Feature Selection Methods For Understanding Business Competitor Relationships

Rahul Gupta[1], Jay Pujara[1], Craig Knoblock[1],
Shushyam Sharanappa[1], Bharat Pulavarti[1],
Gerard Hoberg[1], Gordon Phillips[2]
1: University of Southern California; 2: Dartmouth College

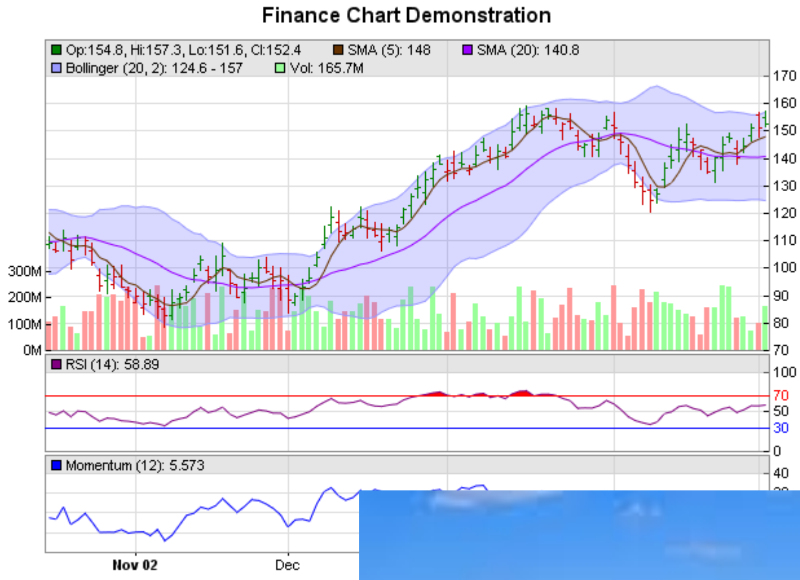
Data Science for Macro-modeling with Financial and Economic Data
6/15/18

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

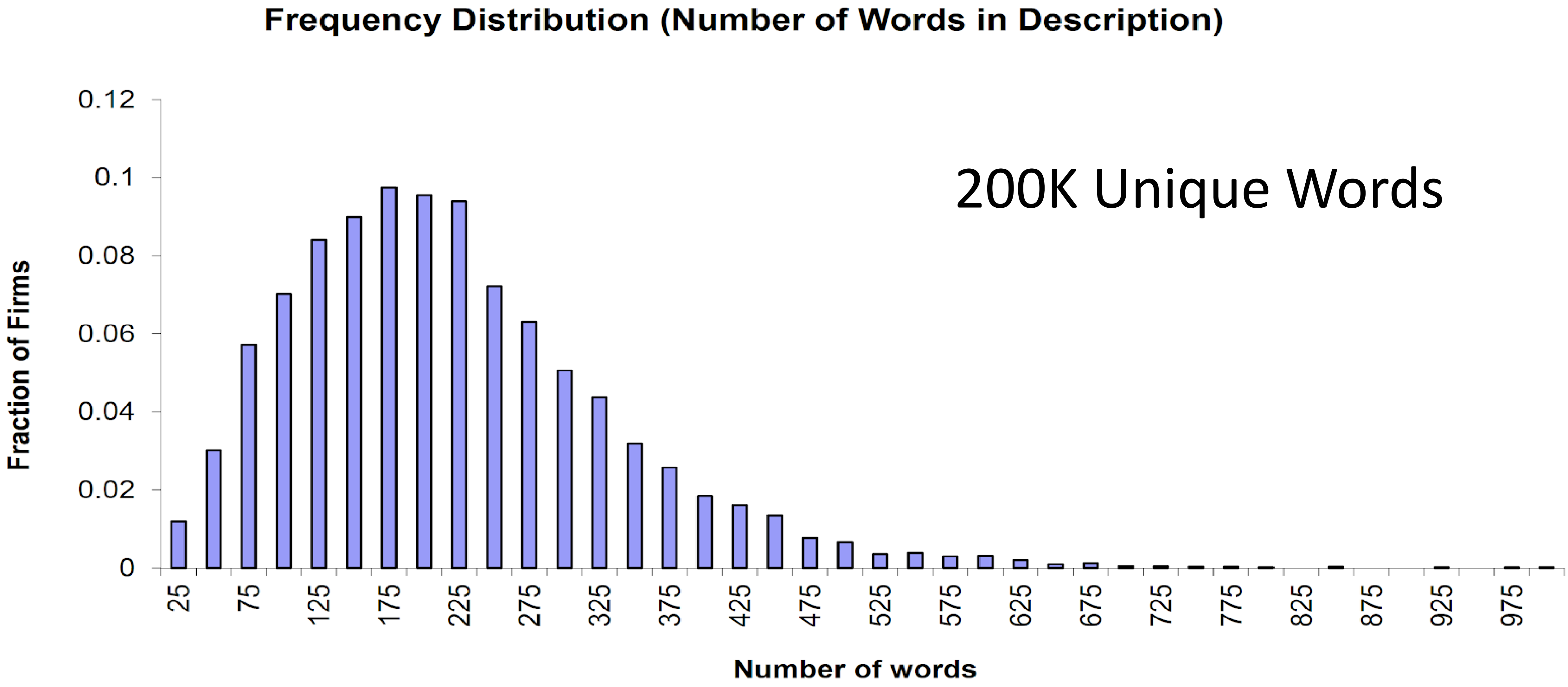
- Approach:
  - Use text from the business descriptions of SEC filings
  - Filter to remove non-noun phrases, locations, frequent terms
  - Use Jaccard similarity of text

- Drawbacks:
  - Restricted to public firms
  - SEC filings lack detail and have limited text

Web Text-Based Network Industry Classification

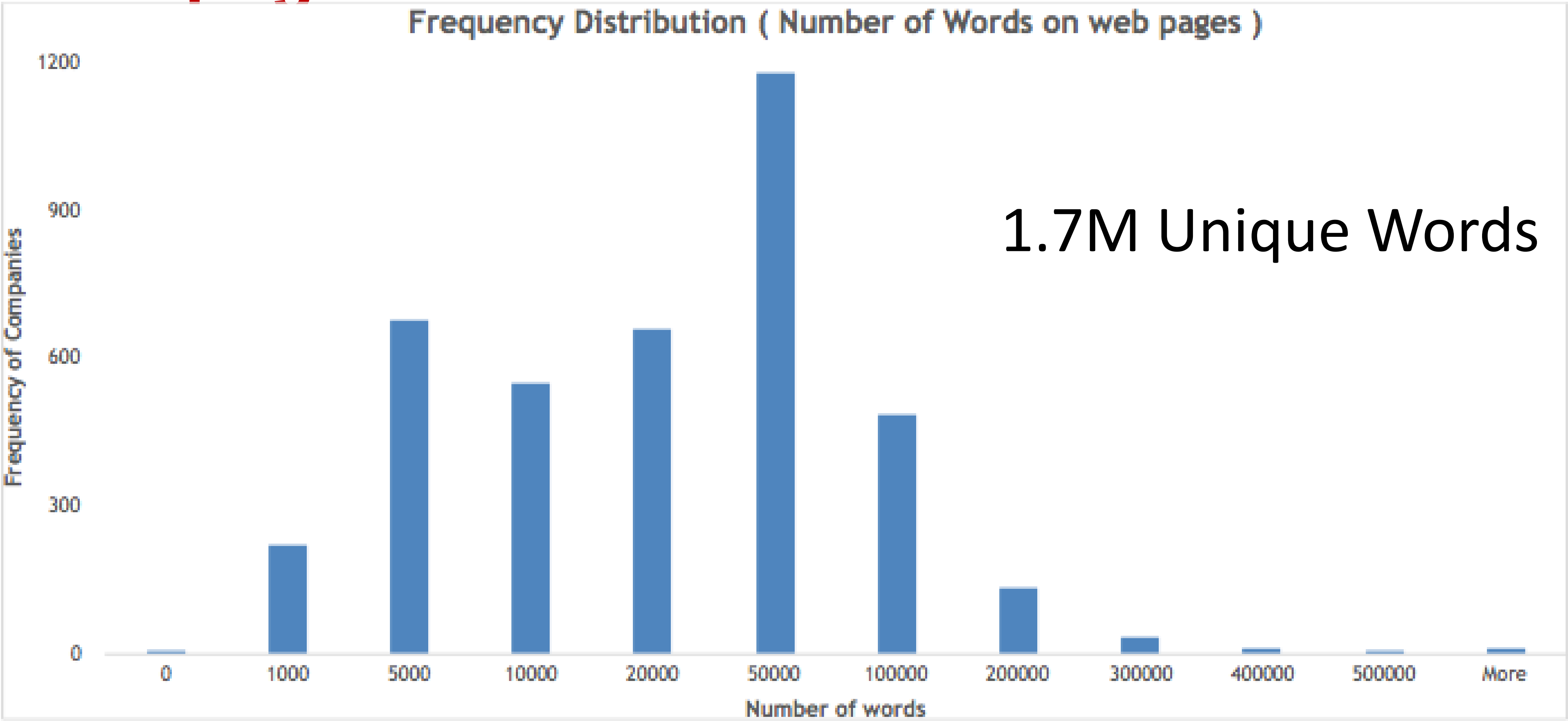
- Key idea: use company webpages instead of SEC filings
- Massive data collection:
  - 400K companies
  - 20 years
  - 8TB compressed text
- Developing more scalable comparison approaches
- Open question: how informative are company webpages?

Comparing SEC filings and Company Webpages



200K Unique Words

Comparing SEC filings and Company Webpages



Comparison of Webpage Words

Industry	N	# words (std. dev)
Chemicals	92	53K (178K)
Cons. Durables	78	38K (42K)
Cons. Nondurables	140	37K (45K)
Energy	156	22K (61K)
Finance	992	16K (26K)
Health	617	25K (27K)
Manufacturing	314	36K (64K)
Misc	432	28K (32K)
Retail	310	68K (119K)
Tech&Bus Equip	622	46K (56K)
Telecom	89	28K (21K)
All	3907	32K (60K)

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

$$tf(t,d) = \sum_{x \in d} fr(x,t) \qquad idf(t) = \log \frac{|D|}{1 + \sum_d I(t,d)}$$

$$fr(x,t) = \begin{cases} 1 & x = t \\ 0 & x \neq t \end{cases} \qquad I(t,d) = \begin{cases} 1 & t \in d \\ 0 & otherwise \end{cases}$$

- 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

$$\hat{F}(c_i) = \lambda\overline{F(R_i)} + c$$

$$R^2 = 1 - \frac{\sum_i (F(c_i) - \hat{F}(c_i))^2}{\sum_i (F(c_i) - \overline{F})^2}$$

Data Issues: Proprietary Terminology

- Terms frequently used by a single company have high rankings:
  - countsbaker
  - geon
  - ultratuf
  - wilflex
  - oncap

Min Companies	R^2
0	0.258
3	0.262
5	0.259
10	0.252

Data Issues: Long words

- kuwaitkyrgyzstanlaoslatvialebanonlesotholiberialibyaliechtensteinlithuanialuxembourgmacaumacedoniamadagascarmalawimalaysiamaldivesmalimaltamarshall
- apioverviewcollectionsprojectsoverviewdeleteeventsprojects
- cashprovidedbyusedinoperatingactivitiesdiscontinuedoperatio ns
- repaymentsofnotespayable

Max Length	R^2
None	0.262
17	0.284
20	0.286
25	0.285

Top-ranked terms by TF-IDF metric

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

Top %	R^2
10	0.289
15	0.286
20	0.220

Comparing Manual and Automatic Feature Selection

Feature Selection Method	R^2
Curated word lists	0.261
Filtered TF-IDF scores	0.286

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