Feature Selection Methods For Understanding Business Competitor Relationships

Rahul Gupta¹, Jay Pujara¹, Craig Knoblock¹, Shushyam Sharanappa¹, Bharat Pulavarti¹, Gerard Hoberg¹, Gordon Phillips²
1: University of Southern California; 2: Dartmouth College

Data Science for Macro-modeling with Financial and Economic Data
6/15/18
Funded by National Science Foundation under grants 1561057 and 1561068.
What is competition?

• Products and differentiation (Hotelling, 1929)

• Production processes and industries (Pearce, 1957)

• Capital structure and financial performance (Fama & French, 1997)

• Co-occurrence in text and queries (Lee+, 2015)
Why do we care about competition?
How Does Data Science Keep Up?

- "Cloud"
- "Ridesharing"
- "Blockchain"
- Need for data-driven approaches that adapt to competition
Prior work: Text-Based Network Industry Classes

• Approach:
  – Use text from the business descriptions of SEC filings
  – Filter to remove non-noun phrases, locations, frequent terms
  – Use Jaccard similarity of text

• Drawbacks:
  – Restricted to public firms
  – SEC filings lack detail and have limited text
Web Text-Based Network Industry Classification

• Key idea: use company webpages instead of SEC filings

• Massive data collection:
  – 400K companies
  – 20 years
  – 8TB compressed text

• Developing more scalable comparison approaches

• Open question: how informative are company webpages?
Comparing SEC filings and Company Webpages

Frequency Distribution (Number of Words in Description)

200K Unique Words
Comparing SEC filings and Company Webpages

1.7M Unique Words
## Comparison of Webpage Words

<table>
<thead>
<tr>
<th>Industry</th>
<th>N</th>
<th># words (std. dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemicals</td>
<td>92</td>
<td>53K (178K)</td>
</tr>
<tr>
<td>Cons. Durables</td>
<td>78</td>
<td>38K (42K)</td>
</tr>
<tr>
<td>Cons. Nondurables</td>
<td>140</td>
<td>37K (45K)</td>
</tr>
<tr>
<td>Energy</td>
<td>156</td>
<td>22K (61K)</td>
</tr>
<tr>
<td>Finance</td>
<td>992</td>
<td>16K (26K)</td>
</tr>
<tr>
<td>Health</td>
<td>617</td>
<td>25K (27K)</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>314</td>
<td>36K (64K)</td>
</tr>
<tr>
<td>Misc</td>
<td>432</td>
<td>28K (32K)</td>
</tr>
<tr>
<td>Retail</td>
<td>310</td>
<td>68K (119K)</td>
</tr>
<tr>
<td>Tech&amp;Bus Equip</td>
<td>622</td>
<td>46K (56K)</td>
</tr>
<tr>
<td>Telecom</td>
<td>89</td>
<td>28K (21K)</td>
</tr>
<tr>
<td>All</td>
<td>3907</td>
<td>32K (60K)</td>
</tr>
</tbody>
</table>
What text should we use?

- Webpages contain all types of text, only some of which is relevant

- Terms used in SEC business descriptions are likely relevant
  - Low coverage, must be extended

- Information retrieval approaches are optimized to find relevant terms
  - High noise, must be filtered
Curated Term Lists

• Start with terms in business descriptions

• Identify frequent or discriminative terms and manually add these to a white list
  – “ethernet carrier”, “sleeper”, “tumor”

• Identify terms that are not relevant and manually add these to a black list
  – “admiralty”, “gardner”, “steinberg”

• Extract only whitelisted terms from webpage text
Term-Frequency, Inverse Document Frequency

- Use traditional information-retrieval metric for text

\[ tf(t, d) = \sum_{x \in d} fr(x, t) \]

\[ idf(t) = \log \frac{|D|}{1 + \sum_d I(t, d)} \]

\[ fr(x, t) = \begin{cases} 
1 & x = t \\
0 & x \neq t 
\end{cases} \]

\[ I(t, d) = \begin{cases} 
1 & t \in d \\
0 & \text{otherwise} 
\end{cases} \]

- Defined over entire corpus (e.g., average TF-IDF of term)
Evaluation Approach

• Data corpus of 3907 publicly traded firms with SEC business descriptions in 2015 10-K filing
• Webpages from Compustat Financial Database, use 500 webpages per company
• Predict asset-adjusted company profits using competitors

\[
\hat{F}(c_i) = \lambda F(R_i) + c
\]

\[
R^2 = 1 - \frac{\sum_i(F(c_i) - \hat{F}(c_i))^2}{\sum_i(F(c_i) - \overline{F})^2}
\]
Terms frequently used by a single company have high rankings:

- countsbaker
- geon
- ultratuf
- wilflex
- oncap

Data Issues: Proprietary Terminology

<table>
<thead>
<tr>
<th>Min Companies</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.258</td>
</tr>
<tr>
<td>3</td>
<td>0.262</td>
</tr>
<tr>
<td>5</td>
<td>0.259</td>
</tr>
<tr>
<td>10</td>
<td>0.252</td>
</tr>
</tbody>
</table>
Data Issues: Long words

- kuwait
- kyrgyzstan
- laos
- latvia
- lebanon
- lesotho
- lithuania
- luxembourg
- macau
- macedonia
- madagascar
- malawi
- malaysia
- maldives
- malta
- marshall

- apioverview
- collections
- projectsoverview
- deleteevents
- projects

- cashprovidedbyusedinoperatingactivitiesdiscontinuedoperations

<table>
<thead>
<tr>
<th>Max Length</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.262</td>
</tr>
<tr>
<td>17</td>
<td>0.284</td>
</tr>
<tr>
<td>20</td>
<td>0.286</td>
</tr>
<tr>
<td>25</td>
<td>0.285</td>
</tr>
</tbody>
</table>
Top-ranked terms by TF-IDF metric

- blog
- accessories
- clinical
- shop
- cloud
- hughes
- loans
- cards
- brands
- loan
- oil

<table>
<thead>
<tr>
<th>Top %</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.289</td>
</tr>
<tr>
<td>15</td>
<td>0.286</td>
</tr>
<tr>
<td>20</td>
<td>0.220</td>
</tr>
</tbody>
</table>
Comparing Manual and Automatic Feature Selection

<table>
<thead>
<tr>
<th>Feature Selection Method</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curated word lists</td>
<td>0.261</td>
</tr>
<tr>
<td>Filtered TF-IDF scores</td>
<td>0.286</td>
</tr>
</tbody>
</table>
Conclusion

• Competitor relationships can be difficult to define or predict

• Company-associated text often contains implicit signals of product offerings, markets, production processes, and strategic goals

• Feature selection is important for identifying the meaningful terms

• Manual feature curation works, but using automated approaches from the information retrieval community performs better