Project 4:
List of Largest Companies In the United States by Revenue

Jeremy Mednik, Karim Naous, Stephen Owen

Creating the Wiki Dataset

Chose the page, List of largest companies in the United States by Revenue

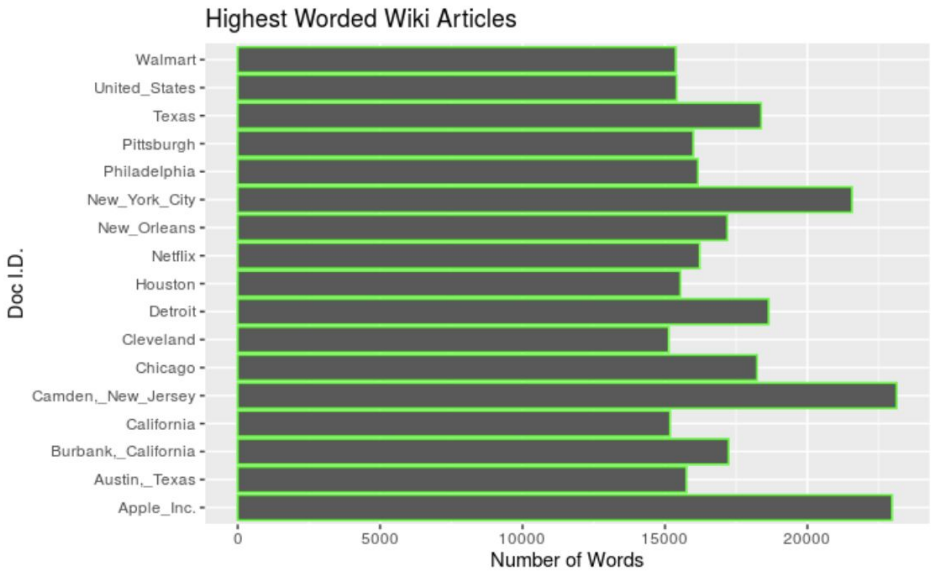
- Found 781 related pages
- Include cities, different companies, states, etc.

	doc_id	text
2	Abbott_Laboratories	Abbott Laboratories is an American multinational me...
3	AbbVie	AbbVie is an American publicly traded biopharmaceut...
4	ABM_Industries	ABM Industries Inc. is a facility management provider ...
5	Activision_Blizzard	Activision Blizzard, Inc. is an American video game ho...
6	ADM_(company)	The Archer-Daniels-Midland Company, commonly kn...
7	Adobe_Inc.	Adobe Inc. (// ə-DOH-bee) is an American multinatio...
8	ADP_(company)	Automatic Data Processing, Inc. (ADP) is an American ...
9	Advance_Auto_Parts	Advance Auto Parts, Inc. (Advance) is an American aut...
10	Advanced_Micro_Devices	Advanced Micro Devices, Inc. (AMD) is an American m...
11	AECOM	AECOM (// ay-ee-) (formerly AECOM Technology Cor...
12	Aerospace_engineering	Aerospace engineering is the primary field of enginee...
13	AES_Corporation	The AES Corporation is a Fortune 500 company that g...
14	Aflac	Aflac Inc. // (American Family Life Assurance Compan...
15	AGCO	AGCO Corporation is an American agricultural machin...
16	Air_Products_&_Chemicals	Air Products and Chemicals, Inc. is an American inter...
17	AK_Steel_Holding	AK Steel Holdings Corporation was a steelmaking co...
18	Akron,_Ohio	Akron (//) is the fifth-largest city in the U.S. state of ...
19	Alaska_Air_Group	Alaska Air Group is an airline holding company based...
20	Albertsons	Albertsons Companies, Inc. is an American grocery co...
21	Alcoa	Alcoa Corporation (a portmanteau of Aluminum Com...

N_Words

Wanted to see which Wiki articles had the highest amount of words…

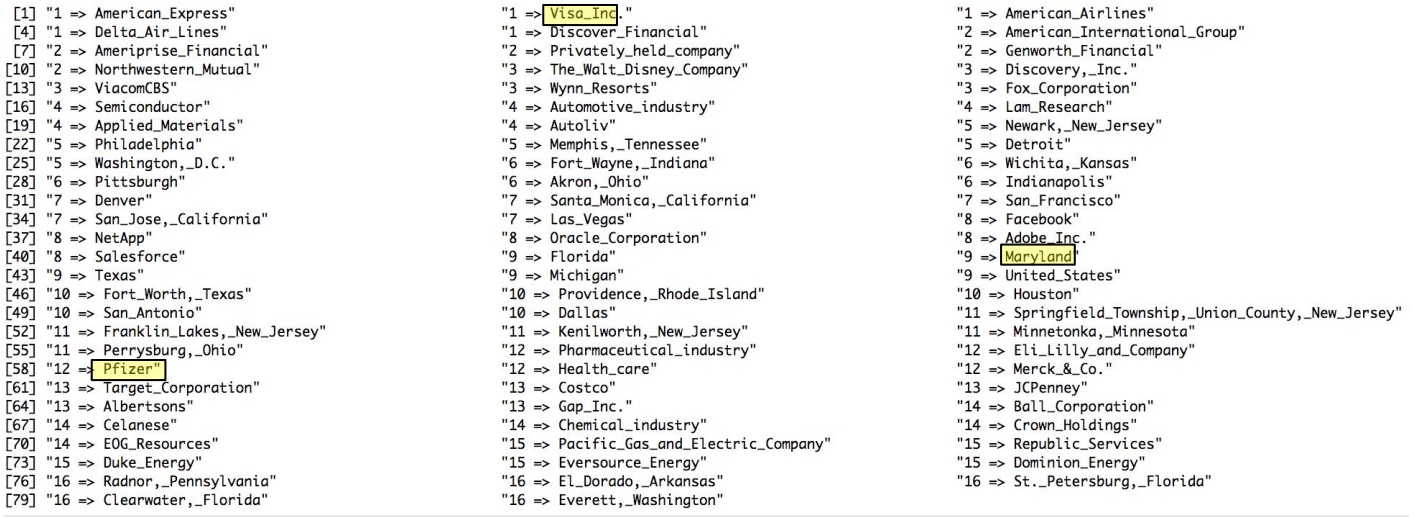
- Highest worded pages…
  - Camden, NJ
  - Apple Inc.
  - New York City
- Surprises
  - U.S. not being highest



Clusters

Split the Wiki articles into different clusters

- 16 clusters



```
 [1] "1 => American_Express"
 [4] "1 => Delta_Air_Lines"
 [7] "2 => Ameriprise_Financial"
[10] "2 => Northwestern_Mutual"
[13] "3 => ViacomCBS"
[16] "4 => Semiconductor"
[19] "4 => Applied_Materials"
[22] "5 => Philadelphia"
[25] "5 => Washington,_D.C."
[28] "6 => Pittsburgh"
[31] "7 => Denver"
[34] "7 => San_Jose,_California"
[37] "8 => NetApp"
[40] "8 => Salesforce"
[43] "9 => Texas"
[46] "10 => Fort_Worth,_Texas"
[49] "10 => San_Antonio"
[52] "11 => Franklin_Lakes,_New_Jersey"
[55] "11 => Perrysburg,_Ohio"
[58] "12 => Pfizer"
[61] "13 => Target_Corporation"
[64] "13 => Albertsons"
[67] "14 => Celanese"
[70] "14 => EOG_Resources"
[73] "15 => Duke_Energy"
[76] "16 => Radnor,_Pennsylvania"
[79] "16 => Clearwater,_Florida"
```

```
"1 => Visa_Inc."
"1 => Discover_Financial"
"2 => Privately_held_company"
"3 => The_Walt_Disney_Company"
"3 => Wynn_Resorts"
"4 => Automotive_industry"
"4 => Autoliv"
"5 => Memphis,_Tennessee"
"6 => Fort_Wayne,_Indiana"
"6 => Akron,_Ohio"
"7 => Santa_Monica,_California"
"7 => Las_Vegas"
"8 => Oracle_Corporation"
"9 => Florida"
"9 => Michigan"
"10 => Providence,_Rhode_Island"
"10 => Dallas"
"11 => Kenilworth,_New_Jersey"
"12 => Pharmaceutical_industry"
"12 => Health_care"
"13 => Costco"
"13 => Gap_Inc."
"14 => Chemical_industry"
"15 => Pacific_Gas_and_Electric_Company"
"15 => Eversource_Energy"
"16 => El_Dorado,_Arkansas"
"16 => Everett,_Washington"
```

```
"1 => American_Airlines"
"2 => American_International_Group"
"2 => Genworth_Financial"
"3 => Discovery,_Inc."
"3 => Fox_Corporation"
"4 => Lam_Research"
"5 => Newark,_New_Jersey"
"5 => Detroit"
"6 => Wichita,_Kansas"
"6 => Indianapolis"
"7 => San_Francisco"
"8 => Facebook"
"8 => Adobe_Inc."
"9 => Maryland"
"9 => United_States"
"10 => Houston"
"11 => Springfield_Township,_Union_County,_New_Jersey"
"11 => Minnetonka,_Minnesota"
"12 => Eli_Lilly_and_Company"
"12 => Merck_&_Co."
"13 => JCPenney"
"14 => Ball_Corporation"
"14 => Crown_Holdings"
"15 => Republic_Services"
"15 => Dominion_Energy"
"16 => St._Petersburg,_Florida"
```

Cluster Lemma

1. Banks, Airlines
2. Financial Institutions, Insurance
3. Entertainment
4. Cars, Research, Engineering
5. Cities
6. Cities
7. Cities
8. Computer Design/ Software
9. States
10. MORE CITIES

1	card; bank; credit; airline; payment
2	insurance; investment; financial; management; firm
3	deal; network; television; customer; stock
4	vehicle; technology; car; manufacturer; production
5	city; population; school; neighborhood; resident
6	city; team; downtown; school; population
7	city; population; downtown; resident; neighborhood
8	user; data; software; technology; computer
9	population; county; government; tax; percent
10	city; population; school; downtown; mile

Cluster Lemma

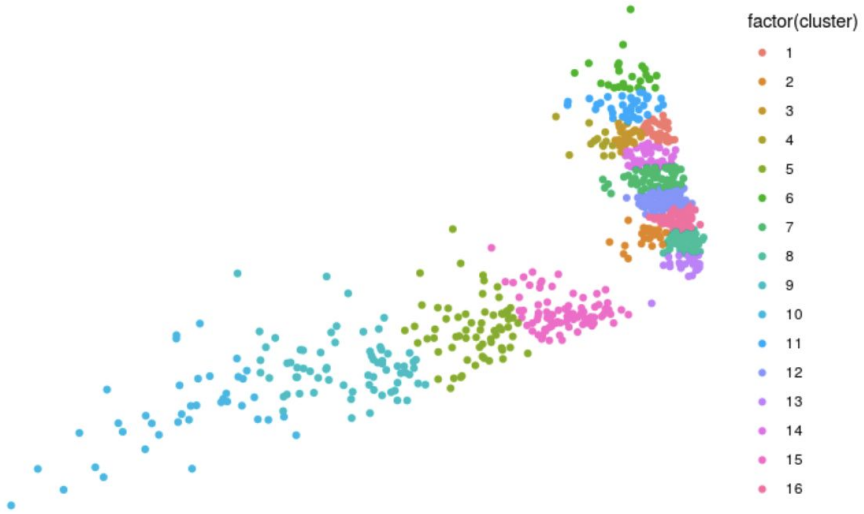
11. Cities
12. Healthcare
13. Large Retail Stores
14. Resource Companies
14. Energy Companies
15. Less Populated Cities

11	age; city; population; household; mile
12	drug; health; care; medical; patient
13	store; location; brand; chain; retail
14	oil; plant; food; chemical; gas
15	gas; power; plant; energy; utility
16	city; downtown; mile; community; age

PCA (16 Clusters)

Wanted to create a PCA with the different clusters...

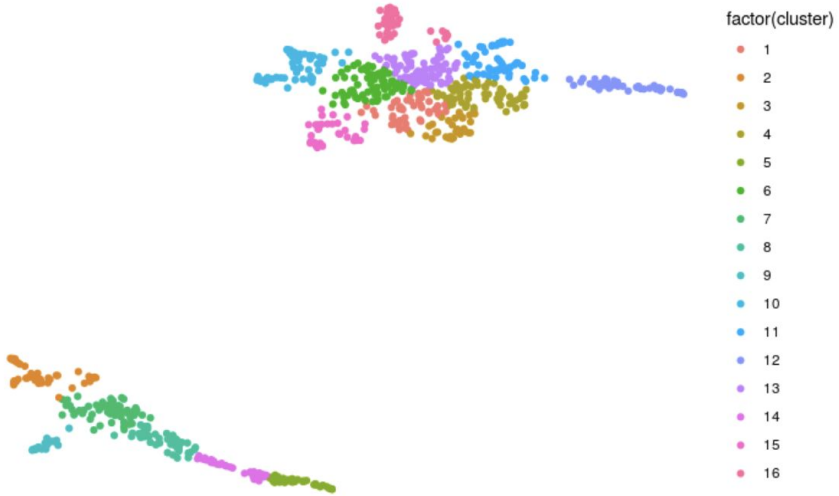
- A lot of the different companies were paired together in clusters.
- Clusters 11 and 6 were the most different from the PCA
  - 11 consisted of insurance and natural resource companies
  - 6 consisted of primarily financial institutions.
  - Makes sense considering the content of those wiki articles.



UMAP

Wanted to create a UMAP with the different clusters...

- Clusters on the bottom are
  - 2, 5, 7, 8, 14
- Clusters at the top are
  - 1, 3, 4, 6, 9, 10, 11, 12, 13, 15, 16
- Clusters on the bottom mainly consist of cities
- Clusters on the top are mainly other types of companies



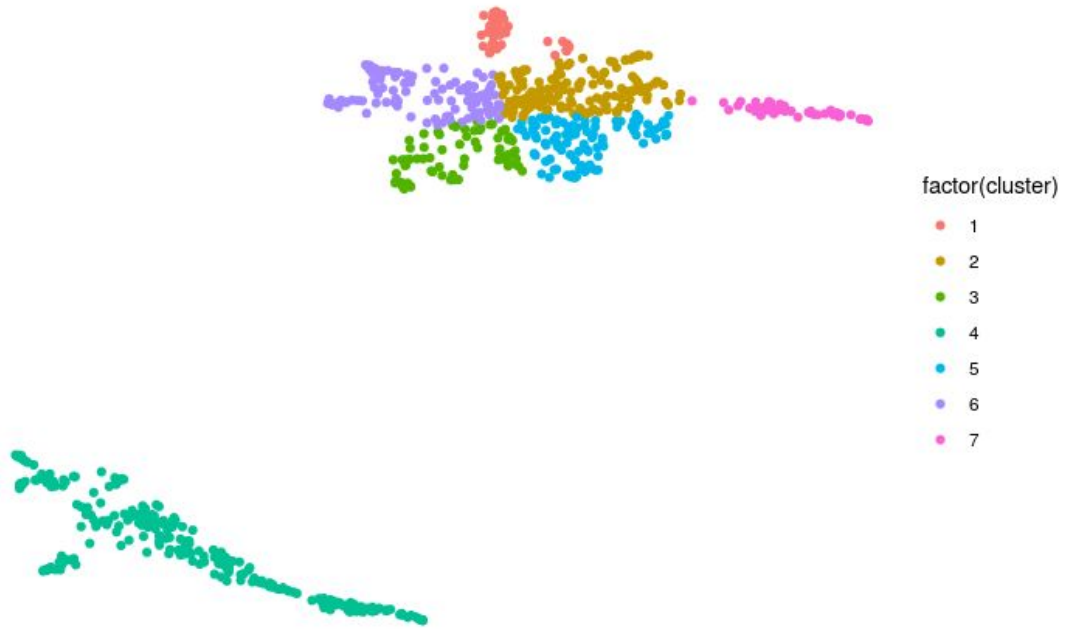
UMAP with 7 Clusters

[3]: Cities, states (ex: “Moline, Illinois”, “Missouri”)

[1]: Finance and insurance companies (ex: “Cincinnati Financial”, “Erie Insurance Group”)

[5]: Big retailers (ex: “Best Buy”, “Costco”)

[7]: Energy (ex: “Dominion Energy”, “ExxonMobil”)

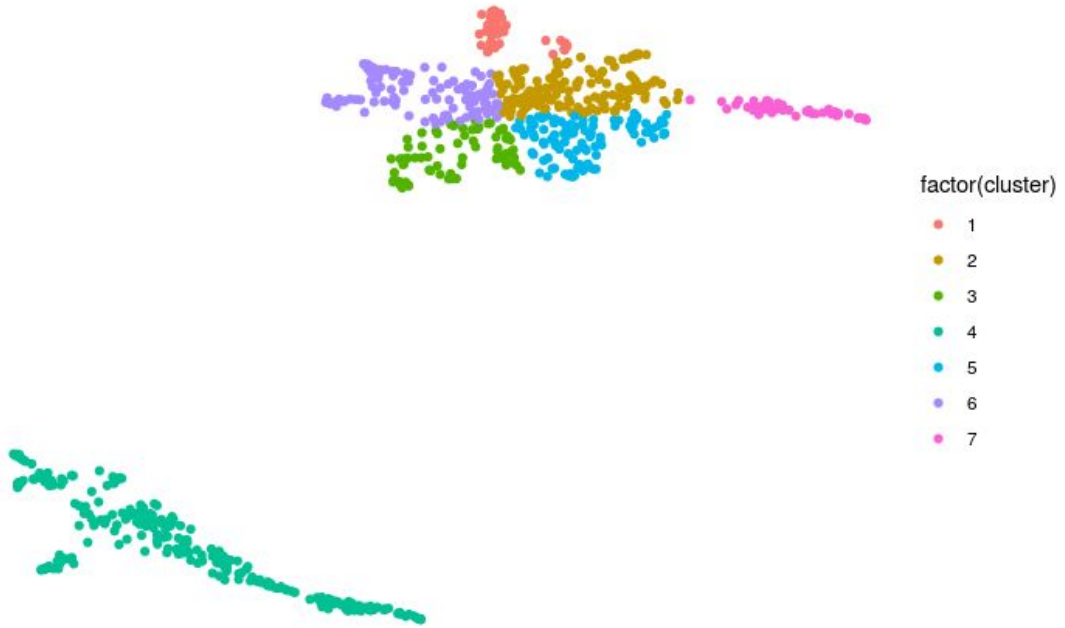


UMAP with 7 Clusters (cont.)

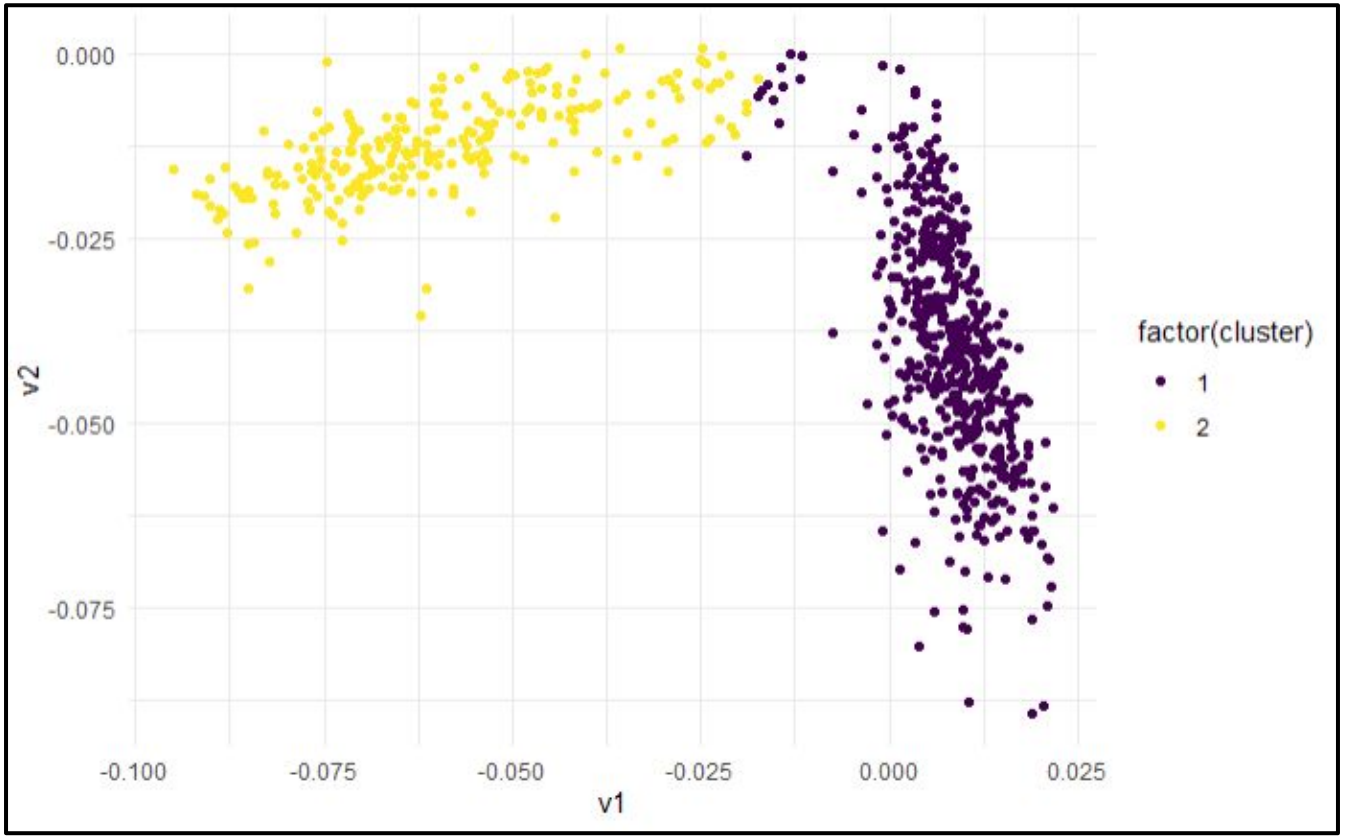
[2]: Manufacturing, Engineering (ex: “Clorox”, “Kraft-Heinz”, “Howmet Aerospace”)

[4]: Mix of technology companies and banks/finance companies (ex: “Alphabet Inc.”, “Capital One”). Could be technology and Fintech companies.

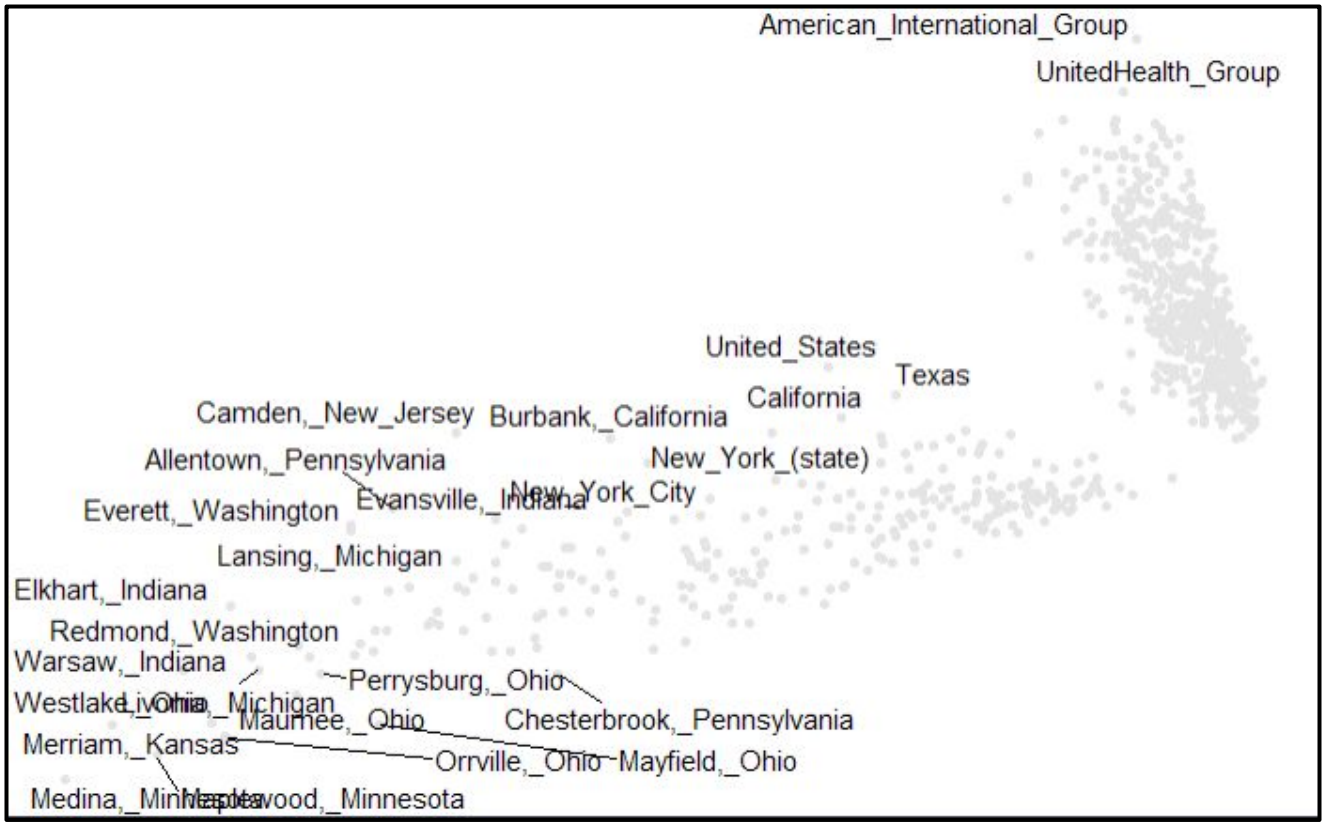
[6]: Technology, Automobile, and Airline companies (ex: “IBM”, “Southwest Airlines”, “Ford Motor Company”)



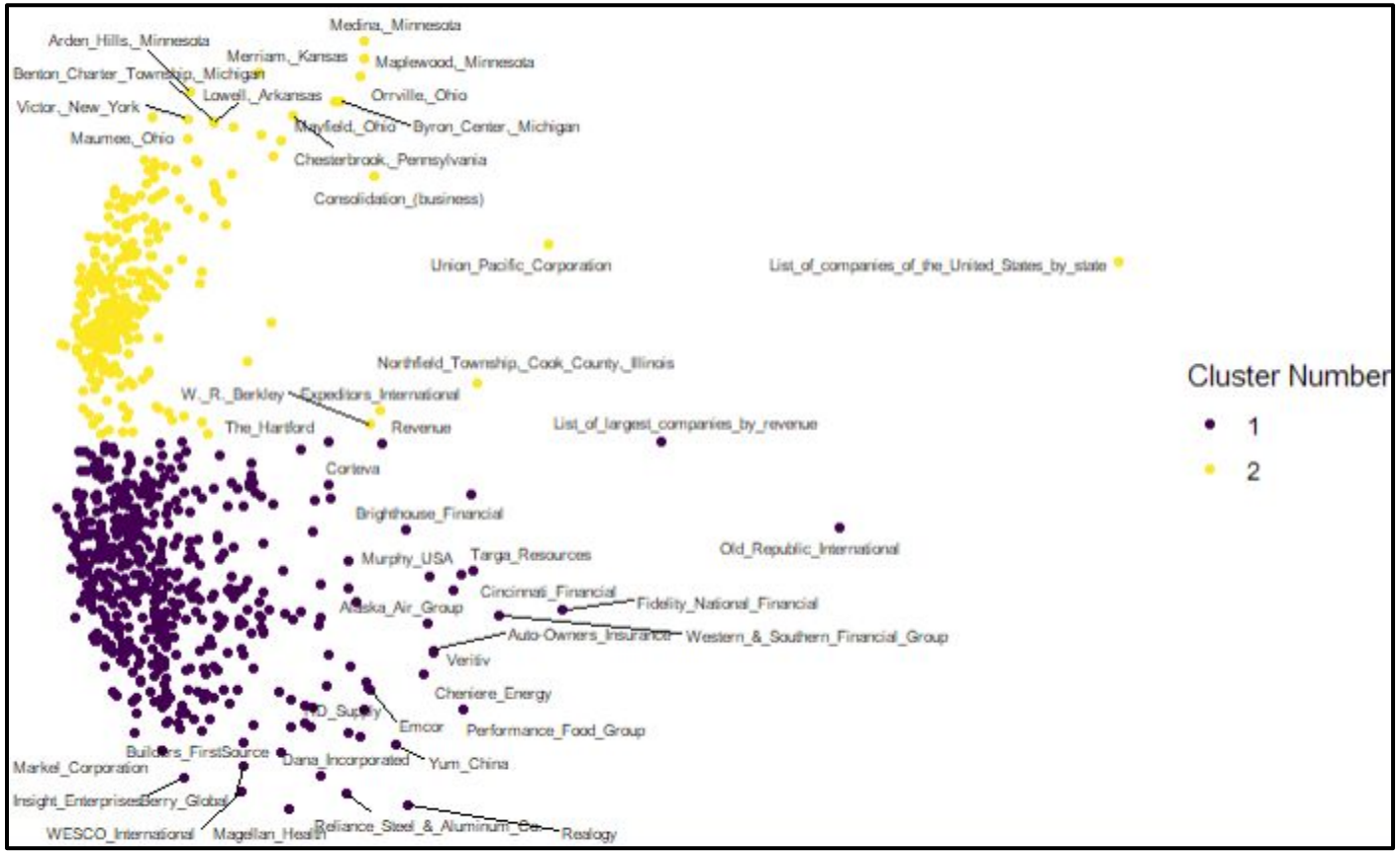
2 Clusters


0.000
-0.025
-0.050
-0.075
v2
-0.100
-0.075
-0.050
-0.025
0.000
0.025
v1
factor(cluster)
1
2

2 Clusters


American_International_Group
UnitedHealth_Group
United_States
Texas
California
Camden,_New_Jersey
Burbank,_California
Allentown,_Pennsylvania
New_York_(state)
Evansville,_Indiana
New_York_City
Everett,_Washington
Lansing,_Michigan
Elkhart,_Indiana
Redmond,_Washington
Warsaw,_Indiana
Perrysburg,_Ohio
Westlake,_Ohio
Livonia,_Michigan
Maumee,_Ohio
Chesterbrook,_Pennsylvania
Merriam,_Kansas
Orrville,_Ohio
Mayfield,_Ohio
Medina,_Minnesota
Maplewood,_Minnesota

2 Clusters (cont.)



Topic Modeling - LDA

1	card; bank; credit; airline; payment; branch; flight; aircraft; account; building
2	insurance; investment; financial; management; firm; asset; stock; fund; bank; acquisition
3	deal; network; television; customer; stock; film; share; cable; best; game
4	vehicle; technology; car; manufacturer; production; engine; model; equipment; ceo; contract
5	city; population; school; neighborhood; resident; district; century; building; center; county
6	city; team; downtown; school; population; local; mile; average; event; center
7	city; population; downtown; resident; neighborhood; south; rate; district; park; temperature
8	user; data; software; technology; computer; customer; digital; application; platform; phone
9	population; county; government; percent; tax; century; region; election; school; country
10	city; population; downtown; school; mile; rock; century; average; region; station

Topic Modeling - LDA (cont.)

topic	proportion of corpus
Cluster 1: card; bank; credit; airline; payment	4%
Cluster 2: insurance; investment; financial; management; firm	13%
Cluster 3: deal; network; television; customer; stock	5%
Cluster 4: vehicle; technology; car; manufacturer; production	9%
Cluster 5: city; population; school; neighborhood; resident	3%
Cluster 6: city; team; downtown; school; population	3%
Cluster 7: city; population; downtown; resident; neighborhood	3%
Cluster 8: user; data; software; technology; computer	5%
Cluster 9: population; county; government; percent; tax	3%
Cluster 10: city; population; downtown; school; mile	3%
Cluster 11: age; city; population; household; mile	12%
Cluster 12: drug; health; care; medical; patient	6%
Cluster 13: store; brand; location; chain; retail	10%
Cluster 14: oil; plant; food; chemical; gas	9%
Cluster 15: gas; power; plant; energy; utility	6%
Cluster 16: city; downtown; mile; community; age	6%