than other influencing factors like regions and people in management levels. We also find that mobile and cloud are popular in today's innovative companies. There are about 27% of the people in these innovative companies have MBA degrees, and higher management level people typically have MBA degrees.

1. Introduction

CrunchBase (www.crunchbase.com), operated by TechCrunch and founded in 2007, is a leading platform to discover innovative companies and the people behind them. CrunchBase is an open database and registered users can make submissions to the database, subject to reviewing and accepting by the moderator (Wikipedia contributors, TechCrunch). There are editors to ensure the database is up to date. There are 4.8 million community edits, 324,000 contributors, 3,100 venture partners, and 1.2 million web visitors each week on the CrunchBase (about.crunchbase.com/company/). Most of the CrunchBase revenue comes from advertising and licensing of data that provides advanced access to its database.

As innovative companies are becoming such a commonplace in today's world, we want to use this database to explore those companies. What attract us most are finding what characteristics like region, industries and people would lead to more funds for companies, as well as getting insights of what these companies are doing. Innovative companies are leaders in developing new technology and driving economic growth, so this analysis would be extremely relevant and interesting.

2. Literature Review

(Liang and Soe-Tsyr 2016) build a social network graph by analyzing data from CrunchBase, which takes people, companies, social links, and funding investment links into consideration. They find out that funding investors are more willing to make investments if the company has a strong social relationship and has less common neighbors.

(Xiang 2012) uses TechCrunch and CrunchBase (TechCrunch's open database) to extract topic features using methods like LDA based on factual features like basic features, financial features, and managerial features to predict M&A behaviors.

(Huang 2008) discusses several methods to compare text similarities such as Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient etc. It also recommends K-

means clustering for dealing with large size text document because it is a fast algorithm with relatively small time complexity.

3. Data

We use the 2013 Snapshot (data.crunchbase.com/docs/2013-snapshot), which contains a replica of the CrunchBase dataset before December 2013. It includes all organizations, people, and product profiles along with funding rounds, acquisitions, and IPOs. There are totally 11 datasets in SQL dump format, in this research, we only used five of them. The name, the number of rows and the size of each dataset are shown below in Table 1.

Table 1. Datasets name, rows, and size

Dataset Name	Number of Rows	Size
cb_objects	462651	339M
cb_funding_rounds	52928	14.7M
cb_people	226709	13.3M
cb_degrees	109610	13.8M
cb_relationships	402878	48.7M

Since our data is in the SQL dump format, we ingested the data into MySQL. We identified some important features that are most relevant to our exploration, and then output two datasets into text format: One is the dataset "companies" that comes from "cb_objects", which contains basic information about companies, and one sample record is shown in Table 2 below. The other is the dataset "people", which is joined by "cb_relationships", "cb_people", and "cb_degrees". The sample records are shown in Table 3.

We found that there are three main problems in the dataset "companies", which are missing values, inconsistent values, and duplicate values. To get a more accurate result, we removed null values. We decided to focus on the companies founded after 2000 and companies with total funding amounts of more than one million dollars (see more detail in section 5.2). We choose the companies founded after 2000, because we are trying to focus on relative new companies that can take advantage of the advance techniques. New technology has been rapidly developing after 1990s, and it takes some time for the new technology to improve and be applied in the real world. The range of the total funding is extremely large from $291 dollars to $5.7 billion dollars with the median funding of $3 million dollars and the average funding of $16.14 million dollars. So, we decide to cut off some small companies, and focus on companies with total funding amounts more than one million dollars.

Table 2. A sample record of the dataset "companies"

id	c:1
name	Wetpaint
normalized_name	wetpaint
category_code	web
status	operating
founded_at	10/17/05
closed_at	NULL
short_description	NULL
description	Technology Platform Company
tag_list	wiki, media-industry, media-platform, social-distribution-system
country_code	USA
state_code	WA
city	Seattle
region	Seattle
first_funding_at	10/1/05
last_funding_at	5/19/08
funding_rounds	3
funding_total_usd	39750000
overview	Wetpaint is a technology platform company that uses its proprietary state-of-the-art technology and expertise in social media to build and monetize audiences for digital publishers. Wetpaint?s own online property, Wetpaint Entertainment, an entertainment news site that attracts more than 12 million unique visitors monthly and has over 2 million Facebook fans, is a proof point to the company?s success in building and engaging audiences. Media companies can license Wetpaint?s platform which includes a dynamic playbook tailored to their individual needs and comprehensive training. Founded by Internet pioneer Ben Elowitz, and with offices in New York and Seattle, Wetpaint is backed by Accel Partners, the investors behind Facebook.

Table 3. Sample records of the dataset "people"

object_id	relationship_object_id	first_name	last_name	affiliation_name	title	degree_type	subject	institution	graduated_at
p:2	c:1	Ben	Elowitz	Blue Nile	Co-Founder/CEO/Board of Directors	BS	Electrical Engineering/Computer Science	University of California, Berkeley	1/1/94
p:2	c:1	Ben	Elowitz	Blue Nile	Co-Founder/CEO/Board of Directors	BS	Applied Mathematics	University of California, Berkeley	1/1/94
p:3	c:1	Kevin	Flaherty	Wetpaint	VP Marketing	BBA	NULL	Washington University in St Louis	NULL
p:3	c:1	Kevin	Flaherty	Wetpaint	VP Marketing	MBA	NULL	Indiana University	NULL
p:5	c:3	Ian	Wenig	Zoho	Senior Director Strategic Alliances	Degree	Advanced Business Professional Course	The Aji Network	1/1/05
p:5	c:3	Ian	Wenig	Zoho	Senior Director Strategic Alliances	BS	Biology, Psychology	McGill University	1/1/86
p:7	c:4	Jay	Adelson	Digg	Chief Executive Officer		Film and Broadcasting, Computer Science	Boston University	1/1/92
p:8	c:4	Owen	Byrne	Digg	Senior Software Engineer	BS	Computer Science	Saint Mary's (Canada)	1/1/86
p:8	c:4	Owen	Byrne	Digg	Senior Software Engineer	MBA	Business	Dalhousie (Canada)	1/1/94
p:9	c:4	Ron	Gorodetzky	Digg	Systems Engineering Manager		Computer Engineering	University of California, San Diego	1/1/03

There are also some problems with the dataset "people". The "relationship object id" comes from the dataset "cb_relationship", and the "affiliation name" comes from the dataset "cb_people" itself. The "relationship object id" should be consistent with the "affiliation name", but it sometimes does not. We kept duplicated values because the same person may get different

degrees from one university or from several universities. For missing values, we removed null values. For inconsistent values, we used filter or the "like" method in MySQL to choose certain words like "CEO", "CFO".

4. Methodology

4.1 SQL

Before our exploration, we assumed that funding is related with regions, industries, and people. So, we used SQL to query the counts and the average funding amount of the innovative companies group by regions, industries. To see the relationship between the funding and people, we also joined the two datasets "companies" and "people" by the company id, so that we can query the counts of people and the average funding amount group by people's institution, which is the university they graduated from.

4.2. Rescale the data

We used GoogleVis to draw heat maps to see the number of innovative companies in different countries and in different states in US. Then we found that the data were very skew, so we used log transformation to rescale the data to get better visualizations. The data of the funding raised amount is also very skew, so again we used log transformation to rescale the data to get a better distribution when plotting the histogram.

4.3. Word cloud

We want to explore more detailed information for the top three industries of the innovative companies in our dataset "companies", so we used the "overview" information to create three word-cloud graphs. We removed the common stop words and used Scala in Spark to calculate the count of each word. After reviewing the results, we also removed some words like "user" and "service" that do not provide special or useful information. Finally, we used the website *WordItOut* (worditout.com/word-cloud) to create the word cloud graphs.

4.4. Association rule

For the dataset "people", we found that 27% of the people have a MBA degree (see more details in section 5.4), so we decided to use association rule to see the relationship between the degree MBA and different titles. We first calculated the support count of MBA, which is 62140. Then we calculated the confidence value by dividing the support count for {MBA, title} by the support count for {title}. Here we chose the "title" as "board of member", "board of director", "founder", "CEO", "CFO", and "VP" (vice president).

4.5. Latent semantic index

Since the dataset "people" is all about text, we also used latent semantic index (LSI) to see the relation among the words in the dataset. We only used four columns of the dataset "people": "title", "degree_type", "subject", and "institution". We used Python packages "NLTK" and "gensim" and ran the program locally with Pycharm, which took us a few minutes to get the result. We first cleaned the document string by transforming the words to lowercase and removing special characters or replacing special characters by white space. Then we tokenized the document string, and turned our tokenized document into a document-term matrix. Then we

used the term frequency-inverse document frequency (tf-idf) in the LSI model (reference: https://radimrehurek.com/gensim/models/lsimodel.html), and we only generated one topic to see the strongest relationship among these words.

4.6. LDA

To get insights of what companies are doing, we further analyzed the overviews in the dataset "companies", which is the text description of companies. We planned to find out major topics by Latent Dirichlet Allocation (LDA). However, LDA is time consuming, and the time spent to run LDA increases tremendously when the number of topics increase even by one. It is NP-hard when the number of topics is large (Sontag & Roy 2009). Therefore, we want to find an optimal small number of topics based on evaluation of number of clusters generated by other faster clustering algorithms such as k-means. Here we chose k-means because k-means is faster with time complexity of $O(ndki)$, where n is the number of d-dimensional vectors, k is the number of clusters and i is the number of times of iterations (Wikipedia contributors, K-means Clustering). In practice, k-means only takes a few minutes while LDA takes 2 to 3 hours (for 5 topics). The evaluation of the number of clusters by k-means provides a good approximation for number of topics in LDA.

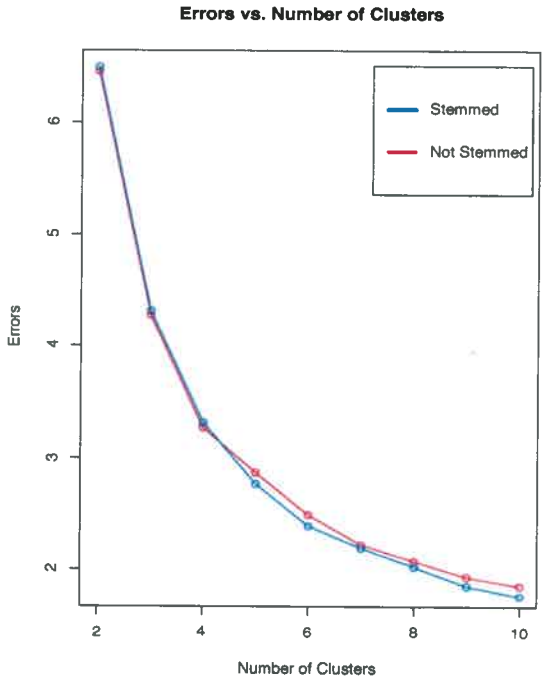


Fig 1. Errors with number of clusters

Figure 1 shows the outputs from k-means clustering implemented by Spark in Scala, specifically, the MLlib library under spark apache. The algorithm can be founded at spark.apache.org ("Clustering - RDD-based API"). For each of the k-means clustering task, it took one to two minutes to run and the input file is "companies" dataset.

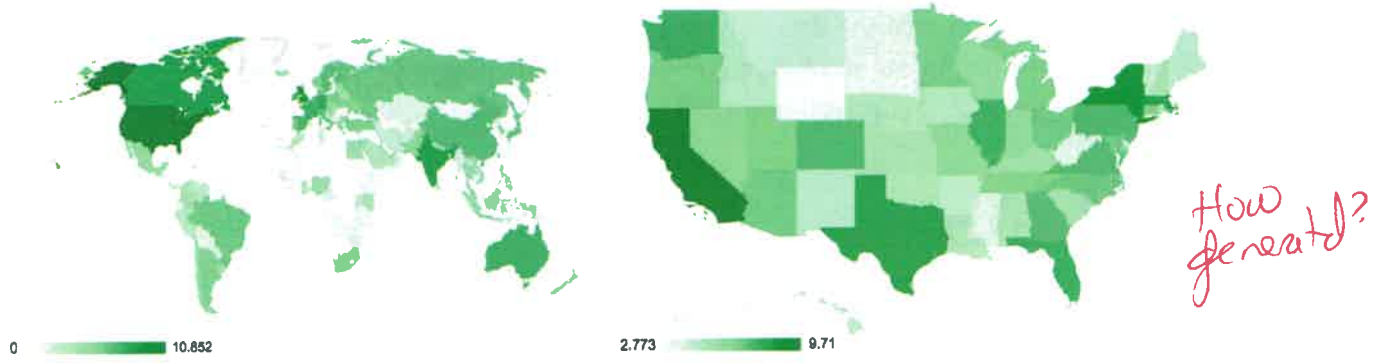
We processed the data for twice: one is stemmed texts, the other is not stemmed texts. The errors (in 10^7) in the Figure 1 stand for within set sum of squared errors that provides a good measurement for the quality of clustering. The smaller the errors, the better the clustering. As we can see from Figure 1, when the number of clusters increase, the errors decrease. The errors for stemmed texts start to be smaller than errors for not stemmed texts when the number of clusters is greater than or equal to 5. When the number of clusters increase from 2 to 5, the errors decrease tremendously. When the number of clusters increase from 5 to 10, the errors decrease only moderately. Thus, cluster number of 5 is considered as an "elbow" point. So, we used 5 as the number of topics for LDA, and we chose stemmed texts of overview as our source texts since it has smaller within set sum of squared errors.

Our LDA model is implemented by Python's "genism", "NLTK" and its "stop_words" and "PorterStemmer" packages, and the algorithm is referred by Barber "Latent Dirichlet Allocation (LDA) with Python". We first tokenized and stemmed the overviews for each company, then we converted the tokenized documents into an id term dictionary. Thirdly, we transformed the tokenized documents into a corpus of document-term matrix, so that we can generate the major topics by the LDA model. We used cluster size of 1 master, 2 cores, and 4 tasks for our analysis and it took 2-3 hours. The results of LDA are shown in section 5.5.

5. Results

5.1 General views

5.1.1 Heat Maps of innovative companies in different regions

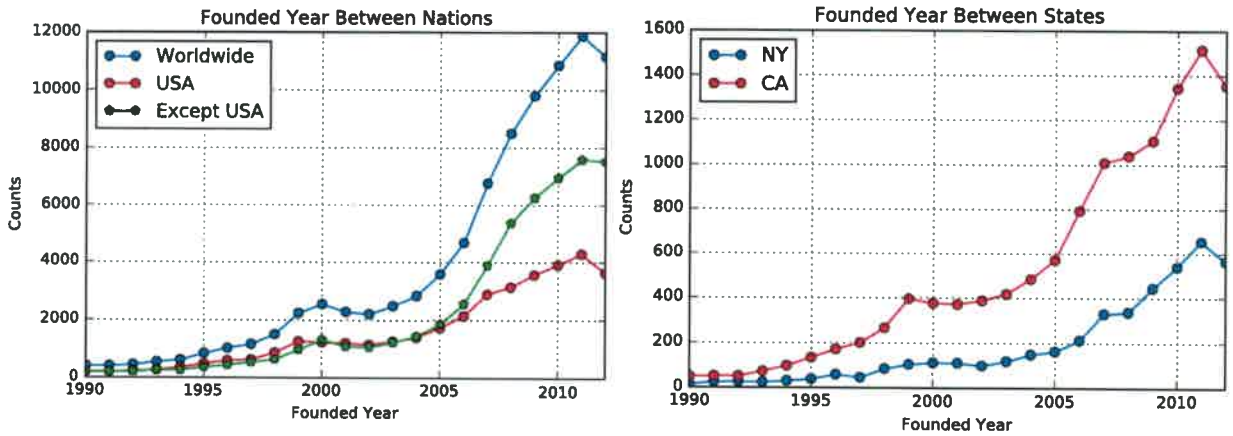


(a) (b)

Fig 2. Heat map of the number of innovative companies in (a) each country (b) each state in US

The heat maps show the counts of innovative companies in the world and in the US from the dataset "companies" from 1990 to 2012. In Figure 2, the darker the color is, the larger the number of innovative companies is in the region. As we used log transformation to rescale the data, the legends represent the log counts. Figure 2(a) shows the heat map of each country, and America, Britain, India, and Canada are the top four countries. Figure 2(b) is the heat map of each state in US, and California, New York, Massachusetts, and Texas are the top four states.

5.1.2. Trends over the past two decades



(a) (b)

Fig 3. Number of founded companies of (a) nations (b) states

The number of innovative companies created by year from 1990 to 2012 is shown in Figure 3. The data are generated from the "companies" dataset. The data in 2013 is truncated because we find they are incomplete. From Figure 3, we can see that the number of companies created in 2012 goes down, and the reason might be that the data in 2012 are also incomplete. In Figure 3(a), the blue line represents the total number of founded companies worldwide, the red line represents the companies created in the USA, and the green line represents in other nations except USA. We can see that the overall trend is increasing after 1990, and it reaches the peak points in 2011 with 11,884 companies founded worldwide and 4,294 companies founded in USA. In Figure 3(b), we compared the data between California and New York, because they are the top two states that has the largest number of companies founded each year. We can see that they both have similar trends with the overall trend in USA.

5.2. Funding

5.2.1 Summary of funding round types

Table 4. Summary of the funding round type data

Funding Round Type	Min (USD)	Max (USD)	Average (USD)	Total Count	Count > 1 million	Percentage > 1 million
angel	1,000	1,499,999	399,649	13163	801	6%
crowdfunding	3,591	20,000,000	1,783,027	111	36	32%
other	1,000	3,835,050,000	11,538,451	4201	2006	48%
post-ipo	10,500	3,200,000,000	186,559,310	87	75	86%
private-equity	2,000	2,600,000,000	27,645,086	1043	840	81%
series-a	291	681,759,114	6,432,685	9873	8246	84%
series-b	1,000	315,000,000	11,911,836	4892	4510	92%
series-c+	1,000	950,000,000	21,738,938	4216	4025	95%
venture	1,000	1,500,000,000	9,069,114	15342	9497	62%

There are many types of funding rounds in our dataset. Except post-ipo, the minimal raised amounts in us dollars are similar, which are around 1000 dollars. According to the average raised amounts, in general, the increasing order of raised amount is angel, series-a, series-b, series-c+. Other funding round types include crowdfunding (which starts in 2010 from our dataset), post-ipo (which is very rare, and Zillow is one example in our dataset), private-equity, and venture. To make our following research more meaningful, we set a threshold of one million us dollars to cut off some small companies, and it seems to be a good threshold. Because it cuts off 94% of angel funding, and there are more than 80% of series-a funding that bigger than one million, and more than 90% of series-b and series-c+ funding amounts are bigger than one million.

5.2.2. Distribution of total funding amounts

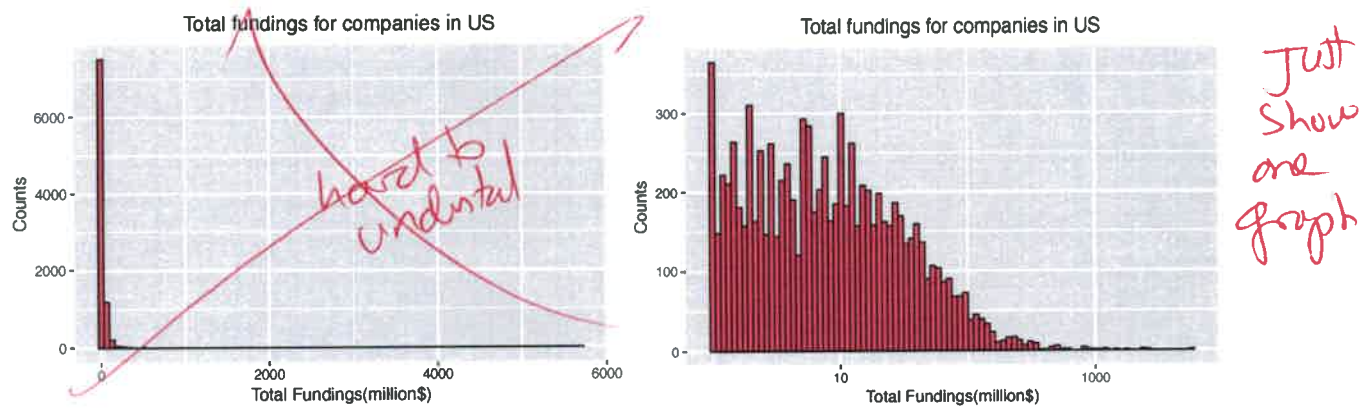


Fig 4. Distribution of total funding amounts > 1million

Figure 4 shows the distribution of the total funding amounts which are bigger than one million. The data are generated from "companies" dataset, and the graph is plotted using R ggplot2 package. Before we rescale the data, the distribution is very skew, which is shown in the left graph of Figure 4. To get a better distribution, we take a log transformation to rescale the data, and the result is shown in the right graph of Figure 4. The average total funding amounts is 16.14 million dollars, while the median total funding amounts is 3 million dollars. Again, this tells us that the distribution is quite skew.

5.3. The relationship between funding amounts and industries, regions, and people

In this section, we are going to verify whether our assumptions that the funding amount has some relationship with industries, regions, and people are right.

5.3.1. Industries with funding amounts

We compared the counts and the average funding amounts of innovation companies group by industries. From Figure 5(a), we can see that software, web, ecommerce, and games video are the top four popular industries. However, when we look at the average funds group by the industries,

which are shown in Figure 5(b), we find that the relationship between the average funds and popularity is somewhat negative. This makes sense because there are too many competitors in the top four industries and these industries do not need high technology or expensive equipment. It is not surprising to find that industries like automotive and nanotech have high average funds. We also find that social ranks on the top.

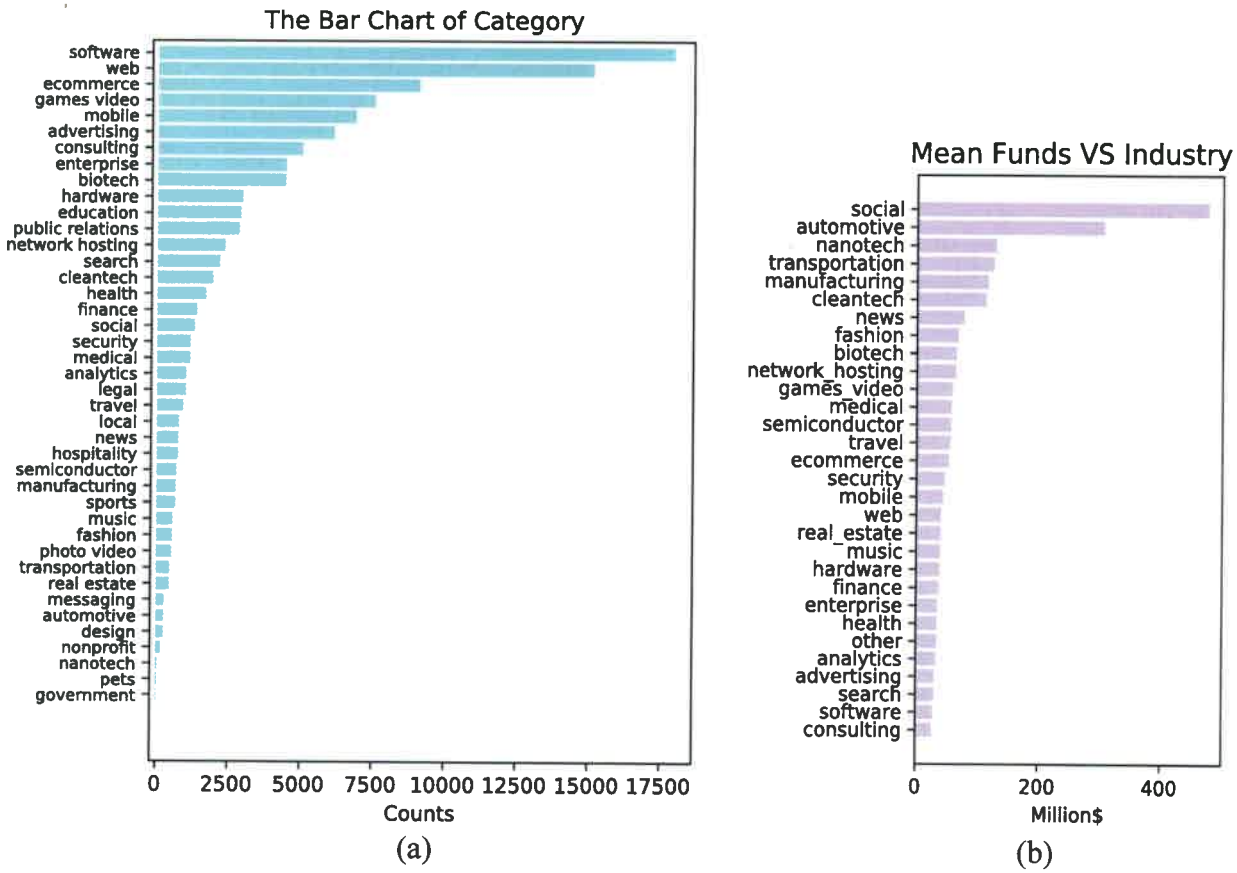
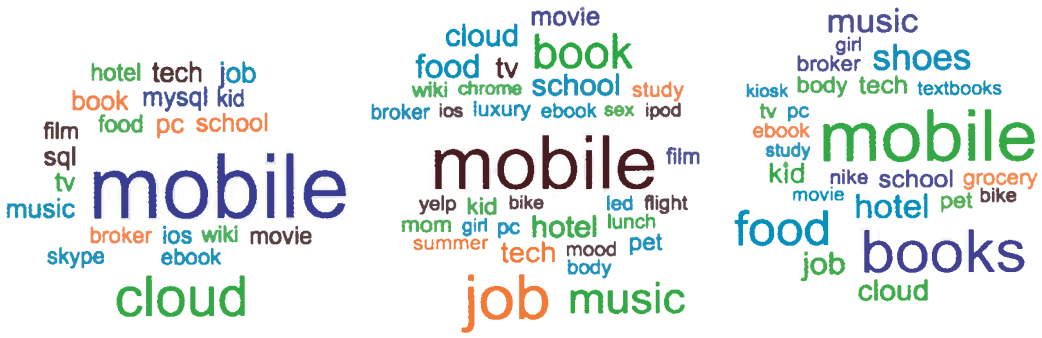


Fig 5. (a) The counts of innovative companies group by industries (b) The average funding amounts of innovation companies group by industries

We also made word-cloud graphs for the top three industries to get the idea what these companies are doing, and the results are shown in Figure 6. We find that mobile and cloud are popular in these companies. Although all three categories are similar in terms of top words (i.e. they all have mobile and music as their most frequent words), there are differences. Companies in software category tend to focus more on "cloud". Companies in web category tend to focus more on "job" and "book". Companies in e-commerce category tend to focus more on "books", "food" and "shoes".



(a) software (b) web (c) e-commerce

Fig 6. Word clouds of category (a) software (b) web (c) e-commerce

5.3.2. States with funding amounts

We calculated the average funding amounts with each state from "companies" dataset, and the top 25 states are showed in Figure 7. We can see that Illinois has the largest average funding amounts, DC ranks the second, and California ranks the third. However, except the top three states, the left states have relatively similar average funding amounts. Therefore, we reach the conclusion that average funding amounts have some, but not much relationship with states.

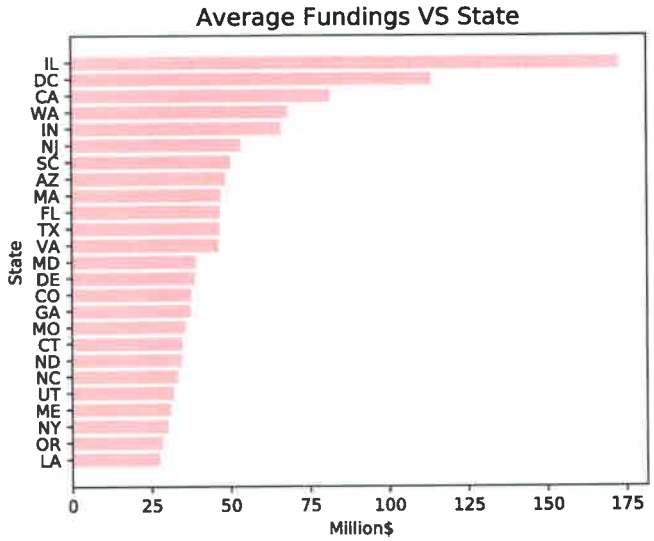


Fig 7. Funding amounts with states

5.3.3. Graduation schools with funding amounts

We used the dataset "people" to see where these people graduated from, and the result is shown in the Figure 8. We can see that universities with high ranks like Stanford University, MIT, and UC Berkeley have more people in our database. It seems like these high ranks universities put more emphasis on innovation than normal universities.

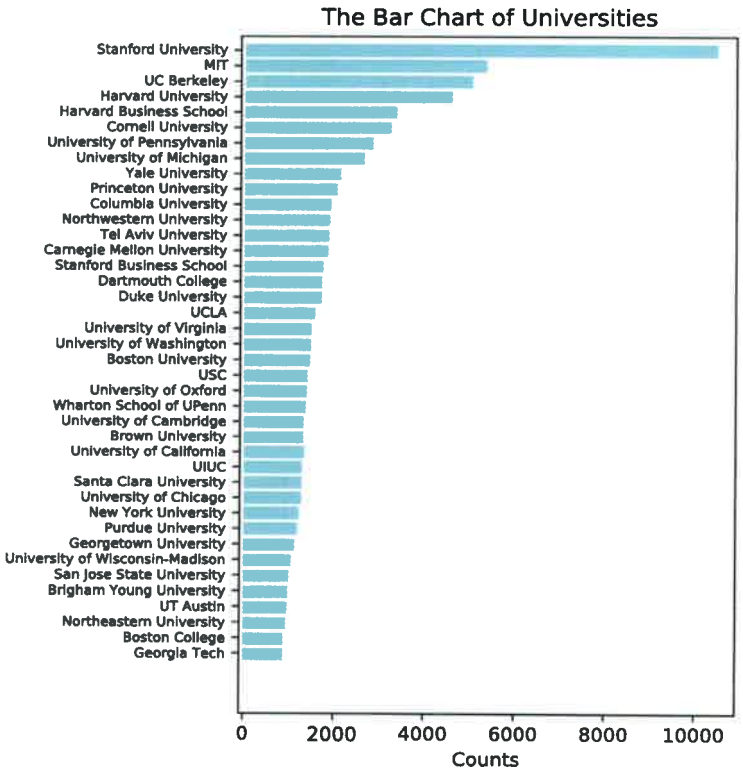


Fig 8. Counts of graduation universities

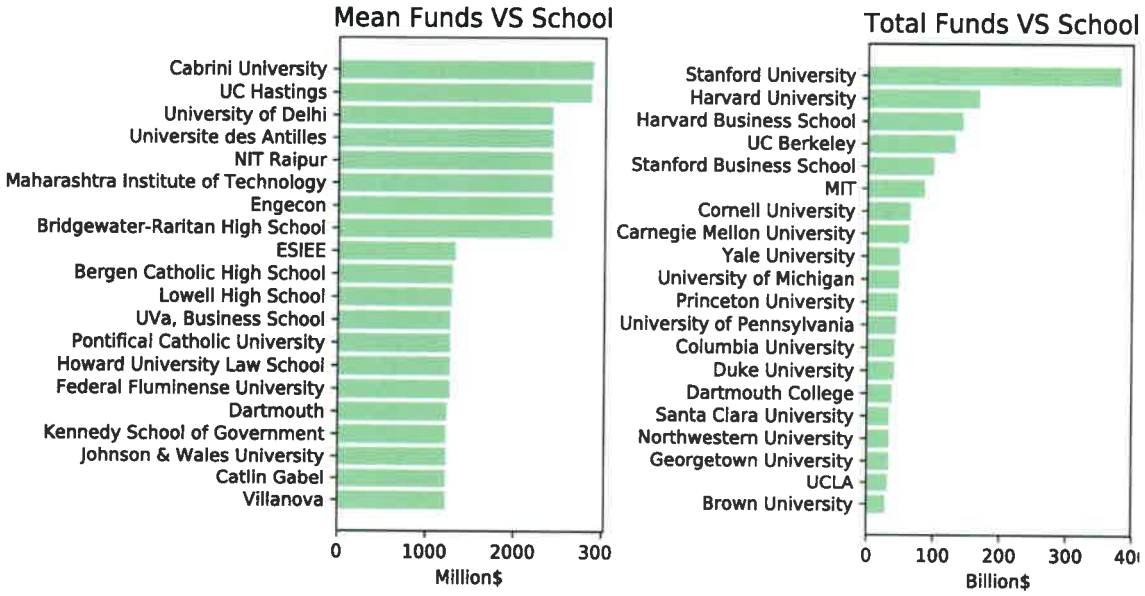


Fig 9. Average and total funding amounts with graduation universities

We compared the average funding amounts and the total funding amounts of the innovation companies group by people's graduation universities to see whether people from high ranks universities help the companies to earn more funding. The average funding amounts with school is shown on the left of Figure 9. We can see that there are not many big-name schools, and some

of them are even high schools. It seems that some people only have a high school degree. The right of Figure 9 shows the relationship of total funding amounts with schools, and there are many big-name universities. Here we did not normalize our data by dividing the number of students in each university, because it is hard to define what number we should use. Therefore, we can know that there is nearly no relationship between the average funding amounts and the graduation schools of the people. However, in well-known universities, more graduates tend to have a place in high management level in innovative companies.

5.4. More exploration about people

We used association rules to find the relationship between the degree MBA and different titles. We only chose the degree type MBA because MBA is much more special than other types like BS and MS. We find that the ratio of the support count of MBA (62,140) is greater than the count of all people who has a degree (359,267), which is 0.17. As there are duplicated people in this dataset (one might have two or more degrees), we estimated that more than 20% people have MBA degrees. If we calculate the count of MBA (62,140) over the number of people, which is the rows of "cb_people" (226,709), then the ratio is 0.27, that is, about 27% of the people in these innovative companies have MBA degrees, which is a large proportion. We find that 23% of the board members hold a MBA degree; 22% of the board of directors hold a MBA degree; 11% of the founders hold a MBA degree; 15% of the CEOs hold a MBA degree; 26% of the CFOs hold a MBA degree; 21% of the vice presidents hold a MBA degree. The results confirm our assumption that earning a MBA degree is useful to be in the management level position.

We also used the latent semantic index (LSI) method to see the relation among the words in the "people" dataset. We only used four columns: "title", "degree_type", "subject", and "institution". We find that the top four correlated words are "MBA", "engineering", "business", and "board". From the results of the association, we already know that "MBA" has a high relation with the word "board" (23% of the board members hold a MBA degree; 22% of the board of directors hold a MBA degree). It is also normal that "MBA" has a high relationship with "business". However, "MBA" also has a high relation with the subject "engineering". It seems that there is a trend that MBA is combining with both engineering and business nowadays.

5.5. More exploration about companies

We used LDA to discover five major topics based on the overviews, which is the text description of companies. Table 5 shows the five topics with 15 words for each topic, and each topic contains a list of stemmed words with their weights listed in descending order. The first topic contains words like "mobile", "game" and "application" as bolded (note that the words are stemmed), so we interpret it as "mobile applications and games". In the same way, we interpret the other four topics as "healthcare and medicine", "office system", "market, data and media", and "online social networks". These five topics are somewhat overlapped with the variable "categories" in the dataset "companies", but not the same.

explain in caption. why are words bolded

what does bold mean?

Table 5. Topics generated by LDA

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Term	**Weight**	**Term**	**Weight**	**Term**	**Weight**	**Term**	**Weight**	**Term**	**Weight**
mobil	0.032	develop	0.014	compani	0.017	provid	0.014	**user**	0.010
develop	0.020	**health**	0.011	provid	0.012	busi	0.014	com	0.009
com	0.019	compani	0.010	**servic**	0.012	**market**	0.012	**onlin**	0.008
game	0.018	product	0.009	**product**	0.011	**servic**	0.012	**social**	0.007
app	0.018	student	0.008	wa	0.011	web	0.011	can	0.007
www	0.014	**medic**	0.008	inc	0.009	manag	0.011	site	0.007
http	0.014	research	0.007	offer	0.009	**solut**	0.010	use	0.006
applic	0.013	**care**	0.007	oper	0.008	softwar	0.009	make	0.006
compani	0.011	educ	0.006	includ	0.008	compani	0.008	will	0.006
wa	0.008	technolog	0.006	found	0.008	technolog	0.008	**peopl**	0.006
company	0.008	use	0.006	ha	0.007	custom	0.008	**share**	0.006
ha	0.008	inc	0.006	base	0.006	develop	0.008	new	0.005
found	0.007	wa	0.006	services	0.006	**data**	0.008	creat	0.005
android	0.006	**patient**	0.005	**system**	0.005	design	0.007	one	0.005
web	0.005	base	0.005	**offic**	0.005	**media**	0.007	find	0.005

6. Conclusion and Future Work

Before we started the exploration, we assume that there are relationships between funding and region, funding and industries, and funding and people. After our exploration, we find that total funding amounts have more to do with industries than regions and the universities that people graduate from. Mobile and cloud are becoming more popular in today's innovative companies. When studying the description of companies, we find that major topics in them include mobile applications and games, healthcare and medicine, office system, market, data, media, and online social networks. Higher management level people typically hold a MBA degree.

There is still a lot to do if we have more time. We want to explore the funding amounts by rounds instead of the total funding amounts. We are also curious about investors who provide funds to innovative companies. We think there must be some relationship between them and we would like to explore it. To enrich our exploration, we would like to crawl the newest dataset from CrunchBase instead of using the "2013 snapshot". There are many limitations of "2013 snapshot". For example, it only provides the data before 2014, and there are lots of missing values and inconsistent values. From the CrunchBase website, we can see that the newest system has updated and the city names are consistent now. It has also changed the content of category. In "2013 snapshot", one company can only belong to one category. However, in the new system, one company can have several categories, which makes more sense in the real world.

Reference

Liang, Yuxian Eugene, and Soe-Tsyr Daphne Yuan. "Predicting investor funding behavior using crunchbase social network features." *Internet Research* 26.1 (2016): 74-100.

Xiang, Guang, et al. "A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch." *ICWSM*. 2012.

Huang, Anna. "Similarity measures for text document clustering." *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 2008.

Sontag, David, and Daniel M. Roy. "Complexity of Inference in Topic Models." (2009): 1-4. Web.

Wikipedia contributors. "TechCrunch." *Wikipedia, The Free Encyclopedia.*, 5 Apr. 2017. Web. 1 May 2017.

Wikipedia contributors. "K-means Clustering." *Wikipedia, The Free Encyclopedia.*, 20 Apr. 2017. Web. 1 May 2017.

Barber, Jordan . "Latent Dirichlet Allocation (LDA) with Python." *Amazonaws*. N.p., n.d. Web. 1 May 2017.

"Clustering - RDD-based API." *Clustering - RDD-based API - Spark 2.1.0 Documentation*. N.p., n.d. Web. 1 May 2017.