Data Science Institute
COLUMBIA UNIVERSITY

Fall 2020 Capstone Project

Progress Report 1

Measuring startup strategy and its evolution

October 23, 2020

Members

Derek Chen - cc4506
Wangzhi Li - wl2737
Bernardo Lopez Vicencio - bl2786
Yinhe Lu - yl4372
Alberto Munguia Cisneros - am5334

Mentor

Prof. Jorge Guzman
Columbia Business School

Table of Contents

I. Introduction

A. Project Scope and Goal

The purpose of this project is to develop a new analysis of the strategy of firms using text-based machine learning. The key insight is that distance in the initial statements made by startup companies can be partially indicative of their strategic positioning to each other, and this, in turn, could be an explicative factor of the future performance of the startup. An early-stage implementation of this idea is already available in the paper *"Measuring Founding Strategy"* by Professor Jorge Guzman and former Columbia student Aishen Li.

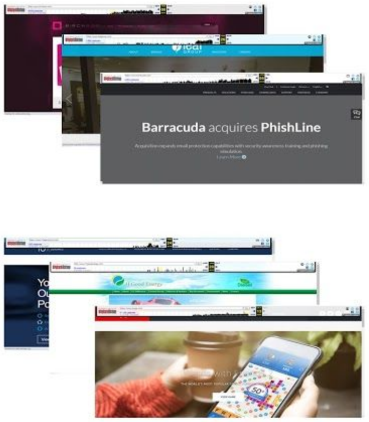
To achieve our goal, we will develop the project in four modules:

- Module I consists of extracting the early statements and other relevant information from a representative sample of startups. To achieve this purpose, we will use the digital library of the Wayback Machine website to access snapshots of early versions of the startup website and web-scrape the information accessible on 1 level depth.
- Module II also consists of extracting the relevant information for public companies that exist around the same year as the startup's foundation. To complete this task, we will follow a similar approach by using the Wayback Machine website to access relevant information for the sample of public companies
- Module III entails the exploration of Natural Language Processing (NLP) techniques to determine the similarity of statements between startups and public companies. In this section, we will investigate the closeness of the statements by exploring dimensions of lexical and semantic similarity. Furthermore, we will propose a measure of similarity that could reflect the strategic positioning of the startups
- Module IV focuses on the implementation of machine learning and statistical analysis to evaluate to what extent the strategy measure proposed in Module III can predict the performance of the startup in terms of economic success.

The entire project will be developed in Python 3 and implemented in an instance in Amazon Web Services (AWS). The expected outcome of the project will be reproducible code and the improvement of the early version of the paper *"Measuring Founding Strategy."*

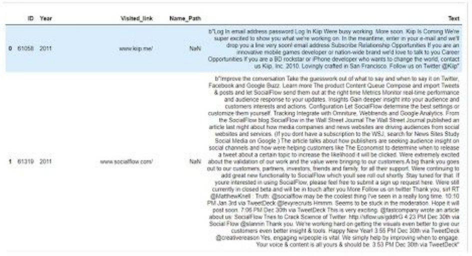
B. Concept Slide

Historical snapshoots
Startups and Public Companies
websites



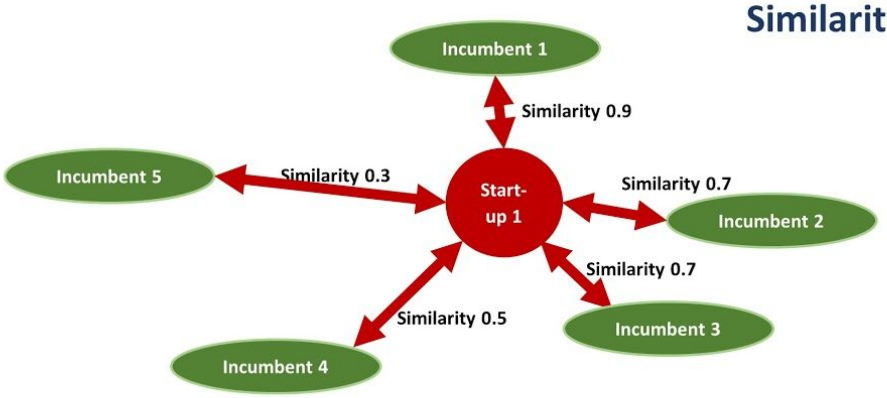
Web scraping

Statements
and relevant information



	ID	Year	Visited_link	Name_Path	Text
0	61058	2011	www.kip.me/	NaN	b"Log In email address password Log In Kip Were busy working. More soon. Kip Is Coming We're super excited to show you what we're working on. In the meantime, enter in your e-mail and we'll drop you a line very soon! email address Subscribe Relationship Opportunities If you are an innovative mobile games developer or nation-wide brand we'd love to talk to you Career Opportunities If you are a BD rockstar or iPhone developer who wants to change the world, contact us Kip, Inc. 2010. Lovingly crafted in San Francisco. Follow us on Twitter @Kip"
1	61319	2011	www.socialflow.com/	NaN	b"Improve the conversation Take the guesswork out of what to say and when to say it on Twitter, Facebook and Google Buzz. Learn more The product Content Queue Compose and import Tweets & posts and let SocialFlow send them out at the right time Metrics Monitor real-time performance and audience response to your updates. Insights Gain deeper insight into your audience and customers interests and actions. Configuration Let SocialFlow determine the best settings or customize them yourself. Tracking Integrate with Omniture, Webtrends and Google Analytics. From the SocialFlow blog SocialFlow in the Wall Street Journal The Wall Street Journal published an article last night about how media companies and news websites are driving audiences from social websites and services. (If you dont have a subscription to the WSJ, search for News Sites Study Social Media on Google.) The article talks about how publishers are seeking audience insight on social channels and how were helping customers like The Economist to determine when to release a tweet about a certain topic to increase the likelihood it will be clicked. Were extremely excited about the validation of our work and the value were bringing to our customers.A big thank you goes out to our customers, partners, investors, friends and family, for all their support. Were continuing to add great new functionality to SocialFlow which youll see roll out shortly. Stay tuned for that. If youre interested in using SocialFlow, please feel free to submit a sign up request here. Were still currently in closed beta and will be in touch after you More Follow us on twitter Thank you, sir! RT @MatthewKnell : Truth: @socialflow may be the coolest thing I've seen in a really long time. 10:10 PM Jan 3rd via TweetDeck @levyrecruits Hmmm. Seems to be stuck in the moderation. Hope it will post soon. 7:06 PM Dec 30th via TweetDeck This is very exciting. @fastcompany wrote an article about us: SocialFlow Tries to Crack Science of Twitter. http://sflow.us/gddfrG 4:23 PM Dec 30th via Social Flow @slarmin Thank you. We're working hard on getting the visuals even better to give our customers even better insight & tools. Happy New Year! 3:55 PM Dec 30th via TweetDeck @creativereason Yes, engaging w/people is vital. We simply help by improving when to engage. Your voice & content is all yours & should be. 3:53 PM Dec 30th via TweetDeck"

NLP Algorithms
Similarity and Strategy
score



ML prediction
of economic
success

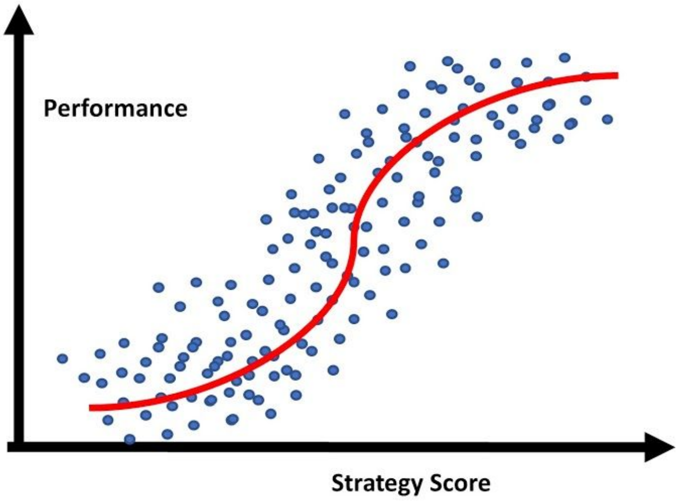


Fig1: Concept Slide of Measuring startup strategy project

II. Theoretical Overview

Strategic management is a field based on the main idea that a better strategy predicts higher firm performance. At least since Porter (1980), a significant portion of strategic management research and teaching has been done on understanding the cause of competitive advantages, e.g. what characteristics of firms and environments generate a better strategy and what choices managers have to make to achieve it. In contrast, little work has been done on algorithmically measuring the degree of competitive advantages itself - even within a single school of thought, such as strategic positioning, there is no specific approach to score a strategy.

The ability to measure strategic positioning would shed light on many debates and questions across the field of strategic management. For example, in recent debates on the importance of positioning for startups, some researchers proposed that the importance of positioning might be strongly downgraded by the significant uncertainty startups face and opportunities for adaptation (Reis, 2011), while others emphasize how founding choices might even determine the path of experimentation (Felin et al. 2019, Gans, Scott & Stern 2019). If we can measure the quality of the strategic positioning of a certain company, we can bridge the future performance and its evolution with its strategy.

Although it is common for human analysts to assess and compare the positioning of companies, the competitive advantage assessment is not a trivial endeavor: given that strategy stems from multiple disciplines (economic, sociology, and psychology), quantitative methods of analysis developed in these fields remain difficult to be synthesized. Furthermore, it is even more challenging to measure the strategic positioning of startups due to the extremely uncertain environment.

In this sense, can we quantitatively measure the entrepreneurial strategies, specifically, strategic positioning? The short answer is: yes, machine learning holds great potential to achieve it. Let’s look back to what proficient human strategy analysts will do: they listen to the companies’ statements, which emphasize their unique value proposition. This method can be perfectly summarized by a quota from Poter (1996), “competitive strategy is about being different.” Along this line, if we can systematically map the difference and value proposition of one company to others, especially to those closest competitors who shape competitive dynamics, would a measure of such distance (i.e. the degree of similarity) reflect the competitive advantages of strategic positioning?

Theoretically, such a method is feasible. Recent research has shown some systematic and observable differences across firms at the founding. Compelling qualitative evidence in the early synthesizer industry, as presented by Anthony et al. (2016), shows how different startups use different messages to create different value propositions. Similarly, a series of recent papers show how a large portion of the variance in growth outcomes across companies can be attributed to simple founding strategies (Guzman & Stern (2015, 2019)). Technically, text similarity in natural language processing (NLP) is mature enough for us to realize our proposed method. Such technique determines the “closeness” between two pieces of text based on both lexical and semantic similarity, i.e. the word-level and meaning-level similarity. And we will introduce these methods in the following section.

Previous research by Li and Guzman (2019) has proposed a prototype of a strategic positioning measuring technique based on machine learning. They showed the feasibility of measuring strategy using NLP techniques and arrived at a positive result that a higher founding strategy score can partially indicate a higher possibility of achieving high growth outcomes and receiving financing themselves, even conditional on location and cohort effects. However, this seminal works has room for improvement: 1) the estimates of founding strategy score are still imprecise; 2) the size of the data set is limited and the choice of text source for public companies is not consistent with that for startups; 3) the conclusions suffer from oversimplified assumptions on the competitions among companies, which ignored geographical constraints; 4) the role of strategy is underestimated through the use of fixed-effects, which excludes the effects of obvious co-founders related to location, cohort, industry, and seed financing year on the estimated strategy score. Based on their work, we intend to improve previous work from the following aspects: 1) to extend the dimensions of the original design of founding strategy score; 2) to explore different text similarity measuring techniques; 3) to upscale the web scraping to enlarge the text database.

III. Measuring Startup Strategy

A. Universe of companies and data sources

Our starting point for this project is the universe of companies that will constitute a representative sample of the US market, over which we will build our similarity and strategy metric and measure it against the economic performance of each startup. The two types of companies that we are going to focus on are Startups and Public companies.

In the case of the startups, we made use of a large population of US startups founded between 2000 and 2019; and that received any amount of funding from venture capital firms. We extracted the database from Preqin, which is an online private equity database that covers all aspects of the industry for all fund types, including buyout, venture capital, mezzanine, distressed, fund of funds, secondaries, natural resources, and others. Furthermore, Preqin provided information that will be vital in further steps of our project, for example, the website of the startup, its general description, economic sector, type and amount of funding received, name and characteristics of the investors.

For public companies, we use the database from Orbis, a database service from the Dutch company Bureau Van Dijk, which provides integrated qualitative and quantitative information from public companies. The population extracted from the Orbis database were all current and former public companies in the US. Also, the database contained critical information such as the company's website, listing and delisting date, market capitalization, economic sector, and financial information.

The following section will focus on an initial exploratory data analysis of the databases for public companies and startups.

1. Startup Database.

Our startup database has 13,704 unique startups that initiated operations between 2000 and 2020. 79.1% of the startups were founded between 2009 and 2018.

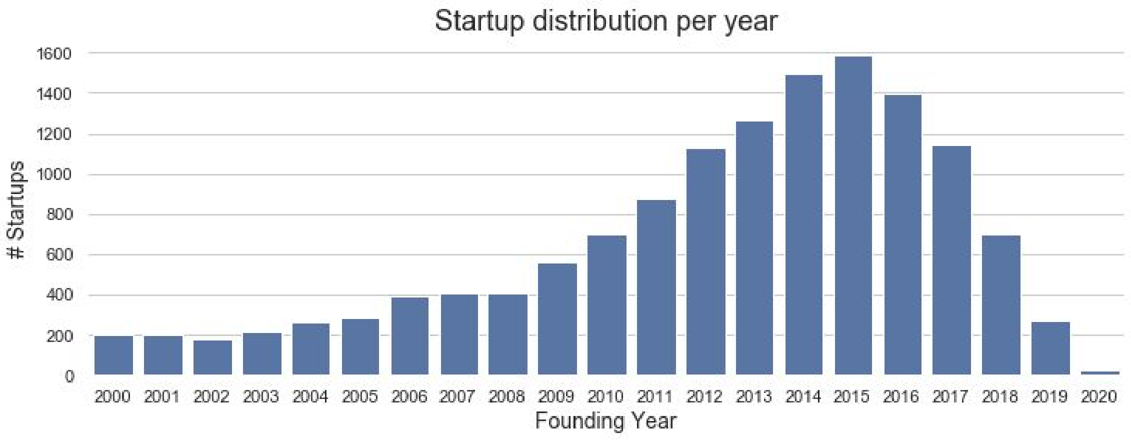


Fig2: Startup absolute distribution per year of foundation

Regarding the distribution of the number of venture capital deals and their corresponding size, we observe that seed funding, early rounds and unspecified funding present the highest frequency in terms of the number of deals and the accumulated amount of funding. However, the average size of funding tends to increase naturally as the startups become more mature.

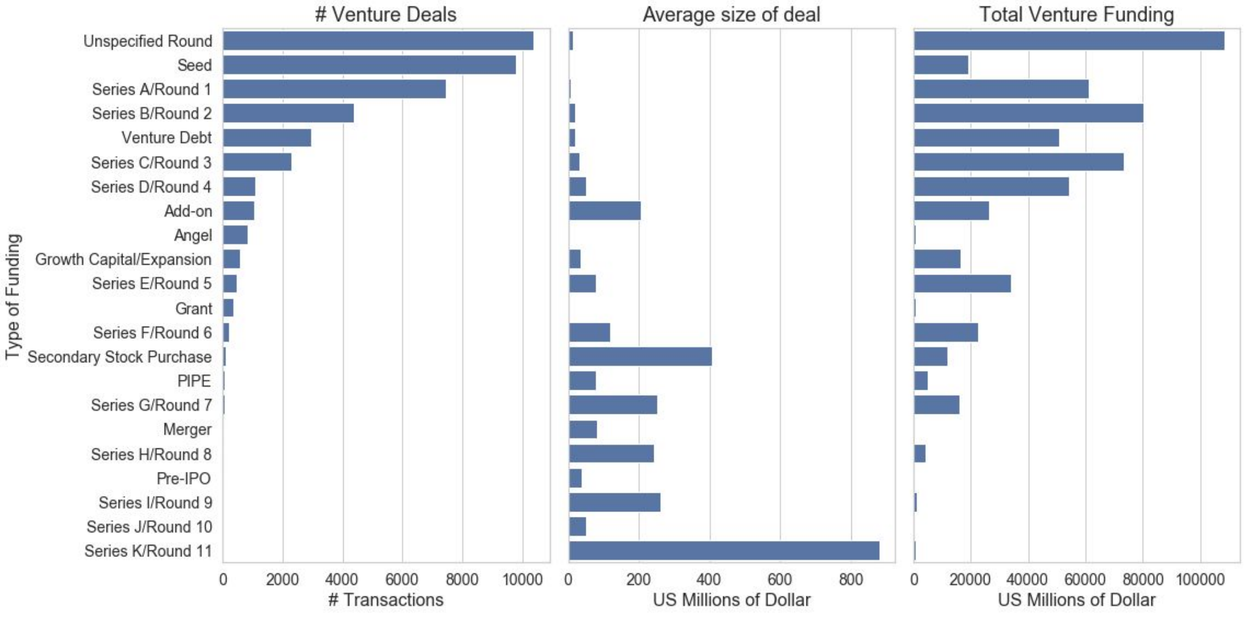


Fig3: Distribution of startups and funding deals per type of venture capital round of investment.

In terms of geographical distribution, California dominates the number of funded startups in a second distant place we observe New York followed by Massachusetts and Texas.

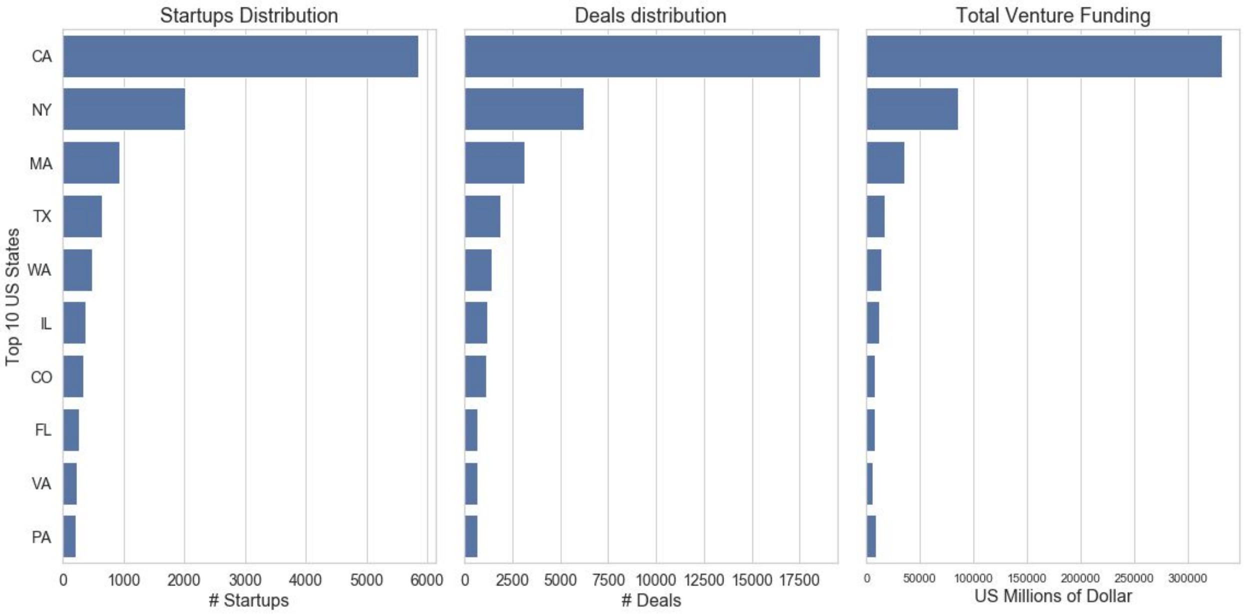


Fig4: Distribution of startups and funding deals per geographical location.

Finally, the startup distribution among industries is biased towards the Information technology industry (9,038 startups, 65.6%), completing consumer discretionary (1,610 startups, 11.7%), and healthcare (952 startups, 6.9%), the top three industries in our startup sample.

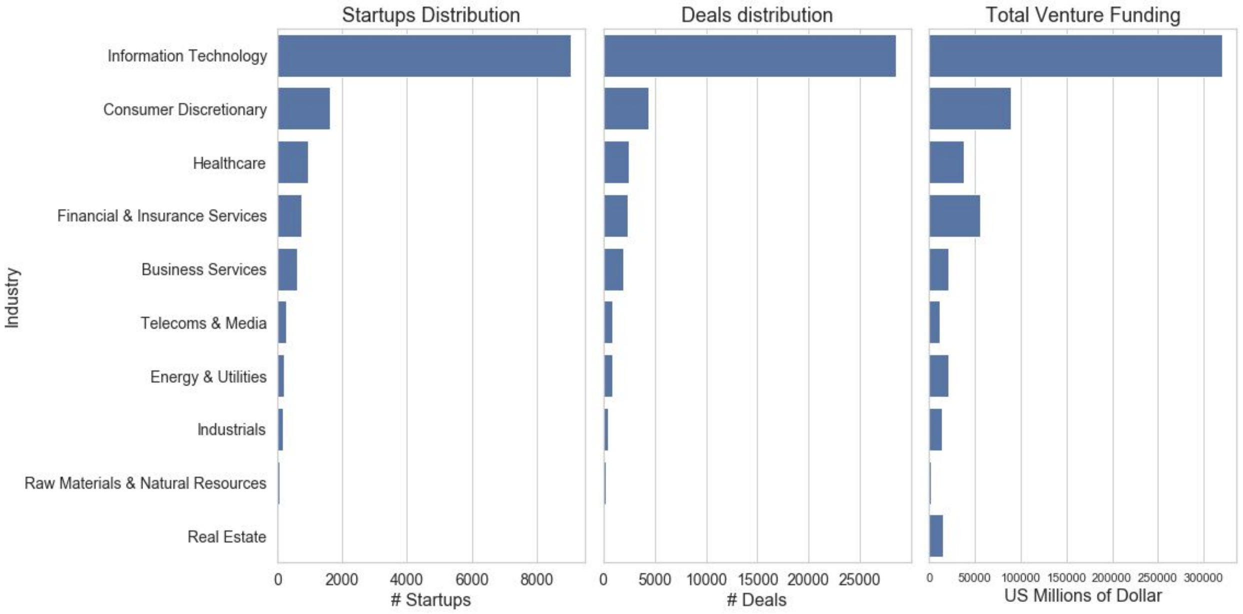


Fig5: Distribution of startups and funding deals per industrial sector.

2. Public Companies Database.

The database of public companies is formed by 22,333 companies that are currently public or were public between 2000 and 2020.

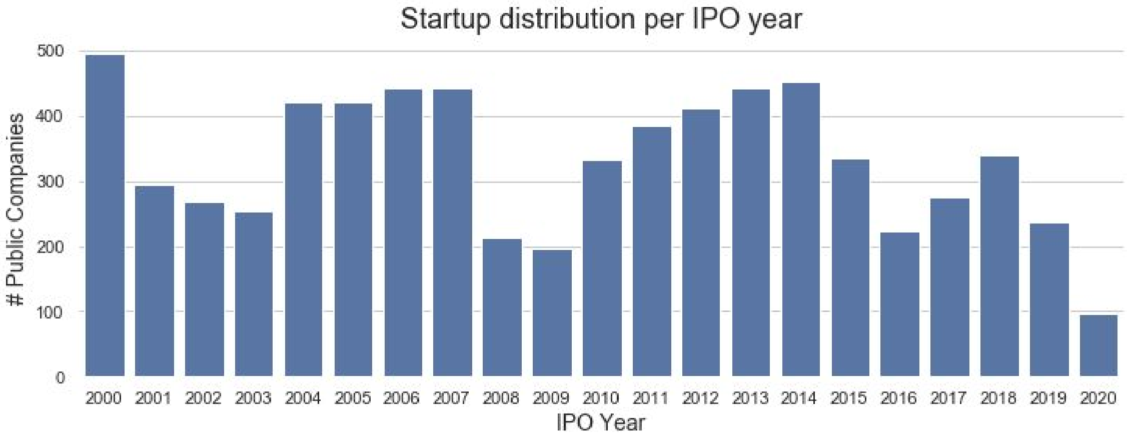


Fig6: Distribution of public companies per initial public offering (IPO) year.

In terms of the industrial sector, we observe that the majority of our public companies belong to the Finance and Insurance sector, 9,035 companies, followed by the manufacturing and Information sector with 4,544 and 1,785 companies respectively.

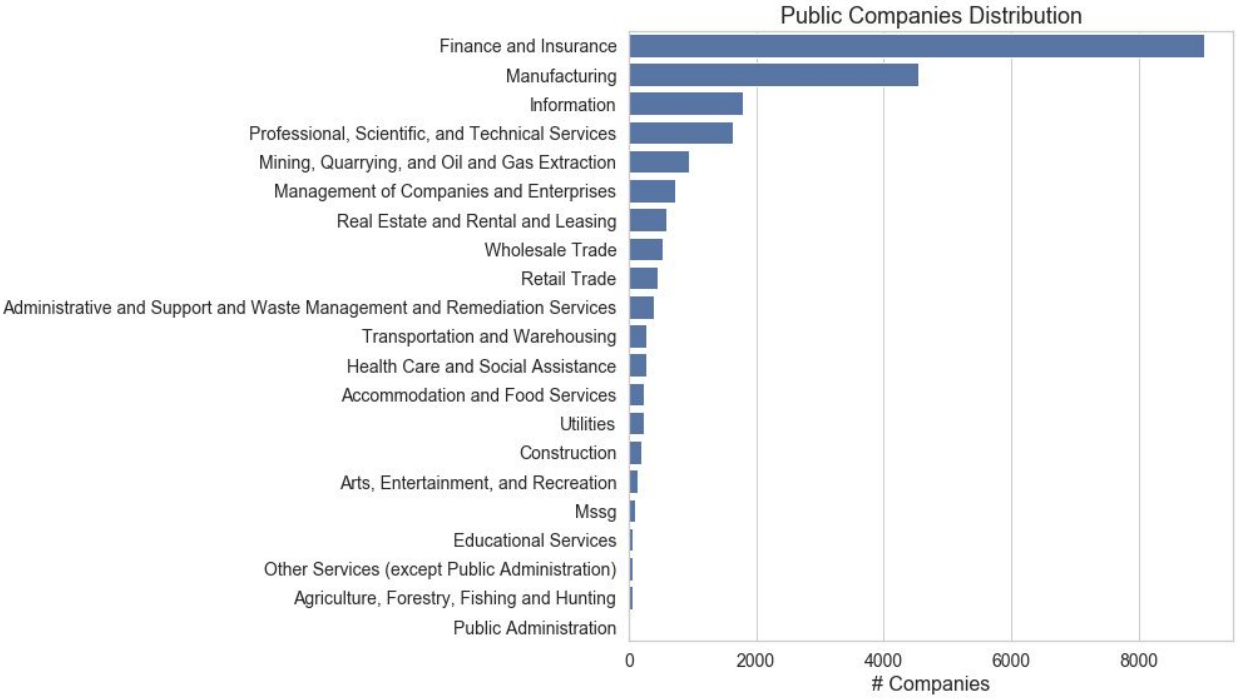


Fig7: Distribution of public companies North American Industry Classification System (NAICS).

From this initial exploratory analysis, we can draw the following conclusions: 1) For the startups, we observe that the population is concentrated in startups originated in California, New York, Massachusetts, and Texas. 2) The information sector dominates the economic orientation of the startups, and it does not coincide with the dominant sector of the public companies, the financial segment. This aspect could have a relevant impact on the next steps of our project, specifically in the results of the similarity measure.

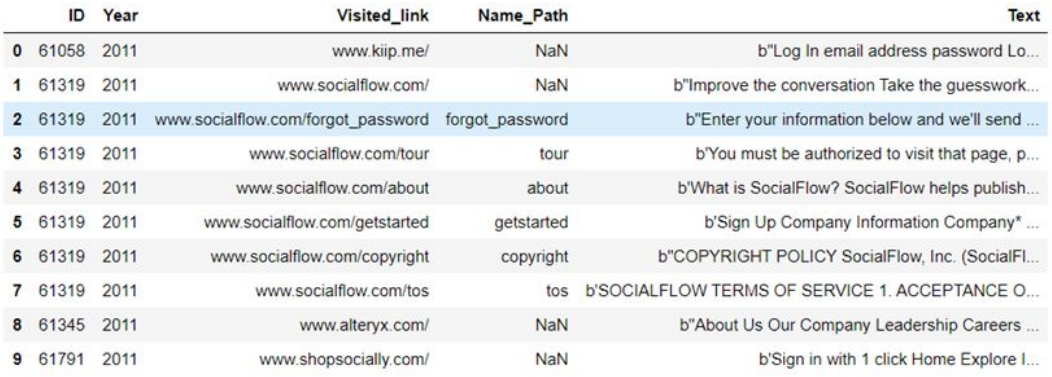
B. Web Scraping and Processing in AWS

This stage of our project consisted of extracting early statements and relevant information for the list of startups and public companies; we mentioned in section II.A., for each year between 2000 and 2020. To get website history data of each startup and public company in our database, we used the Wayback Machine, an online platform funded by the Internet Archive (archive.org), which provides access to a digital library containing over 330 billion web-page snapshots occurring in history. The web-page snapshots are taken several times a year for each unique domain on the internet. We implemented a web-scraping script to query website history data for companies from Wayback Machine. For the current stage of our project, we decided to implement a pilot test and scrape and download the initial website of each startup company in 2011.

Our web-scraping script will visit the historical version of the company websites and search for the content on the home page and all the first level links. The web-scraping algorithm that we implemented will follow the next steps:

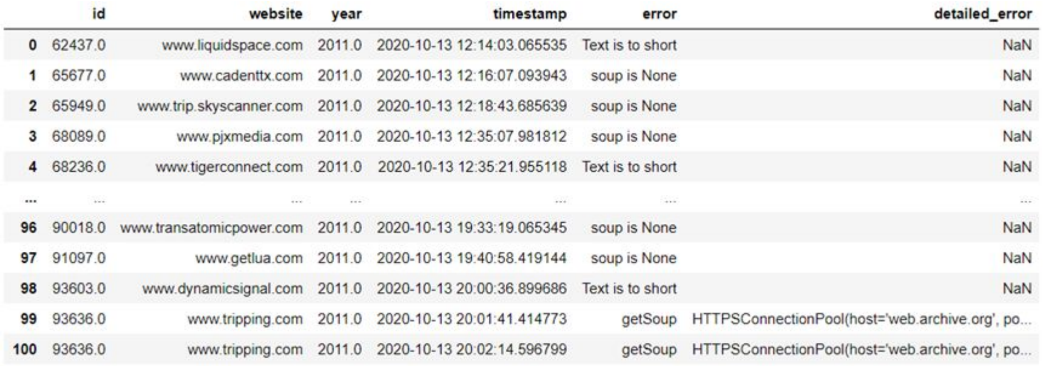
- As input, it will take the website of the company and the year of creation for the startups or a valid IPO year in the case of public companies.
- Through the Wayback Machine API, our code obtains the first timestamp of the following target year and will determine which of the available snapshots in the Wayback Machine archive we will web-scrape.
- Then the code visits the company's website and gets all the available links that are at one level depth on the website. We scrape all the data contained in the HTML file of each link, convert it into text and store it as a text file and in a structured database in a CSV.

Moreover, our web-scraping procedure has some crucial steps to get the information as clean and as consistent as possible. The scraper will automatically ignore the footer section of the website and error links like goDaddy.com or Error404 pages. We also implemented error log functionality to track errors that occurred during the scraping process; an error log table is generated in a CSV format at the end of each run. Some common bugs we encountered are texts from pages that are too short (less than 50 characters), or the scraper exceeds 30 redirect links and HTTPS responds too slow. As an example, here we have a snapshot of the output table and the error log.



	ID	Year	Visited_link	Name_Path	Text
0	61058	2011	www.kiip.me/	NaN	b"Log In email address password Lo...
1	61319	2011	www.socialflow.com/	NaN	b"Improve the conversation Take the guesswork...
2	61319	2011	www.socialflow.com/forgot_password	forgot_password	b"Enter your information below and we'll send ...
3	61319	2011	www.socialflow.com/tour	tour	b'You must be authorized to visit that page, p...
4	61319	2011	www.socialflow.com/about	about	b'What is SocialFlow? SocialFlow helps publish...
5	61319	2011	www.socialflow.com/getstarted	getstarted	b'Sign Up Company Information Company* ...
6	61319	2011	www.socialflow.com/copyright	copyright	b"COPYRIGHT POLICY SocialFlow, Inc. (SocialFl...
7	61319	2011	www.socialflow.com/tos	tos	b'SOCIALFLOW TERMS OF SERVICE 1. ACCEPTANCE O...
8	61345	2011	www.alteryx.com/	NaN	b"About Us Our Company Leadership Careers ...
9	61791	2011	www.shopsocially.com/	NaN	b'Sign in with 1 click Home Explore I...

Fig8: Web-scraping output, structured database.



	id	website	year	timestamp	error	detailed_error
0	62437.0	www.liquidspace.com	2011.0	2020-10-13 12:14:03.065535	Text is to short	NaN
1	65677.0	www.cadenttx.com	2011.0	2020-10-13 12:16:07.093943	soup is None	NaN
2	65949.0	www.trip.skyscanner.com	2011.0	2020-10-13 12:18:43.685639	soup is None	NaN
3	68089.0	www.pjxmedia.com	2011.0	2020-10-13 12:35:07.981812	soup is None	NaN
4	68236.0	www.tigerconnect.com	2011.0	2020-10-13 12:35:21.955118	Text is to short	NaN
...	...	...	...	...	...	...
96	90018.0	www.transatomicpower.com	2011.0	2020-10-13 19:33:19.065345	soup is None	NaN
97	91097.0	www.getlua.com	2011.0	2020-10-13 19:40:58.419144	soup is None	NaN
98	93603.0	www.dynamicsignal.com	2011.0	2020-10-13 20:00:36.899686	Text is to short	NaN
99	93636.0	www.tripping.com	2011.0	2020-10-13 20:01:41.414773	getSoup	HTTPSConnectionPool(host='web.archive.org', po...
100	93636.0	www.tripping.com	2011.0	2020-10-13 20:02:14.596799	getSoup	HTTPSConnectionPool(host='web.archive.org', po...

Fig9: Error log from web-scraping process .

Another challenging part of our project is the scaling up of the web-scraping procedure concerning the universe of startups and public companies. Since we will be running our web-scraping script on a much larger dataset (approximately 30,000 companies in total). Running on a local machine will take too much time and computation power; we will have to run our script on a virtual machine/cloud server. We launched our first version scraping script on Amazon Elastic Compute Cloud (EC2) worker node, which provides scalable computing capacity in the AWS cloud and eliminates the need to invest in time and computer power running on our local machine.

The instance type of our EC2 worker node is t2.large, which contains 2vCPU, 64-bit Arm cores, and 12GB memory. We did several tests on the capability and speed of the EC2 worker node. Our initial test was based on two local machines and one EC2 worker node. The two local machines did a run on the first 500 companies in the Preqin database dataset, and the EC2 worker node completed the run on the rest 191 companies. There were 691 companies in 2011 in the Preqin database dataset, and we got around 460 companies' history website data from scrapping the 691 companies on Wayback Machine. We then did a full run on EC2 worker node using the same web-scraping script and obtained a record of 451 companies' history website data. The conversion rate is about 66%, which proves the feasibility of collecting our data through EC2 worker nodes.

C. Measuring Similarity and Strategy Score

To establish a relationship between a startup's strategy metric and its future performance, we have to compute a strategy score for each startup. Our measurement approach builds on the idea that written statements by firms partially reflect their strategic positioning[1]. The steps to create the strategy measure are:

- Create appropriate word embeddings, to transform the value statements of the company and other relevant information from text files into vector representations. To achieve this goal, we use two NLP techniques, TF-IDF and Word2Vec. The input for this task is the information we obtain from our web-scraping process.
- Measure the similarity between the statements of each company. Let be σ_{ij} the similarity measure between statements, s_i and s_j , of any pair of companies.

  $$\sigma_{ij} = h\,(s_i,\ s_j),\ \sigma_{ij} \in [0,\ 1]$$

  For this stage of our project, we will use as a function of similarity h, the cosine similarity, defined as:

  $$h(s_i,\ s_j) = s_i \cdot s_j \ / \ \| s_i \| \ \ \| s_j \|$$

  where s_i and s_j are vector representations of the company's statements. Companies with a value of similarity equal to 1 have completely equivalent statements, while companies with a similarity of 0 have no relationship to each other. Companies with partial similarity are in between.

- The next step consists of the creation of the strategy measure, for that purpose first we compute the complement of the similarity measure, which is the distance between statements s_i and s_j , this distance reflects how different is the value proposition between two companies.

  $$d_{ij} = 1 - \sigma_{ij}$$

  Then, we average the distance among the closest statements for each startup. and the result will be the strategy measure of the startup.

  $$\widehat{S}_i = \tfrac{1}{5} \sum_{j \in J^5} d_{ij}$$

Notice that the procedure described above is our starting point, further work will be made in the following steps of the project regarding the adequacy of the word embedding procedures and the strategy measure. In the following subsections, we will present a brief explanation of the embedding algorithms, TF-IDF and word2Vec, and the

[1] A more extensive explanation could be found in Guzman, J., & Li, A. (2019). Measuring Founding Strategy

results from an early implementation of the strategy score procedure that we have described in the steps above.

1. TF-IDF: Term Frequency - Inverse Document frequency

The TF-IDF algorithm is the combination of two different algorithms, term frequency and inverse document frequency. It assigns a weight to every word in the document, which is calculated using the frequency of that word in the document and frequency of the documents with that word in the entire corpus of documents.

$$TF\ IDF(t) \ = \ TF(t) \bullet IDF(t)$$

where

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$IDF(t) = ln\ \left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right)$$

Term frequency TF represents how common a word is, while IDFs how unique or rare a word is. TF-IDF is intended to reflect how important a word is to a document in a collection or corpus. In tokenization we convert groups of sentences into token . It is also called text segmentation or lexical analysis

2. Word2Vec

Word2vec is an algorithm that learns word associations in a corpus. There are two main learning algorithms in word2vec: continuous bag of words and continuous skip gram. A model that has already been trained is able to receive a word or a sentence and output a 300 dimension vector. Such vectors are created in a way that the cosine similarity between vectors indicate the semantic similarity between words or sentences.

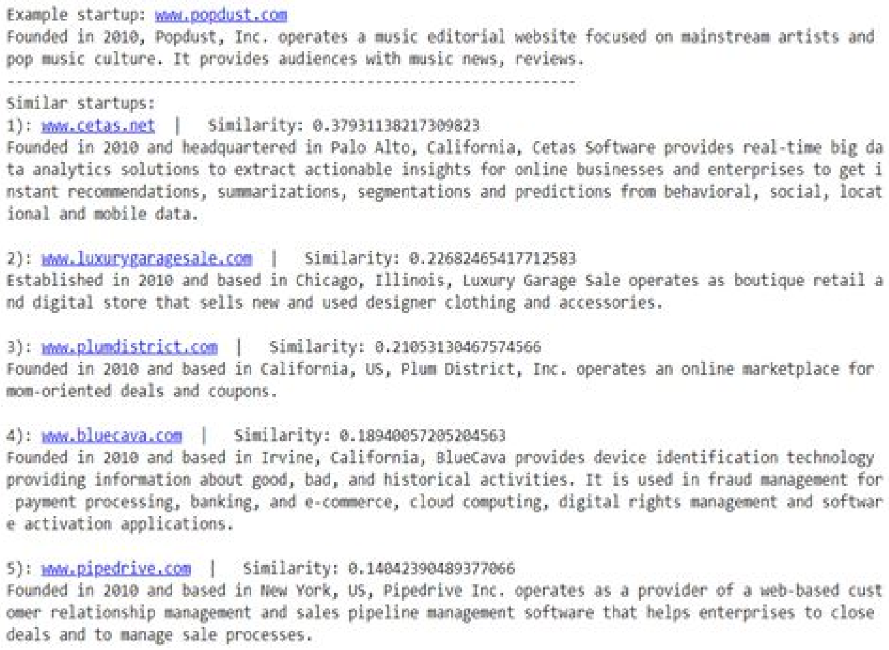
In our Word2Vec implementation, we are using the python library spacy. This library has a pretrained model “en_core_web_lg”, that is a English multi-task Convolutional Neural network trained on OntoNotes, with GloVe vectors trained on Common Crawl. This model can assign word vectors, POS tags, dependency parses and named entities.

One issue we found while implementing word2vec is that spacy’s library only allows strings of a maximum 1,000,000 characters. For strategy statements that are longer than that we are creating many vectors and then we do a weighted average.

3. Results

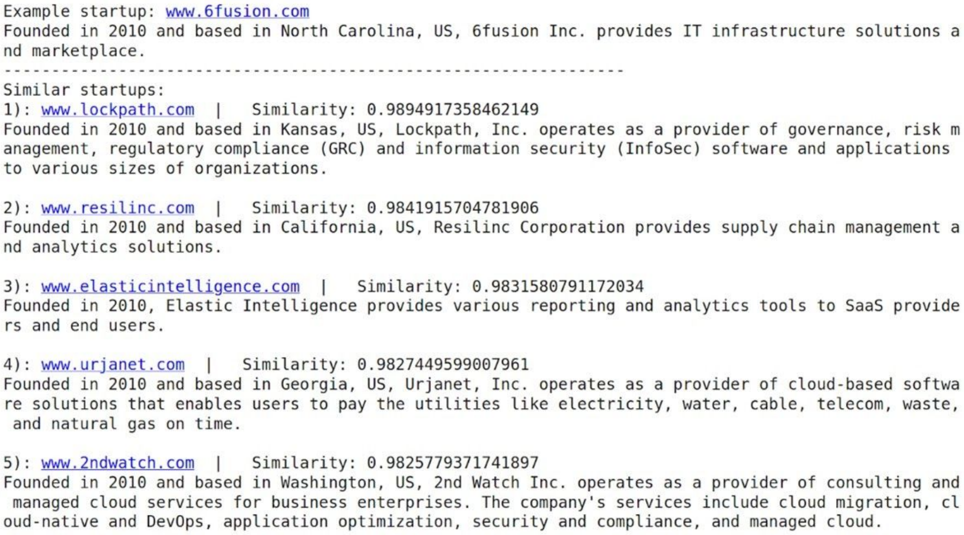
In this subsection, we will present some results as examples of the measurement strategy procedure with TF-IDF and Word2Vec. From this early implementation, we notice that the magnitude of the similarity and the strategy score between the two embedding methodologies are different. Word2vec, based on semantic similarity, gives higher results in the similarity score, with values close to 0.98; in contrast with TF-IDF, which is an embedding method based on lexical similarity, gives values that range from 0.15 to 0.38. Further work needs to be developed to explain the differences.

- Top 5 similar startups, TF-IDF and Word2Vec



```
Example startup: www.popdust.com
Founded in 2010, Popdust, Inc. operates a music editorial website focused on mainstream artists and
pop music culture. It provides audiences with music news, reviews.
----------------------------------------------------------------
Similar startups:
1): www.cetas.net  |   Similarity: 0.37931138217309823
Founded in 2010 and headquartered in Palo Alto, California, Cetas Software provides real-time big da
ta analytics solutions to extract actionable insights for online businesses and enterprises to get i
nstant recommendations, summarizations, segmentations and predictions from behavioral, social, locat
ional and mobile data.

2): www.luxurygaragesale.com  |   Similarity: 0.22682465417712583
Established in 2010 and based in Chicago, Illinois, Luxury Garage Sale operates as boutique retail a
nd digital store that sells new and used designer clothing and accessories.

3): www.plumdistrict.com  |   Similarity: 0.21053130467574566
Founded in 2010 and based in California, US, Plum District, Inc. operates an online marketplace for
mom-oriented deals and coupons.

4): www.bluecava.com  |   Similarity: 0.18940057205204563
Founded in 2010 and based in Irvine, California, BlueCava provides device identification technology
providing information about good, bad, and historical activities. It is used in fraud management for
 payment processing, banking, and e-commerce, cloud computing, digital rights management and softwar
e activation applications.

5): www.pipedrive.com  |   Similarity: 0.14042390489377066
Founded in 2010 and based in New York, US, Pipedrive Inc. operates as a provider of a web-based cust
omer relationship management and sales pipeline management software that helps enterprises to close
deals and to manage sale processes.
```

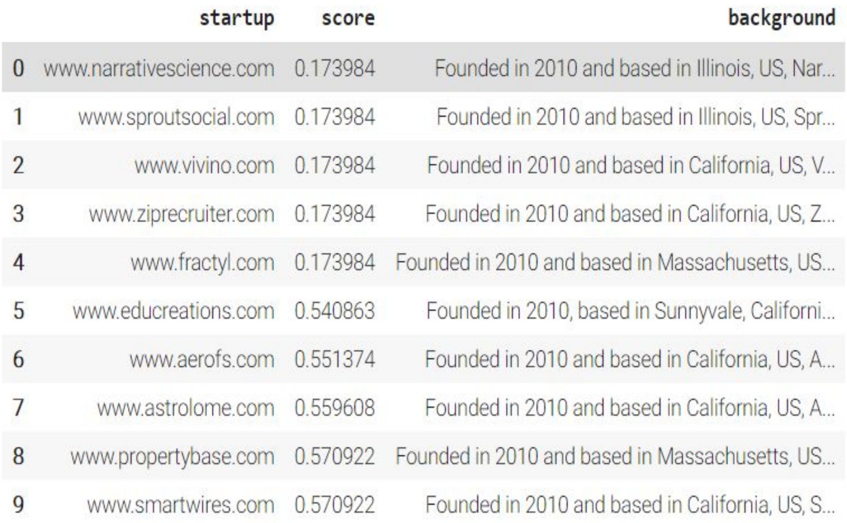
Fig13: Top 5 Similar Start-ups using TF-IDF



```
Example startup: www.6fusion.com
Founded in 2010 and based in North Carolina, US, 6fusion Inc. provides IT infrastructure solutions a
nd marketplace.
-----------------------------------------------------------------
Similar startups:
1): www.lockpath.com  |   Similarity: 0.9894917358462149
Founded in 2010 and based in Kansas, US, Lockpath, Inc. operates as a provider of governance, risk m
anagement, regulatory compliance (GRC) and information security (InfoSec) software and applications
to various sizes of organizations.

2): www.resilinc.com  |   Similarity: 0.9841915704781906
Founded in 2010 and based in California, US, Resilinc Corporation provides supply chain management a
nd analytics solutions.

3): www.elasticintelligence.com  |   Similarity: 0.9831580791172034
Founded in 2010, Elastic Intelligence provides various reporting and analytics tools to SaaS provide
rs and end users.

4): www.urjanet.com  |   Similarity: 0.9827449599007961
Founded in 2010 and based in Georgia, US, Urjanet, Inc. operates as a provider of cloud-based softwa
re solutions that enables users to pay the utilities like electricity, water, cable, telecom, waste,
 and natural gas on time.

5): www.2ndwatch.com  |   Similarity: 0.9825779371741897
Founded in 2010 and based in Washington, US, 2nd Watch Inc. operates as a provider of consulting and
 managed cloud services for business enterprises. The company's services include cloud migration, cl
oud-native and DevOps, application optimization, security and compliance, and managed cloud.
```

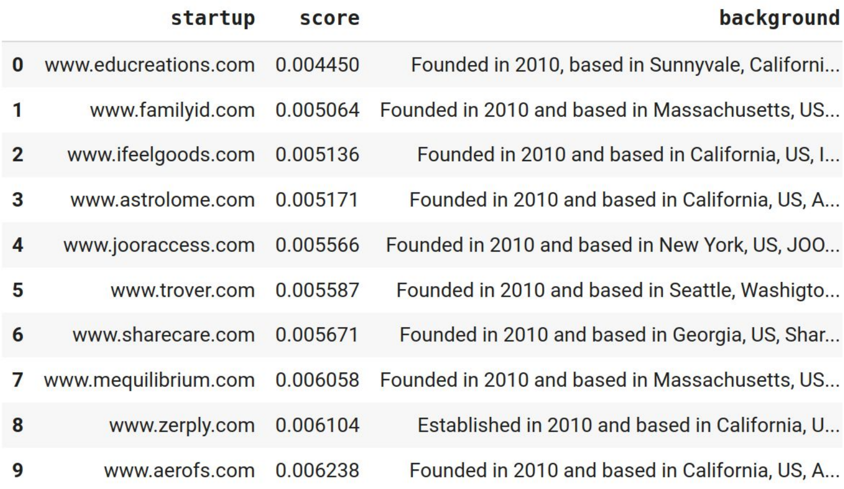
Fig14: Top 5 Similar Start-ups using Word2Vec

- Top 10 strategy score for a given company with TF-IDF and Word2Vec



	startup	score	background
0	www.narrativescience.com	0.173984	Founded in 2010 and based in Illinois, US, Nar...
1	www.sproutsocial.com	0.173984	Founded in 2010 and based in Illinois, US, Spr...
2	www.vivino.com	0.173984	Founded in 2010 and based in California, US, V...
3	www.ziprecruiter.com	0.173984	Founded in 2010 and based in California, US, Z...
4	www.fractyl.com	0.173984	Founded in 2010 and based in Massachusetts, US...
5	www.educreations.com	0.540863	Founded in 2010, based in Sunnyvale, Californi...
6	www.aerofs.com	0.551374	Founded in 2010 and based in California, US, A...
7	www.astrolome.com	0.559608	Founded in 2010 and based in California, US, A...
8	www.propertybase.com	0.570922	Founded in 2010 and based in Massachusetts, US...
9	www.smartwires.com	0.570922	Founded in 2010 and based in California, US, S...

Fig15: Strategy Score with TF-IDF



	startup	score	background
0	www.educreations.com	0.004450	Founded in 2010, based in Sunnyvale, Californi...
1	www.familyid.com	0.005064	Founded in 2010 and based in Massachusetts, US...
2	www.ifeelgoods.com	0.005136	Founded in 2010 and based in California, US, I...
3	www.astrolome.com	0.005171	Founded in 2010 and based in California, US, A...
4	www.jooraccess.com	0.005566	Founded in 2010 and based in New York, US, JOO...
5	www.trover.com	0.005587	Founded in 2010 and based in Seattle, Washigto...
6	www.sharecare.com	0.005671	Founded in 2010 and based in Georgia, US, Shar...
7	www.mequilibrium.com	0.006058	Founded in 2010 and based in Massachusetts, US...
8	www.zerply.com	0.006104	Established in 2010 and based in California, U...
9	www.aerofs.com	0.006238	Founded in 2010 and based in California, US, A...

Fig16: Strategy Score with Word2Vec

IV. Next Steps

We are currently at the 40% mark of our timeline project. So far, we have gained a deep insight into the universe of companies that we are studying, built a functional version of a web scraper code, and run a successful pilot test on an AWS server. Furthermore, we have implemented two NLP methodologies, TF-IDF, and Word2Vec, to determine the similarity and strategic measure for a small sample of startups. Nevertheless, some challenges lie ahead for the next steps of the project.

- A larger-scale web-scraping. We have the intention to web-scrape the entire universe of startups and public companies, which can be computationally demanding. To achieve this task, we plan to implement a parallel computing procedure to scale up our code.
- More NLP techniques. To find the optimal NLP technique for our task, we will continue the research and implementation of NLP techniques to create similarity measures, and compare its performance. So far, we have implemented TF-IDF and Word2Vec for the word embeddings and cosine similarity for the similarity metric; however, we believe that other algorithms such as GLOVE and BERT could constitute interesting paths to explore. Furthermore, the implementation of the NLP algorithms on the entire sample of companies will be a technical challenge in terms of the use of hardware.
- A new design of the strategy score and following performance prediction models. To effectively measure the strategic positioning, the design of strategy score is critical and its influence can be profound. The current strategy score only considers the similarity in text across several of the most similar companies, which, however, ignores the fact that some successful strategies do share some similarities. Based on the new strategy score, we can explore some Machine Learning models such as linear regression, random forest, logistic regression, boosting, which could predict the performance as a dependent variable of the strategy score of the startup. Moreover, information, such as geographical location, can be incorporated into the modeling process to build a more accurate predictor of the performance of the startups in the early rounds of funding.

Github

The code for all the sections of the project is available in a Github repository https://github.com/derekcoding1/StartupStrategy

- **EDAV:** EDAV Startups _Public/Exploratory Analysis.ipynb
- **Web scraping**: Webscraping_V2_with_Table.py & Web Scraping Public Companies V1.py
- **NLP:** nlp/word2vec.ipynb

Reference

Anthony, C., Nelson, A. J. & Tripsas, M. (2016), *'"who are you?. . . i really wanna know": Product meaning and competitive positioning in the nascent synthesizer industry'*, Strategy Science 1(3), 163–183.

Felin, T., Gambardella, A., Stern, S. & Zenger, T. (2019), *Lean startup revisited*, Technical Report 26278, Medium.
URL: https://medium.com/@teppofelin/lean-startup-revisited-c81fb8719614

Gans, J., Scott, E. & Stern, S. (2019), *Entrepreneurial Strategy*, Manuscript.

Guzman, J., & Li, A. (2019). Measuring Founding Strategy. *Available at SSRN 3489585*.

Guzman, J. & Stern, S. (2015), *'Where is silicon valley?'*, Science 347(6222), 606–609.

Guzman, J. & Stern, S. (2019), *'The state of american entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 32 us states, 1988-2014'*, American Economic Journal: Economic Policy .

Porter, M. E. (1980), *Competitive strategy: Techniques for analyzing industries and competitors,* Simon and Schuster.

Porter, M. E. (1996), 'What is strategy?', Harvard Business Review 6(74), 61–78.

Reis, E. (2011), *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses,* Currency.