Measuring Startup Strategy and Its Evolution

Wen Fan, Zhiyi Guo, Yujie Wang,
Fan Wu, Xu Xu
Mentor: Professor Guzman

Presentation Video: https://youtu.be/9Q9REVIfM9k

Project Introduction

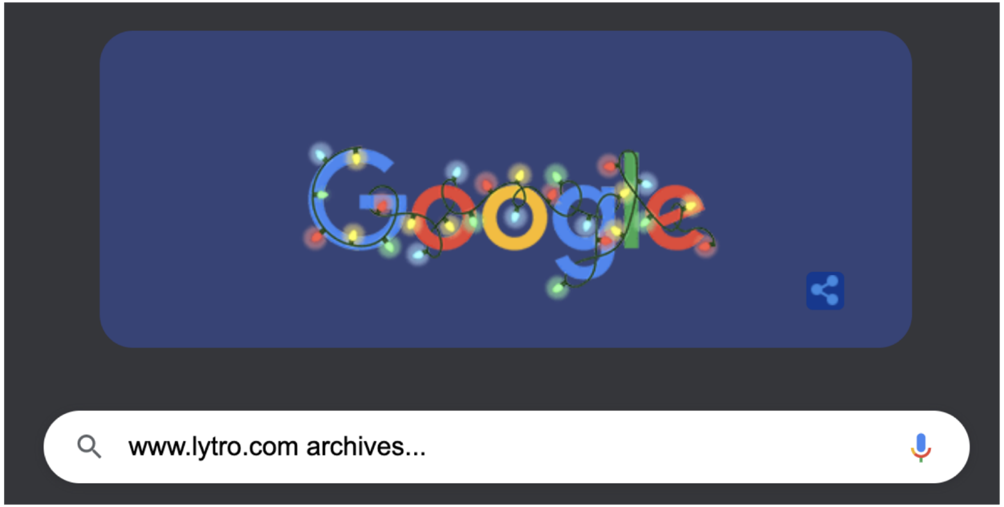
‘Pivots’

Time stamps of company’s website over time

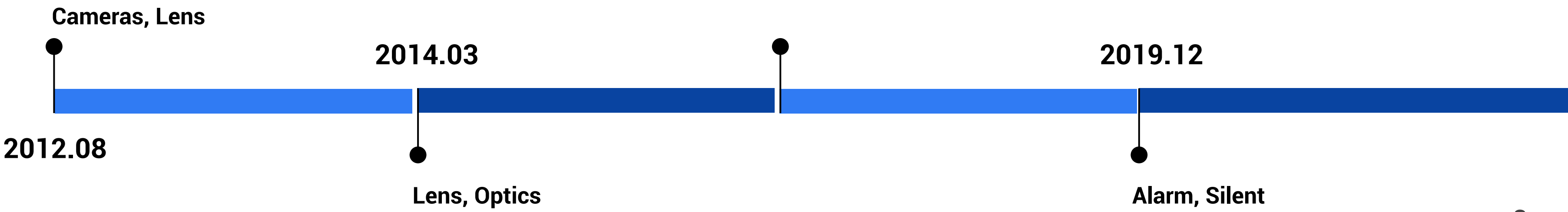
Goal

- Identify its turning points in strategies along the path
- Analyze the evolution of a company since its founding

Concept



Given the screenshots of the website over time...



Steps

Data Preprocessing	Modeling of Topics	Metrics of Change	Model of Choice	Distribution
• Scrape data from WayBackMachine.com • Convert HTML to text • Segmentation, cleaning, normalization, lemmatization	• Model for text processing: bag-of-words • Model for topics analysis: LSA, LDA	Need to find a metric to detect changes • Cosine Similarity • Jensen-Shannon • Topics over time	LDA with three metrics combined: turning points picked by at least two models, validated using 50 companies	Distribution of 600 companies

Data Preprocessing

- 13704 companies with founding dates
- Scraped monthly screenshots of homepage from WayBackMachine.com
- Downloaded data was in HTML form, convert into text
- Segmentation, cleaning, normalization, lemmatization
- Bag-of-words model

Modeling

Bag-of-words Model

Topic Models

- ❑ Latent Semantic Analysis (LSA)
- ❑ Latent Dirichlet Allocation (LDA)

Bag-of-words Model

- Bag of Words ：
  - Count of word occurrences in the document
- TF-IDF model ：
  - Importance of words

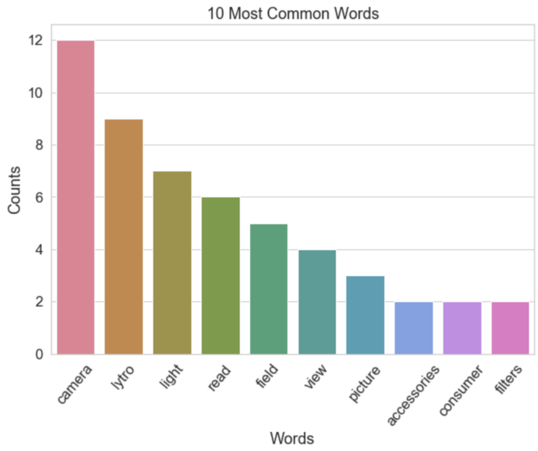
occurrences of term in document

total documents

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_j}$$

tf-idf score

documents containing word



Latent Semantic Analysis (LSA)

- Unsupervised model
- Assume that similar topics make use of similar words and each document is composed of several topics
- Build a document-term matrix for each company's website text data for each month
- Dimensionality reduction using singular value decomposition (SVD)
- Dense representation of semantic features that we need to derive possible topics
- Give us important topics and words for our text analysis
- Disadvantage: lacks interpretable embeddings and have lower accuracies than LDA model

Latent Dirichlet Allocation (LDA)

- ❑ Unsupervised model
- ❑ Generates topic distributions to each monthly website data to find out which topic is close to the website information
- ❑ Then, generates word distributions to each topic to see which word contributes most to the topic
- ❑ Tries various topics to improve accuracy
- ❑ Has distribution outputs which makes easy to compare document similarity or make recommendations

	topic_words	timestamp
50	picture light field javascript enabled view free	201401
51	lytro online learn free store apple shipping	201404
52	bundle photo lytro new camera app sharing	201404
53	light field lytro picture javascript enabled view	201404
54	lytro picture way capture dimension deeper des...	201407

Sample LDA results of lytro.com

Metrics

Cosine Similarity

Jensen-Shannon Similarity

Topics Over Time

Cosine Similarity

- ❑ Treated the topic with its important key words as a non-zero vector
- ❑ Computed the inner product with the previous month to get the measure of similarity between months.
- ❑ 0 indicates totally different; 100 indicates identical
- ❑ If similarity score is lower than the threshold, we treat the month as a turning point
- ❑ Run our model with different threshold values to tune similarity threshold parameter

month	similarity
200810	100.000000
200811	100.000000
201005	100.000000
201008	100.000000
201102	0.000000
201103	7.715167
201107	75.865814
201110	27.399831
201201	74.456944

Jensen-Shannon Distance Similarity (JSD)

- ❑ Measures the divergence between two distributions rather than vectors
- ❑ Exact outputs from the LDA model
- ❑ Symmetric, more stable, eliminates potential errors
- ❑ 0 indicates two distributions are the same; 1 indicates they are totally unlike
- ❑ Need to find large JSD in our case
- ❑ Compute JSD between any two adjacent dates
- ❑ Any pair whose JSD is larger than 0.83 will be marked as a turning point

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

JSD mathematical definition

origin	timestamp	js_distance
201204	201207	0.830424
201404	201407	0.831697
201407	201410	0.831928

Sample JSD results of lytro.com

Topics Over Time

- ❑ Find dominant topic for each month
- ❑ Find the turning points

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	dominant_topic
201304	0.010000	0.780000	0.010000	0.010000	0.130000	0.010000	0.010000	0.010000	0.010000	0.010000	1
201307	0.010000	0.900000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	1
201310	0.010000	0.920000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	1
201401	0.010000	0.920000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	1
201404	0.010000	0.920000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	1
201407	0.010000	0.080000	0.010000	0.010000	0.010000	0.010000	0.860000	0.010000	0.010000	0.010000	6
201410	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.920000	0.010000	0.010000	0.010000	6
201501	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.920000	0.010000	0.010000	0.010000	6
201504	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.880000	0.010000	0.010000	0.010000	6
201507	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.880000	0.010000	0.010000	0.010000	6

- ❑ Compare the changes in the topic with website homepage
- ❑ Optimize the model:
- ❑ GridSearch the best LDA model
- ❑ Run the LDA model multiple times, and keep turning
- ❑ points appeared over threshold

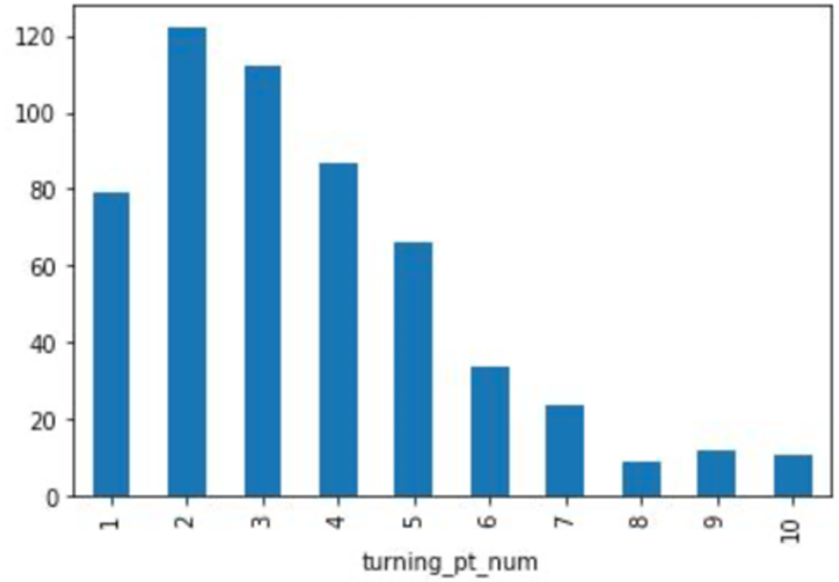
Figure. Website snapshots for 201404 and 201407

Comparison Between Models

- ❑ Apply our three models separately to 50 companies
- ❑ Check the accuracy of these results manually using WayBackMachine, label “1” and “0”
- ❑ Accuracy for none of the model is high
- ❑ Test the common turning points found by at least 2 models and calculate the accuracy
- ❑ For turning points found by exactly two models, the accuracy is 50.75%
- ❑ For turning points found by all three models, the accuracy is 61.11%
- ❑ Accuracy still not as high as we expected, but more acceptable
- ❑ Decide to combine three methods by using turning points found by at least two models

Distribution

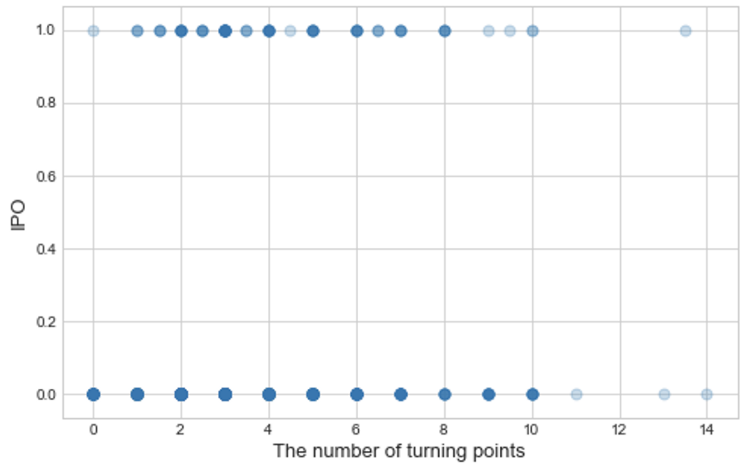
- ❑ Ran all three methods separately for the same 600 companies
- ❑ Cosine and topic-over-time have a higher percentage of overlap than cosine with Jensen and topic-over-time with Jensen
- ❑ Distribution of 600 companies' turning points is slightly skewed to the right
- ❑ Use *KstestResult* function to test normality
- ❑ The distribution is indeed not normal and is right skewed



Sample: combined results with number of turning points

IPO

- ❑ Try to explore the cause-and-effect relationship between the number of turning points and IPO
- ❑ Label 0: non-IPO
- ❑ Label 1: IPO
- ❑ Dealt with imbalanced data
- ❑ Logistic regression model: ~58% accuracy
- ❑ Decision tree model: ~68% accuracy



Thank you!