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

Goal

- Identify its turning points in strategies along the path
- Analyze the evolution of a company since its founding
Given the screenshots of the website over time...
Steps

Data Preprocessing
- Scrape data from WayBackMachine.com
- Convert HTML to text
- Segmentation, cleaning, normalization, lemmatization

Modeling of Topics
- Model for text processing: bag-of-words
- Model for topics analysis: LSA, LDA

Metrics of Change
- Need to find a metric to detect changes
  - Cosine Similarity
  - Jensen-Shannon
  - Topics over time

Model of Choice
- LDA with three metrics combined: turning points picked by at least two models, validated using 50 companies

Distribution
- 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

\[ w_{i,j} = t f_{i,j} \times \log \frac{N}{d f_j} \]

- TF-IDF formula
- Diagram showing the count of occurrences in the document and the total documents containing the 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

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

<table>
<thead>
<tr>
<th>month</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>200810</td>
<td>100.000000</td>
</tr>
<tr>
<td>200811</td>
<td>100.000000</td>
</tr>
<tr>
<td>201005</td>
<td>100.000000</td>
</tr>
<tr>
<td>201008</td>
<td>100.000000</td>
</tr>
<tr>
<td>201102</td>
<td>0.000000</td>
</tr>
<tr>
<td>201103</td>
<td>7.715167</td>
</tr>
<tr>
<td>201107</td>
<td>75.865814</td>
</tr>
<tr>
<td>201110</td>
<td>27.399831</td>
</tr>
<tr>
<td>201201</td>
<td>74.456944</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>origin</th>
<th>timestamp</th>
<th>js_distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>201204</td>
<td>201207</td>
<td>0.830424</td>
</tr>
<tr>
<td>201404</td>
<td>201407</td>
<td>0.831697</td>
</tr>
<tr>
<td>201407</td>
<td>201410</td>
<td>0.831928</td>
</tr>
</tbody>
</table>

Sample JSD results of lytro.com
Topics Over Time

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

- 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
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
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!