Measuring startup strategy and its evolution

Capstone Project

By

Derek Chen - cc4506

Wangzhi Li - wl2737

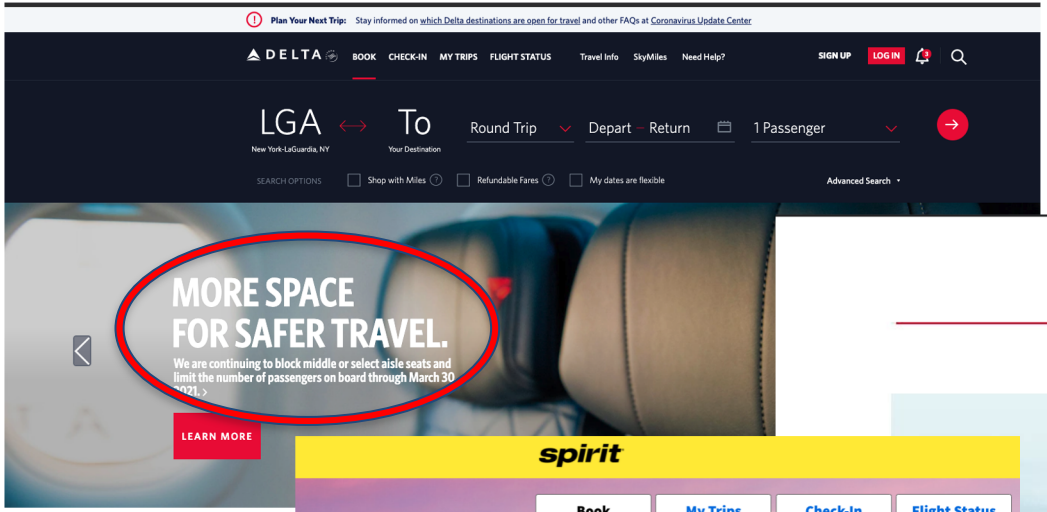
Bernardo Lopez Vicencio - bl2786

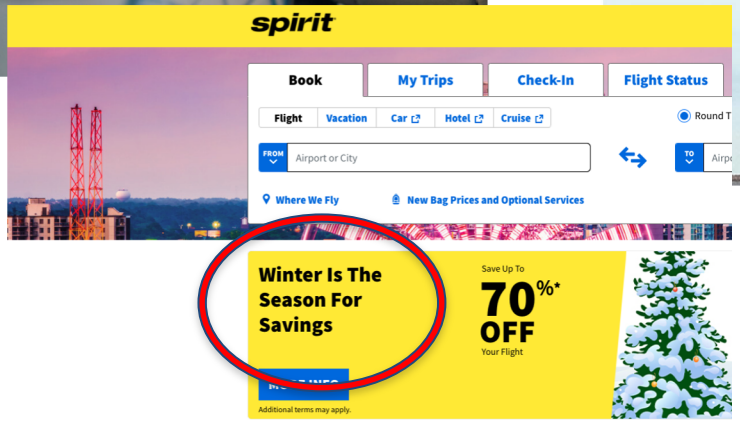
Yinhe Lu - yl4372

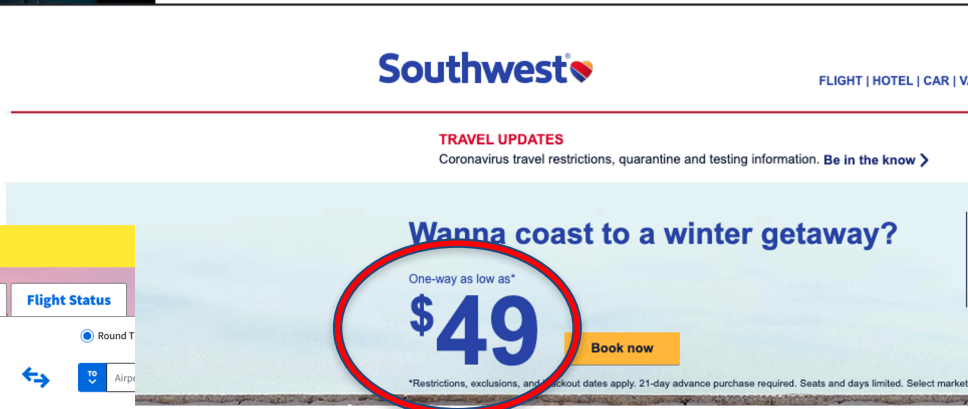
Alberto Munguia Cisneros - am5334

Introduction

What can we tell about a company from its website?







A key question to ask

"Competitive strategy is about being different."

-- Michael E. Porter

How can we quantitatively measure the strategic positioning?

Our 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.

- The expected outcome of the project will be reproducible code and the improvement of the early version of the paper “Measuring Founding Strategy.” by Prof. Jorge Guzman and Aishen Li, which is the base of this project.

Concept Overview

Extraction of initial statements and relevant information from a representative sample of startups and public companies

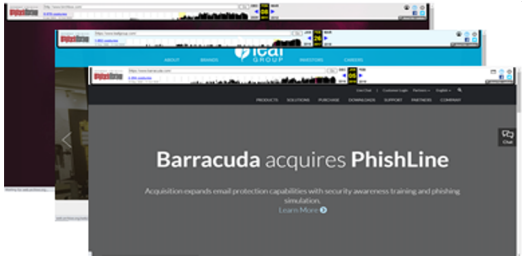
Data standardization and cleaning to eliminate noisy and atypical data

Exploration of Natural Language Processing (NLP) techniques to determine the similarity of statements between companies and build a strategy score

Exploration of different performance variables and elaboration of statistical analysis to determine the predictive power of the strategy score on startup performance

Concept Overview

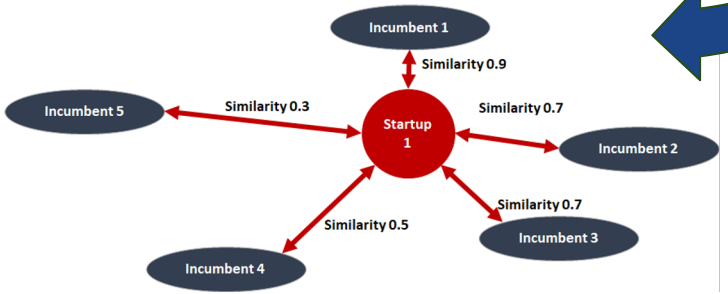
Web-Scraping



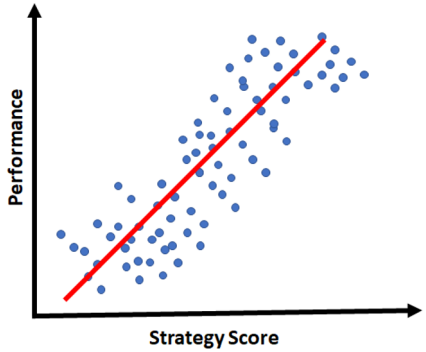
Data Standardization and Cleaning

	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 soonl 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, sirl 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/gddfrQ 4:23 PM Dec 30th via Social Flow @slannn 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"

Strategy Score



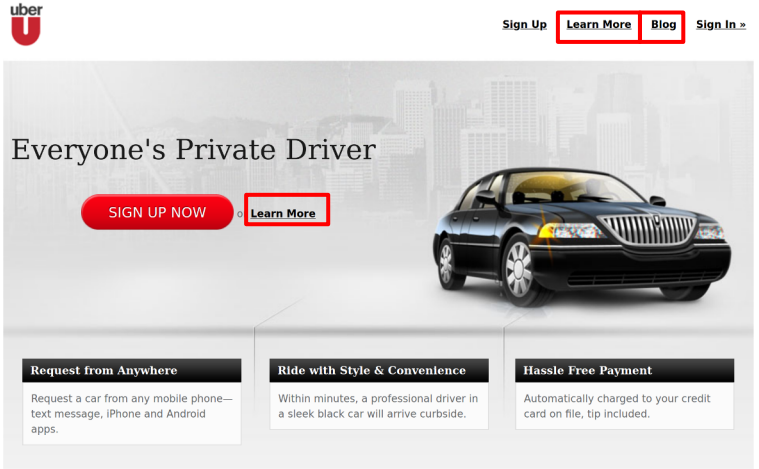
Performance Prediction



Data Sources and Web-scraping

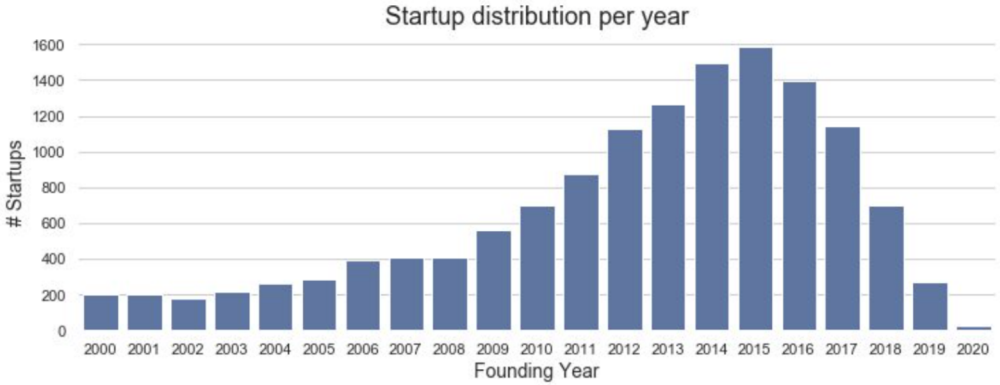
Wayback Machine

- Snapshot of target year
- Available links on level one depth

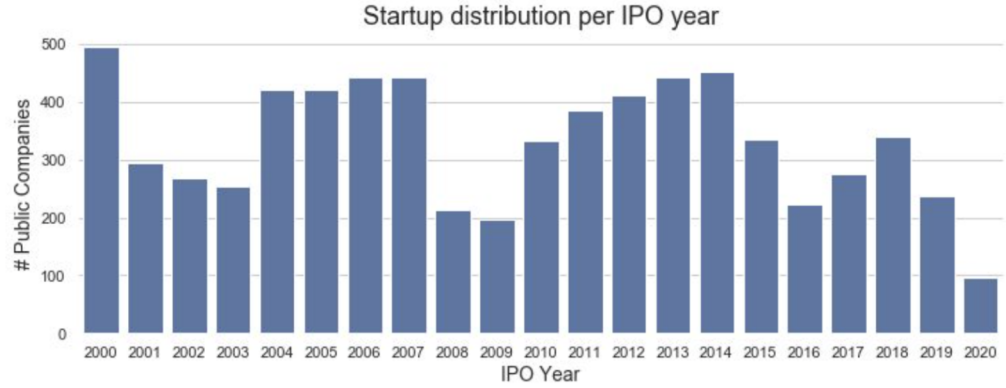


Data Sources

- 13,704 unique startups(2000-2020)
- Database from Preqin
- Information sector is the dominance

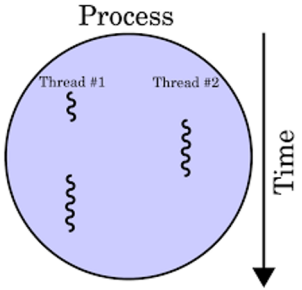


- 22,333 public companies
- Database from Orbis
- Financial segment is the dominance



Web-Scraping Process

- EC2 worker with 12GB RAM
  → 4 virtual machine with extended RAM,25GB
- Multi-threading technique
- **218.6 days** of continuous processing
  → finished collecting the entire universe of companies(more than 70 GB) in **two weeks**



Type	# companies	Estimated time processing(hrs)	Estimated time processing (days)
Startups	13,677	174.8	7.3
Public Companies	396,896	5,071.4	211.3
Total	**410,573**	**5,246**	**218.6**

Size sample	Estimated time process(hrs)		Performance (times)
	Initial	Multithreading	
200	2.6	0.6	4.6
1,500	19.2	0.9	21.7

Data processing and result

- **Data structure**
  Concentrate all text in one single CSV file for each year.

- **Language Consistency**
  Use the compact language detector v3 from Google to detect the language and filter those text that were not English.

- **Outliers**
  Drop all companies whose website produces a text with more than 1 million words.

Year	Total # of companies	# of startups	# of public companies	# of websites in english	# of websites in other languages
2000	221	49	172	205	16
2001	3975	51	3924	3755	220
2002	4055	44	4011	3865	190
2003	4656	104	4552	4404	252
2004	4655	142	4513	4457	198
2005	3680	147	3533	3562	118
2006	5357	181	5176	5202	155
2007	5656	221	5435	5475	181
2008	5237	211	5026	5068	169
2009	5390	294	5096	5215	175
2010	6292	430	5862	5995	297
2011	6304	539	5765	6137	167
2012	6744	768	5976	6524	220
2013	7331	894	6437	7106	225
2014	7623	1014	6609	7334	289
2015	7815	1079	6736	7532	283
2016	6803	885	5918	6549	254
2017	7442	842	6600	7163	279
2018	5361	507	4854	5160	201
2019	5676	231	5445	5457	219

Strategy Score

Strategy Score

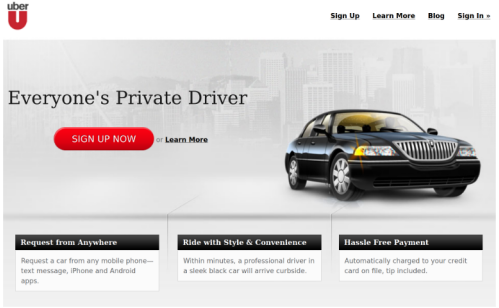
Professor Guzman defined the ***strategy score*** as a measure of **distance** between the **strategic positioning** of companies at their **foundation**.

$\hat{S}_i$

High strategy score

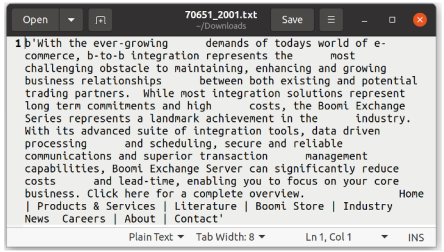
High expected profit

Challenge:



Numeric score

From text to the strategy score



Text

$$A = (w_1, w_2, \ldots, w_n)$$

Word embeddings

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity between each pair of companies

$$d_{A,B} = 1 - similarity(A,B)$$

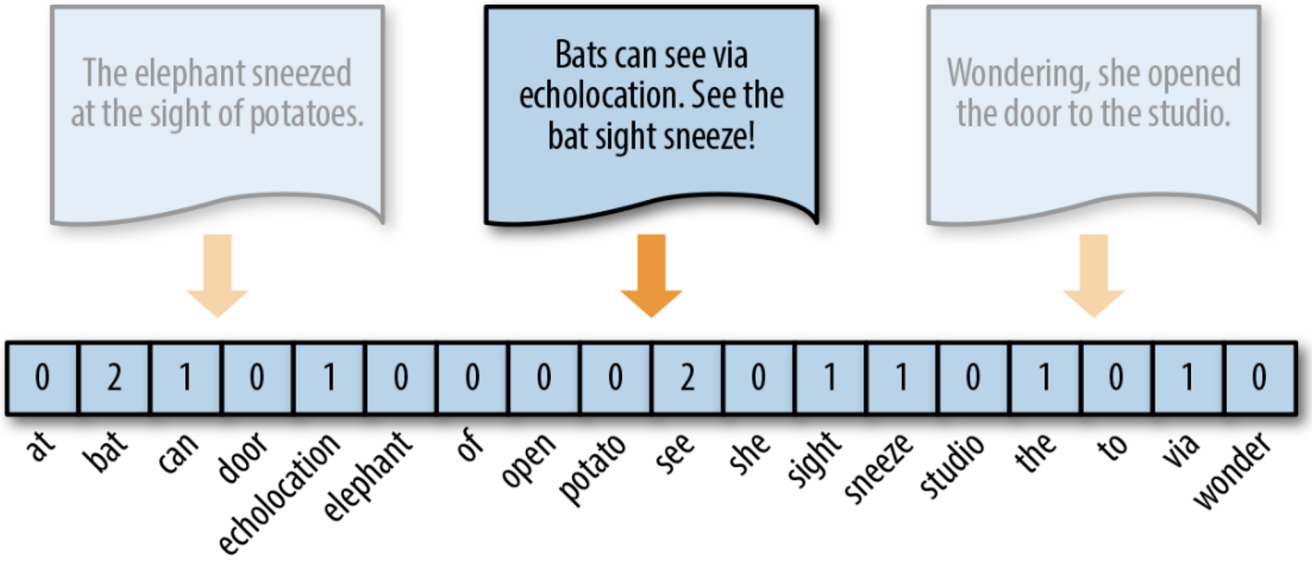
Get the distances

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

Compute the strategy score for each startup using the closest companies

Word Embeddings

- Word embedding: words or phrases from the vocabulary are mapped to vectors of real numbers



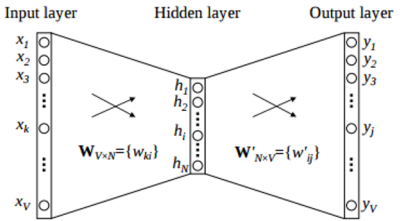
Word Embeddings (cont'd)

tf-idf

Assigns a weight to every word:

tf-idf(w) = term_frequency(w) x inverse_document_frequency(w)

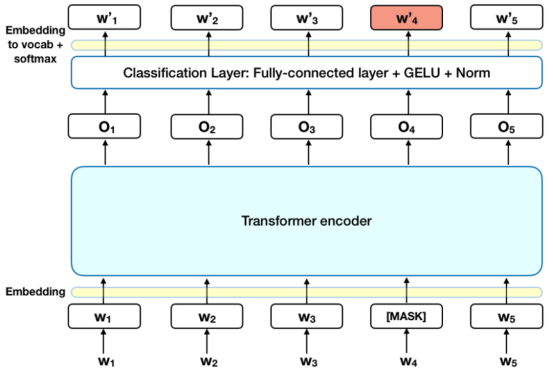
word2vec



word2vec model architecture

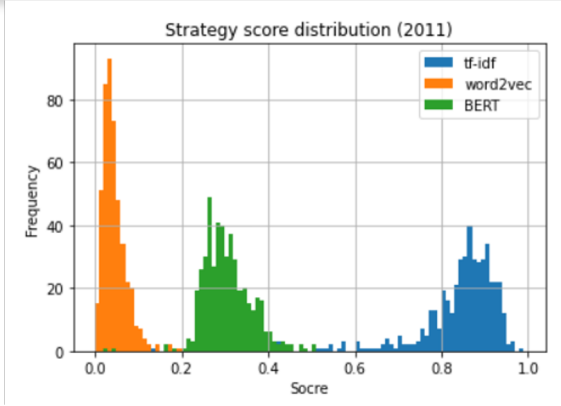
BERT

BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers



Strategy Score- Comparison

- Advantages of TF-IDF-based strategy score:
  - The highest variance
  - Simplicity
    - As a baseline model
    - Shortest running time
  - Served as a comparison with Prof. Guzman's previous work
  - Significantly correlated with Word2Vec-based and BERT-based strategy score.



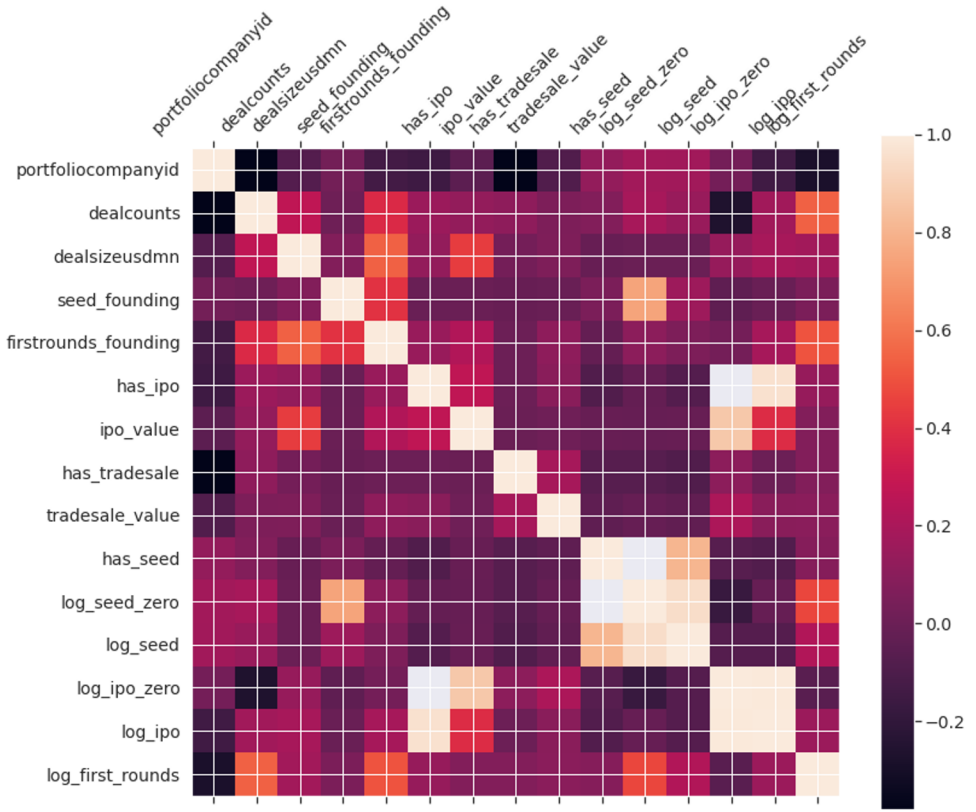
	tf-idf	word2vec	Bert
Mean	0.833	0.045	0.305
Variance	0.013	0.001	0.003

	TF-IDF	*word2vec*	*bert*
TF-IDF	-	*0.22****	*0.26****
word2vec	*0.22****	-	*0.36****
bert	*0.26****	*0.36****	-

Performance prediction

Design of Dependent Variables

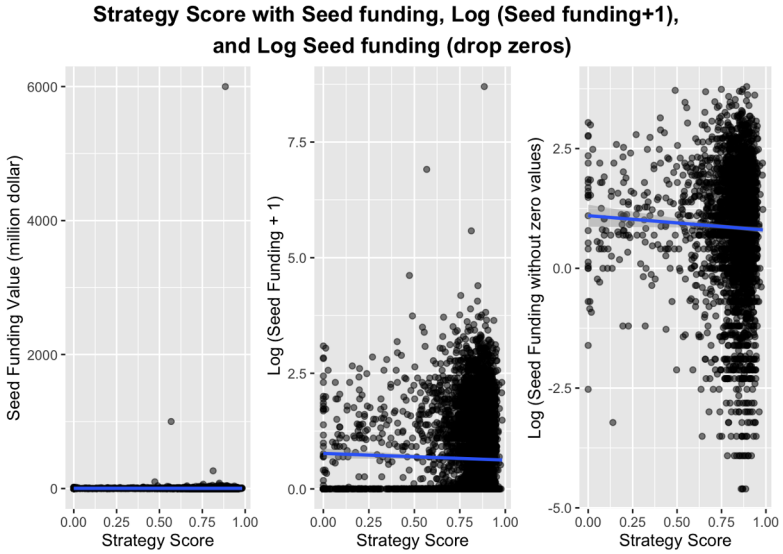
- Continuous Variables (including log version)
  - Seed Funding
  - First Rounds of Funding
  - IPO value
  - Trade Sale value
- Binary Variables
  - Has Seed Funding
  - Has IPO
  - Has Trade Sale



Detailed definition of each dependent variable can be found in the Appendix

Regression Results

- Seed funding
  - Our linear regression model with fixed-effects shows a **positive** relationship between our strategy score and the value of seed funding

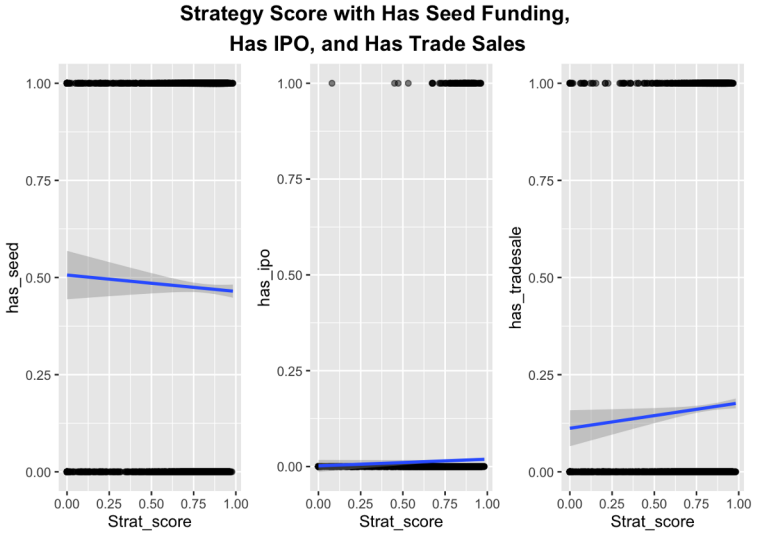


Linear Regression: Log(Seed founding + 1)

	Dependent variable:			
	log_seed			
	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)
Strat_score	-0.144**	-0.004	0.030	0.022
	(0.066)	(0.084)	(0.044)	(0.049)
Score Year F.E.	No	Yes	Yes	Yes
City F.E.	No	No	Yes	Yes
Industry F.E.	No	No	No	Yes
Observations	7,463	7,463	7,463	7,463
R^2	0.001	0.086	0.169	0.178
Adjusted R^2	0.0005	0.084	0.081	0.090
Residual Std. Error	0.866 (df = 7461)	0.830 (df = 7442)	0.831 (df = 6750)	0.826 (df = 6741)
Note:				*p<0.1; **p<0.05;** p<0.01*

Regression Results

- Has Seed, Has IPO, and Has Trade Sales
  - Our logistic regression models show a significantly **positive** relationship between strategy score and good outcomes of startups.



Logistic Regression: Has Seed, Has IPO, and Has Trade sales

	Dependent variable:		
	has_seed	has_ipo	has_tradesale
	(1)	(2)	(3)
Strat_score	-0.042	0.017*	0.065**
	(0.038)	(0.010)	(0.028)
Constant	0.506***	0.002	0.112***
	(0.032)	(0.008)	(0.024)
Observations	7,463	7,463	7,463
Log Likelihood	-5,405.308	4,945.220	-3,193.073
Akaike Inf. Crit.	10,814.620	-9,886.439	6,390.146
Note:			*p<0.1; **p<0.05;** p<0.01

Conclusion and Next Steps

Conclusion and Next Steps

Conclusion

- Multi-threading and concurrency techniques can significantly speed up the web scraping process.
- Compared with other text embedding methods, TF-IDF show advantages in simplicity, variance, and representativeness.
- TF-IDF-based strategy score shows a positive relationship with seed funding and good outcomes of startups, such as IPO and trade sales.

Next Steps

- Measure performance using Word2vec-based and BERT-based strategy scores
- Investigate the unexplained part of TF-IDF-based model
- Explore new strategy score definition

Appendix

- Definition of Dependent Variables
  - Seed Funding
    - Seed_founding: Million of dollars invested in seed funding.
    - Log_seed: logarithm of (Seed_founding variable + 1)
    - Log_seed_zero: logarithm of non-zero values in Seed_founding variable
  - First Rounds of Funding
    - firstrounds_founding: Million of dollars invested in the seed funding and in the first four rounds (i.e. Series A, Series B, Series C, and Series D).
    - log_first_rounds: logarithm of (firstrounds_founding variable + 1)
  - IPO value
    - ipo_value: Value of the company (in millions of dollars) when it IPO
    - log_ipo: logarithm of (ipo_value variable + 1)
    - log_ipo_zero: logarithm of non-zero values in ipo_value variable
  - Trade Sale value (tradesale_value): Value of the company (in millions of dollars) when it was sold.
  - Has Seed Funding (has_seed): Binary variable that indicates if a startup has seed funding.
  - Has IPO (has_ipo): Binary variable that indicates if a startup has IPO.
  - Has Trade Sale (has_tradesale): Binary variable that indicates if a startup has an exit of type “Trade Sale”. These are startups that have been sold.