Text Analysis for Economics and Finance

Ruben Durante
ICREA-UPF, BGSE, IPEG, CEPR

DIW, October 2020

Sentiment analysis with short text: VADER

- Valence Aware Dictionary and sEntiment Reasoner (VADER): lexicon-based sentiment analysis tool particularly suited for social media content
- A sentiment lexicon is a list of lexical features (e.g., words) labeled according to their semantic orientation as either positive, negative, or neutral
- It has been quite successful dealing with social media texts, newspaper editorials, movie and product reviews.
- It does not only classify text as positive, negative, or neutral, but also provides a composite score that combines all three
- Does not require any training data; constructed from a generalizable human-curated sentiment lexicon. subcategory

Hassan et al. (QJE, 2019): firm-level political risk

- Another example of application of dictionary methods
- But the construction and the use of the dictionary is innovative
- Goal: to construct a measure of political risk faced by individual US firms
- Data: 178,173 transcripts of quarterly earnings conference calls
- Idea: measure the share of the conversations between call participants and firm management centered around risks associated with political matters

Hassan et al. (QJE, 2019): firm-level political risk

- Another example of application of dictionary methods
- But the construction and the use of the dictionary is innovative
- Goal: to construct a measure of political risk faced by individual US firms
- Data: 178,173 transcripts of quarterly earnings conference calls
- Idea: measure the share of the conversations between call participants and firm management centered around risks associated with political matters

Hassan et al. (QJE, 2019): firm-level political risk

- Another example of application of dictionary methods
- But the construction and the use of the dictionary is innovative
- Goal: to construct a measure of political risk faced by individual US firms
- Data: 178,173 transcripts of quarterly earnings conference calls
- Idea: measure the share of the conversations between call participants and firm management centered around risks associated with political matters

Hassan et al. (QJE, 2019): constructing the dictionary

- Training library of political text archetypical of discussions of political topics, **P**
- Training library of non-political text, archetypical of discussion of non-political topics, **N**
- **P**: William T. Bianco and David T. Canon, *American Politics Today*
- **N**: Robert Libby, Patricia A. Libby, and Daniel G. Short's, *Financial Accounting*
- Each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P} \backslash \boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Hassan et al. (QJE, 2019): constructing the dictionary

- Training library of political text archetypical of discussions of political topics, **P**
- Training library of non-political text, archetypical of discussion of non-political topics, **N**
- **P**: William T. Bianco and David T. Canon, *American Politics Today*
- **N**: Robert Libby, Patricia A. Libby, and Daniel G. Short's, *Financial Accounting*
- Each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P} \backslash \boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Hassan et al. (QJE, 2019): constructing the dictionary

- Training library of political text archetypical of discussions of political topics, **P**
- Training library of non-political text, archetypical of discussion of non-political topics, **N**
- **P**: William T. Bianco and David T. Canon, *American Politics Today*
- **N**: Robert Libby, Patricia A. Libby, and Daniel G. Short's, *Financial Accounting*
- Each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P} \setminus \boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Hassan et al. (QJE, 2019): constructing the dictionary

- First key statistics for them is $f_{b,\boldsymbol{P}}/B_{\boldsymbol{P}}$
  - $f_{b,\boldsymbol{P}}$: frequency of bigram b in the political training library
  - $B_{\boldsymbol{P}}$ is the total number of bigrams in the political training library
  - **What is this?**
  - Relative term frequency of b in P. Similar to $tf_{i,j}$
- Second key statistics is $\mathbf{1}[b \in \boldsymbol{P} \setminus \boldsymbol{N}]$
  - Where $\mathbf{1}[\cdot]$ is an indicator function.
  - This is an extreme way of doing an idf_b across libraries. **Why?**
  - idf_b would give more weight to terms that are "special" to library $\boldsymbol{P}$, i.e. not as frequent in $\boldsymbol{N}$.
  - Here the weight is set to 0 for all terms in $\boldsymbol{P}$ that are also in $\boldsymbol{N}$.

Hassan et al. (QJE, 2019): constructing the dictionary

- First key statistics for them is $f_{b,\boldsymbol{P}}/B_{\boldsymbol{P}}$
  - $f_{b,\boldsymbol{P}}$: frequency of bigram b in the political training library
  - $B_{\boldsymbol{P}}$ is the total number of bigrams in the political training library
  - **What is this?**
  - Relative term frequency of b in P. Similar to $tf_{i,j}$
- Second key statistics is $\mathbf{1}[b \in \boldsymbol{P} \backslash \boldsymbol{N}]$
  - Where $\mathbf{1}[\cdot]$ is an indicator function.
  - This is an extreme way of doing an idf_b across libraries. **Why?**
  - idf_b would give more weight to terms that are "special" to library $\boldsymbol{P}$, i.e. not as frequent in $\boldsymbol{N}$.
  - Here the weight is set to 0 for all terms in $\boldsymbol{P}$ that are also in $\boldsymbol{N}$.

Hassan et al. (QJE, 2019): constructing the dictionary

Table 2: Top 120 political bigrams used in construction of $PRisk_{i,t}$

Bigram	$(f_{b,\mathbb{P}}/B_{\mathbb{P}}) \times 10^5$	Frequency	Bigram	$(f_{b,\mathbb{P}}/B_{\mathbb{P}}) \times 10^5$	Frequency
the constitution	201.15	9	governor and	26.79	11
the states	134.29	203	government the	26.39	56
public opinion	119.05	4	this election	25.98	26
interest groups	118.46	8	political party	25.80	5
of government	115.53	316	american political	25.80	2
the gop	102.22	1	politics of	25.80	5
in congress	78.00	107	white house	25.80	21
national government	68.03	7	the politics	25.80	31
social policy	62.16	1	general election	25.22	30
the civil	60.99	64	and political	25.22	985
elected officials	60.40	3	policy is	25.22	135
politics is	53.95	7	the islamic	25.04	1
political parties	51.61	3	federal reserve	24.63	119
office of	51.02	58	judicial review	24.04	6
the political	51.02	1091	vote for	23.46	6
interest group	48.09	1	limits on	23.46	53
the bureaucracy	48.09	1	the faa	23.28	22
and senate	46.33	19	the presidency	22.87	2
government and	44.57	325	shall not	22.87	4
for governor	41.48	2	the nation	22.87	52
executive branch	40.46	3	constitution and	22.87	3
support for	39.88	147	senate and	22.87	28
the epa	39.15	139	the va	22.65	77
civil service	27.56	2	and party	18.77	2
government policy	27.56	52	governor in	18.76	1
federal courts	27.56	1	state the	18.26	35
argued that	26.98	8	executive privilege	18.18	1
the democratic	26.98	7	of politics	18.18	4
islamic state	26.92	1	the candidates	18.18	11
president has	26.86	7	national security	18.18	59

Hassan et al. (QJE, 2019): using the dictionary

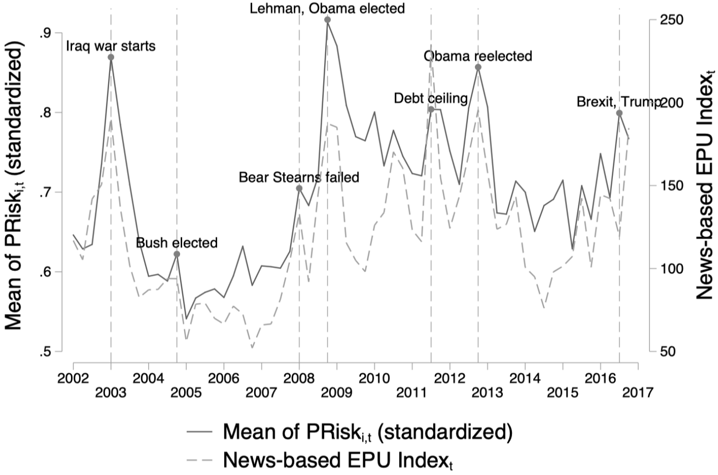
- Count the number of instances where political bigrams are used in conjunction with synonyms for "risk"
- Conference-call transcript of firm i in quarter t into a list of bigrams contained in the transcript $b = 1, ..., B_{it}$.

$$PRisk_{it} = \frac{\sum_{b=1}^{B_{it}} \left(\mathbf{1}\left[b \in \boldsymbol{P} \backslash \boldsymbol{N}\right] \times \mathbf{1}\left[|b - r| < 10\right] \times \frac{f_{b,\boldsymbol{P}}}{B_{\boldsymbol{P}}} \right)}{B_{it}}$$

- r is the position of the nearest synonim for risk or uncertainty

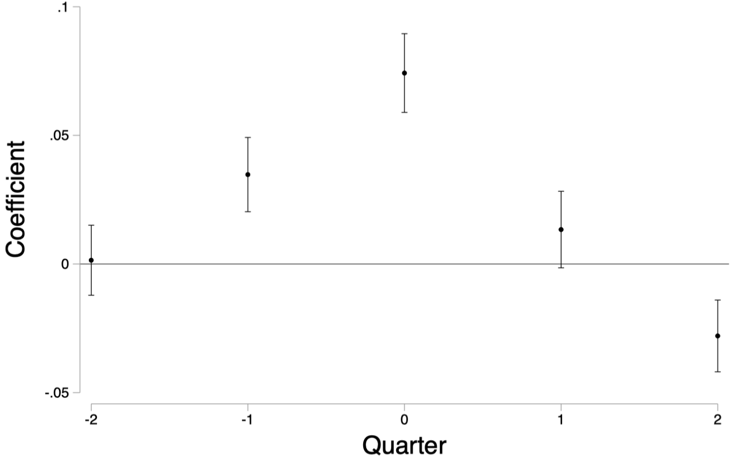
Hassan et al. (QJE, 2019): using the dictionary

Figure 1: Variation in $PRisk_{i,t}$ over time and correlation with EPU



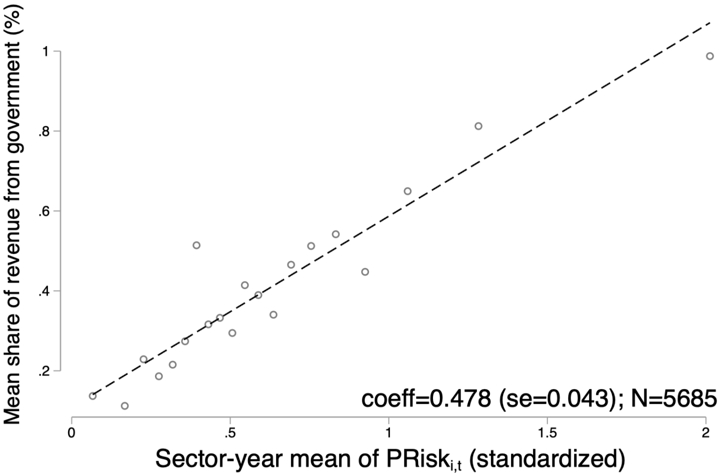
Hassan et al. (QJE, 2019): using the dictionary

Figure 2: Variation in $PRisk_{i,t}$ around federal elections

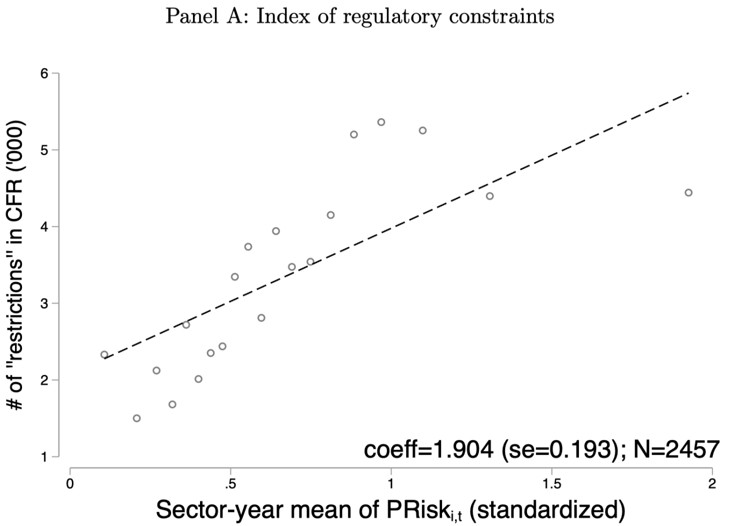


Hassan et al. (QJE, 2019): using the dictionary

Panel B: Share of revenue from federal government



Hassan et al. (QJE, 2019): using the dictionary



Documents as vectors

- In the document-feature-matrix each document is represented by a row-vector
- Each vector contains the (weighted) frequencies of each feature in the document
- Idea: these vectors can be used to measure the similarity/distance between documents

Property of distance measures

- Let A and B be any two documents in a set and $d(A, B)$ be the distance between A and B

1. $d(x, y) \geq 0$: the distance between any two points must be non-negative
2. $d(A, B) = 0$ iff $A = B$: the distance between two documents must be zero if and only if the two objects are identical
3. $d(A, B) = d(B, A)$: distance must be symmetric
4. $d(A, C) \leq d(A, B) + d(B, C)$ must satisfy the triangle inequality

Euclidean distance

Between document A and B where j indexes their features, where y_{ij} is the value for feature j of document i

- Euclidean distance is based on the Pythagorean theorem
- Formula

$$\sqrt{\sum_{j=1}^{j}(y_{Aj} - y_{Bj})^2}$$

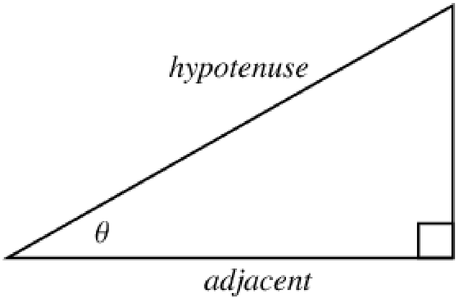
- In vector notation:

$$\|\mathbf{y}_A - \mathbf{y}_B\|$$

- Can be performed for any number of features J – the number of columns in the document-feature matrix, i.e., the number of feature types in the corpus

A geometric notion of "distance"

In a right angled triangle, the cosine of an angle θ or $\cos(\theta)$ is the length of the adjacent side divided by the length of the hypotenuse



We can use the vectors to represent the text location in a *n*-dimensional vector space and compute the angles between them

Cosine Similarity: Idea

- Each document is a non-negative vector in an *n*-space (size of the common dictionary) and it defines a *ray*
  - Closer rays form smaller angles
  - The furthest rays are orthogonal
- $cos(0) = 1$ and $cos(\pi/2) = 0$
- Distance monotonically increases on $\{0, \pi/2\}$ ->
  Cosine or similarity monotonically decreases on $\{1,0\}$
- Nice property: independent of document length, because it deals only with the angle of the vectors

Cosine Similarity: Idea

- Each document is a non-negative vector in an *n*-space (size of the common dictionary) and it defines a *ray*
  - Closer rays form smaller angles
  - The furthest rays are orthogonal
- $cos(0) = 1$ and $cos(\pi/2) = 0$
- Distance monotonically increases on $\{0,\pi/2\}$ -> Cosine or similarity monotonically decreases on $\{1,0\}$
- Nice property: independent of document length, because it deals only with the angle of the vectors

Cosine similarity: Formula

$$\text{cos\_sim}(y_A, y_B) = \frac{y_A \cdot y_B}{||y_A||||y_B||}$$

- y_A and y_B are vectors representing documents.
- The operator $\cdot$ is the dot product, or $\sum_j y_{A_j} y_{B_j}$
- $||y_A||$ is the vector norm of the features vector y for document A, such that $||y_A|| = \sqrt{\sum_j y_{A_j}^2}$
- $+1$ means identical documents; 0 means no words in common.
- Note: Using tf-idf to down-weight terms that appear in many documents usually gives better results.

Example text

Hurricane Gilbert swept toward the Dominican Republic Sunday , and the Civil Defense alerted its heavily populated south coast to prepare for high **winds**, heavy **rains** and high seas.

The **storm** was approaching from the southeast with sustained **winds** of 75 mph gusting to 92 mph .

"There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday .

Cabral said residents of the province of Barahona should closely follow **Gilbert** 's movement .

An estimated 100,000 people live in the province, including 70,000 in the city of Barahona , about 125 miles west of Santo Domingo .

Tropical **Storm Gilbert** formed in the eastern Caribbean and strengthened into a **hurricane** Saturday night

The National **Hurricane** Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north , longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

The National Weather Service in San Juan , Puerto Rico , said **Gilbert** was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the **storm**.

The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6p.m. Sunday.

Strong **winds** associated with the **Gilbert** brought coastal flooding , strong southeast **winds** and up to 12 feet to Puerto Rico 's south coast.

Example text: selected terms

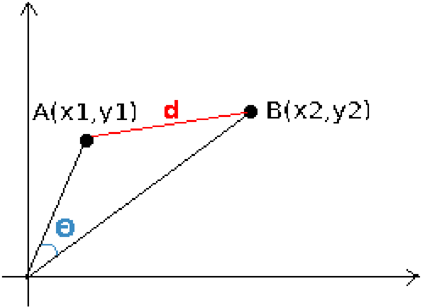
- Document 1
  Gilbert: 3, hurricane: 2, rains: 1, storm: 2, winds: 2

- Document 2
  Gilbert: 2, hurricane: 1, rains: 0, storm: 1, winds: 2

- Cosine similarity = 0.9438798

Relationship to Euclidean distance

- Cosine similarity measures the similarity of vectors with respect to the origin
- Euclidean distance measures the distance between particular points of interest along the vector

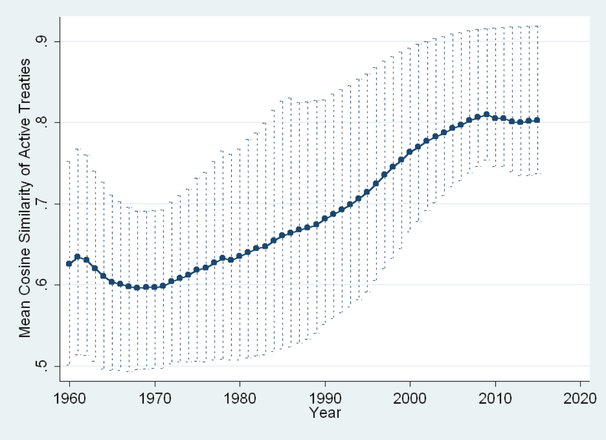


Other measures

- Edit distance refers to the number of operations required to transform one string into another for strings of equal length
- Common edit distance: the Levenshtein distance
- Example: the Levenshtein distance between "kitten" and "sitting" is 3
  - kitten → sitten (substitution of "s" for "k")
  - sitten → sittin (substitution of "i" for "e")
  - sittin → sitting (insertion of "g" at the end).
- Hamming distance: for two strings of equal length, the Hamming distance is the number of positions at which the corresponding characters are different
- Not common, as at a textual level this is hard to implement and possibly meaningless

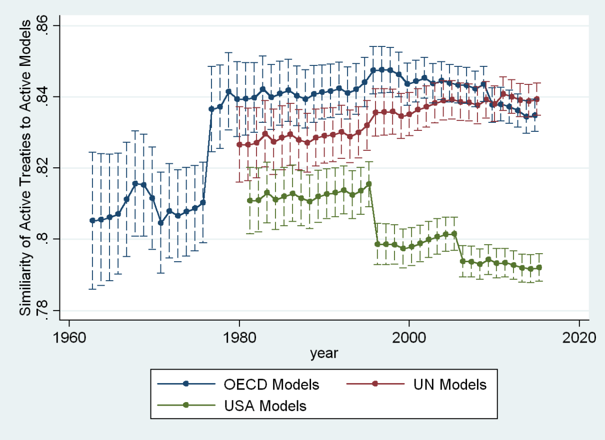
Tax Treaties have converged in language

Ash and Marian (2018)



Average cosine similarity between active treaties by year

Influence of Model Treaties over Time



- **OECD** and **UN** treaty models most influential.

Abrahamson and Barber

The Evolution of National Constitutions (QJPS 2019)

- Corpus: Comparative Constitutions Project:
  - A repository of current and historical constitutions across countries and provinces.
  - 1297 constitutions, 185 countries, 1789-2010
- Annotations (1329 features):
  - e.g. structure of executive, amendment process, election process, legislative composition

Colonial Path Dependence

Table 4: Between estimates of colonial history and constitutional similarity.

Distance from:	(1) UK	(2) France	(3) Spain
Former British colony	**−0.48**	−0.36	0.41
	(0.12)	(0.07)	(0.10)
Former French colony	−0.14	**−0.40**	0.02
	(0.11)	(0.07)	(0.10)
Former Spanish colony	0.31	0.31	**−0.33**
	(0.13)	(0.09)	(0.10)
Other colonies	−0.03	−0.17	0.08
	(0.17)	(0.11)	(0.14)
N	190	190	190

In each model the dependent variable is the average absolute distance of each country's constitution from the country listed at the top of the column. For example, Model 1 shows the average distance from the UK constitution. Negative coefficients indicate more similarities. The omitted category in each model is countries that were never colonized. Robust standard errors shown below OLS coefficients.

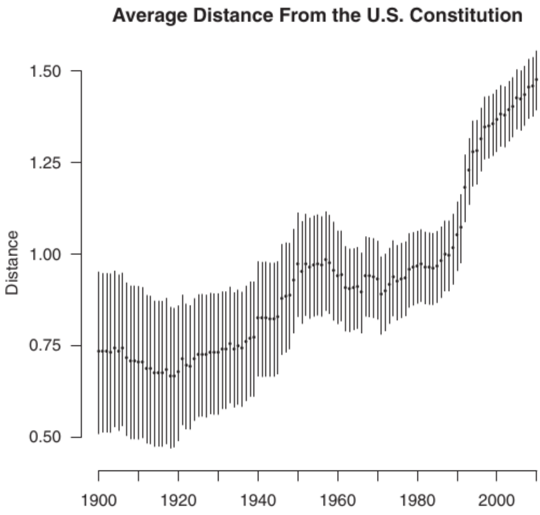


Figure 5: Similarity of constitutional systems to the United States over time.

Text analysis of patent innovation

"Measuring technological innovation over the very long run", Kelly et al. (2019)

- Goal:
  - Construct a new measure of novelty and impact of innovations based on similarity and distance between the text of patents
- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - Date, inventor, backward citations
  - Text (abstract, claims, and description)
- Text pre-processing:
  - Drop HTML markup, punctuation, numbers, capitalization, and stopwords
  - Remove terms that appear in less than 20 patents
  - 1.6 million words in vocabulary.

Text analysis of patent innovation

"Measuring technological innovation over the very long run", Kelly et al. (2019)

- Goal:
  - Construct a new measure of novelty and impact of innovations based on similarity and distance between the text of patents
- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - Date, inventor, backward citations
  - Text (abstract, claims, and description)
- Text pre-processing:
  - Drop HTML markup, punctuation, numbers, capitalization, and stopwords
  - Remove terms that appear in less than 20 patents
  - 1.6 million words in vocabulary.

Text analysis of patent innovation

"Measuring technological innovation over the very long run", Kelly et al. (2019)

- Goal:
  - Construct a new measure of novelty and impact of innovations based on similarity and distance between the text of patents
- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - Date, inventor, backward citations
  - Text (abstract, claims, and description)
- Text pre-processing:
  - Drop HTML markup, punctuation, numbers, capitalization, and stopwords
  - Remove terms that appear in less than 20 patents
  - 1.6 million words in vocabulary.

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\text{\# of patents prior to } p}{\log\left(1 + \text{\# documents prior to } p \text{ that include } w\right)}$$

  - Down-weights words that appeared frequently before a patent, but up-weights new words

- For each patent:
  - Compute cosine similarity to all future patents, using BIDF of earlier patent
- 9m$\times$9m similarity matrix = 30TB of data
  - Enforce sparsity by setting similarity $<$ .05 to zero (93.4% of pairs).

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w, p) = \frac{\text{\# of patents prior to } p}{\log\left(1 + \text{\# documents prior to } p \text{ that include } w\right)}$$

  - Down-weights words that appeared frequently before a patent, but up-weights new words

- For each patent:
  - Compute cosine similarity to all future patents, using BIDF of earlier patent

- 9m$\times$9m similarity matrix = 30TB of data
  - Enforce sparsity by setting similarity $<$ .05 to zero (93.4% of pairs).

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\text{\# of patents prior to } p}{\log\left(1 + \text{\# documents prior to } p \text{ that include } w\right)}$$

  - Down-weights words that appeared frequently before a patent, but up-weights new words
- For each patent:
  - Compute cosine similarity to all future patents, using BIDF of earlier patent
- 9m$\times$9m similarity matrix = 30TB of data
  - Enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

where $F(j)$ is the set of future patents (in, e.g., next 100 years).

- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

where $F(j)$ is the set of future patents (in, e.g., next 100 years).

- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

  where $F(j)$ is the set of future patents (in, e.g., next 100 years).

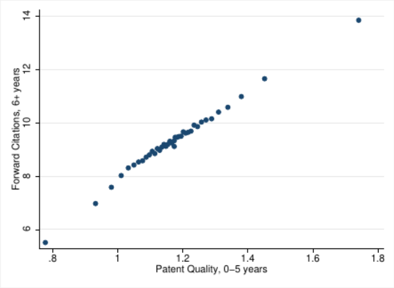
- A patent has high quality if it is novel and impactful:

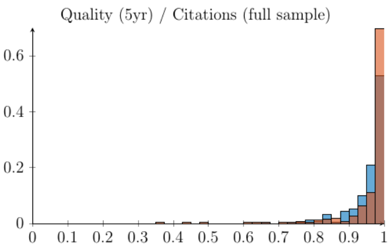
$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Validation

1. For pairs with higher $\rho_{i,j}$, patent j is more likely to cite patent i.
2. Patent office assigns 3-digit technology class code; similarity is significantly higher within class compared to across class.
3. Higher quality patents get more cites:

Validation (cont.)

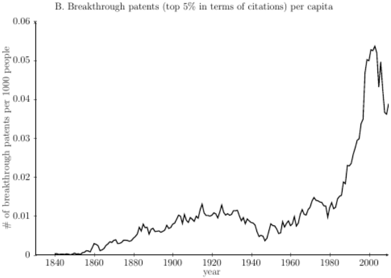


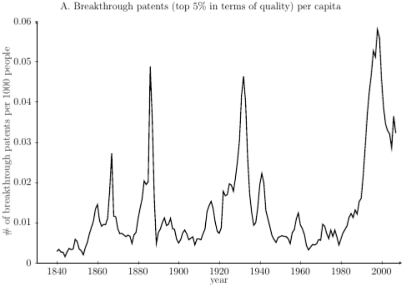


Most Innovative Firms

Assignee	First Year	# Breakthroughs
General Electric	1872	3,457
Westinghouse Electric Co.	1889	1,762
Eastman Kodak Co.	1890	2,244
Western Electric Co.	1899	1,222
AT&T (includes Bell Labs)	1899	5,645
Standard Oil Co.	1900	1,212
Dow Chemical Co.	1902	1,235
Du Pont	1905	3,353
International Business Machines	1908	14,913
American Cyanamid Co.	1909	690
Universal Oil Products Co.	1919	590
RCA	1920	3,222
Monsanto Company (inc. Monsanto Chemicals)	1921	902
Honeywell International, inc.	1928	872
General Aniline & Film Corp.	1929	1,181
Massachusetts Institute of Technology	1935	504
Philips	1939	1145
Texas Instruments	1960	2,088
Xerox	1961	2,198
Applied Materials	1971	510
Digital Equipment	1971	1,101
Hewlett-Packard Co.	1971	2,661
Intel	1971	2,629
Motorola, inc.	1971	4,129
Regents of the University of California	1971	823
United States Navy	1945	791
NCR	1973	737
Advanced Micro Devices	1974	1,195
Apple Computer	1978	864

Breakthrough patents per capita





Breakthrough patents and firm profits