Preliminary Report: Natural Language Processing Analysis of LGO Theses

Kerry Weinberg[1,*]

[*]kerryweinberg@alum.mit.edu

ABSTRACT

MIT's Leaders for Global Operations program theses afford a unique view into the evolution of industry research in manufacturing and operations over the past 32 years. Through natural language processing analyses of the theses we can evaluate overall trends and shifts in research across industry sectors and over time. SPECTER, a recently developed natural language processing embedding technique, was applied to the entire corpus of theses. Analyses of similarity in embedded theses show that this technique can adequately represent documents in the manufacturing and operations research space. Clustering analysis of embedded representations of theses as well as topic modeling reveals both topics and trends which remain consistent over time and across industry verticals as well as distinct shifts in specific topics in pre and post Great Recession periods. These results suggest that industry research in application of lean has remained consistent over time and while machine intelligence topics may not have been referred to as such, application of data driven methods to manufacturing and operations has been consistent over time. Natural Language Processing of industry research encapsulated in 30 years of LGO theses identifies both consistent trends and areas of focus for firms as well as those sensitive to major economic shocks such as the Great Recession.

Introduction

LGO Theses represent a unique data set to study industry research in manufacturing and operations over the past 32 years. Given the significant impact of globalization, explosion of services industry, and automation, one would expect to identify distinct topics and themes that emerge over time periods. In order to ascertain the degree to which these topics might be identified in the theses data set, a natural language processing pipeline was applied to extract a latent representation of the theses which could be analyzed by various machine learning techniques. SPECTER, a newly published algorithm from the Allen Institute of AI, has been shown to be successful in various natural language processing tasks that require semantic understanding of long form scientific documents (e.g. publications)[1] Given the industry specific language utilized in LGO theses, we hypothesized that using such an embedding technique might prove more successful than standard natural language processing methods that are trained on generic English language corpuses (e.g. Word2Vec). Analysis of the embedded representation of the theses can then be complemented with topic modeling, in this case performed using Amazon's Comprehend service. Together, clustering analysis of the embedded theses, topics identified from Comprehend, and meta-data on the theses (e.g. Year, Industry), can be analyzed to identify trends across verticals and over time.

Methods and Results

0.1 Data Engineering

Theses contained in the analysis were present in pdf form and text was extracted from them using the AWS Textract service. Metadata on the theses such as MIT student name, engineering department, company name, and class year were available from the MIT LGO IT department. Categorization of companies to different industries was performed manually based on human judgement and can be seen in Table 2.

0.2 SPECTER

SPECTER is a recently published algorithm from the Allen Institute for AI which generates embeddings to accomplish a variety of natural language processing tasks. Of note, SPECTER has been trained on a corpus of published literature from a variety of academic fields. SPECTER embeddings were generated for the theses to facilitate downstream analyses. Individual theses were evaluated based on cosine similarity of the embeddings to evaluate the performance of the embeddings.

0.3 K-Means Clustering

K-Means Clustering was applied to the SPECTER embedded representation of the theses to evaluate whether the semantic similarity of theses was specific to a time frame or would be identified over time. For example, theses in the early 2000's

might have referred to data driven analysis of manufacturing data as "statistical modeling" while these days that would more likely be referred to as "machine learning". Analyzing the SPECTER embeddings using clustering enables a summarized representation of a subset of theses which can be evaluated over time periods and across industry verticals. A total of 10 clusters were identified with the optimal K=10 found using the elbow method.

0.4 Topic Modeling

While the SPECTER embeddings can be used to address a number of natural language processing tasks, topic modeling can be helpful in gaining a rapid understanding of the type of language and topics referenced in the data set. Hence, AWS Comprehend service was used to perform a topic modeling analysis on the data set. Analyses were conducted using both a target number of topics of 15 and 30. The results reported here are based on a total of 30 topics as a finer grained assessment of the topics was desired and descriptive plots focus primarily on a subset of these topics that were identified as important features in the decision tree explanation described below.

0.5 Decision Tree Explanation

In order to better understand why different theses were being identified as semantically similar to each other using their SPECTER embedding, a decision tree was used to be able to identify the key drivers behind why theses were clustered into different clusters. The input features to the decision tree analysis were class year, industry, and topics identified through the topic modeling analyses. The target label was the cluster. Decision Tree Analysis at multiple maximum depths was performed and overall consistency in features that drove splits is reported. An example of a decision tree is provided in 1. The features driving the cluster assignment were consistently 2009 class year and a number of topics, namely 1, 9, 10, 11, 19, 23, 27, 28, and 29. 2012 and 2016 class years were also identified as key time frames in a subset of the analyses.

0.6 Exploratory Data Analysis of Topics and Clusters across Industry Sectors and over Time

Visualization of the breakdown of theses based on topics, clusters, industries, and class year was performed to characterize key time points as well as semantic similarity of theses across industry sectors and over time. Figures 2 and 3 show how industry representation in theses as well as topics identified (a sub-selection of topics as identified through the decision tree explainer analysis is shown) have evolved over the past 32 years. Of note, prior to 2006 a majority of theses cannot be adequately linked to an industry. This is due to missing internship company information in the dataset which cannot be mapped to a higher level industry category.

Manual curation of the industry associated with theses during this time period could benefit the analyses through increased sample size. Nevertheless, topic identification via AWS Comprehend and clusters identified using K-Means on SPECTER embeddings should not be impacted by any lack of industry information of the theses as these tasks were performed on the thesis text and not thesis metadata. Figure 4 shows that overall class year does not itself explain the differences in the clusters identified, suggesting that semantically, research remains overall relatively consistent over the years.

Nevertheless, a few clusters differ meaningfully from the rest in terms of median class year represented, notably Cluster 7, with the oldest median class year and Cluster 4 with the most recent median class year. Further evaluation of cluster representation provides an assessment of whether specific clusters over represent or under represent different industries, as seen in figures 5, 7, and 6.

These results suggest Cluster 4, having a greater representation of theses associated with the healthcare industry as well as a more recent median class year may overall represent semantic concepts and research more prevalent to the healthcare industry which LGO has had more partner companies in the healthcare industry particularly in the past 15 years. In contrast, Cluster 7, with the lowest median class year contains a number of theses associated with the auto industry and thus may overall represent concepts more associated with issues present to that time frame, which again could be more reflective of the greater prevalence of auto industry partner companies with LGO at that time.

Analysis of cluster representation across industries over time shows differences in representation across industries as well as during distinct time periods as seen in figures 8, 10 and 9. Lastly, a detailed breakdown of topics identified in theses based on industry can be seen in both pre and post great recession periods as shown in figures 12 and 11. Of note, since a single thesis can have multiple topics identified through topic modeling, the counts of total theses exceed the number of absolute theses present in the dataset.

Discussion

0.7 Great Recession impact on research

Through descriptive analysis of topics, industry, and clusters identified over time as well as through the results of a decision tree explanation for the clusters identified reveal both consistent trends in research conducted as well as areas of research which define research in pre and post Great Recession periods. Topics related to reducing product complexity and associated

manufacturing costs seen in topic 9, topic 1 referring to demand forecasting and prediction, and topic 11 related to manufacturing capacity and constraints appear to occur relatively consistently over time. In contrast, other topics, such as topic 10, appear to be primarily driven by one company, in this case, Amazon, rather than across industries and hence suggest there is research that is specific to Amazon being identified.

The decision tree explainer conducted at multiple depths resulted in 2009 as a key class year in explaining the difference in clusters identified using K-Means clustering on the SPECTER embeddings. Since the SPECTER embeddings use only the text present in the theses and do not explicitly take into account the time frame (although certainly the date may be listed explicitly in thesis), this suggests that semantically research differs in a pre-recession time period vs. post-recession time period. Descriptive analysis of the sub-selected topics identified as important in the decision tree explainer, reveal trends in topics occurring more frequently in theses pre 2009 vs. post 2009. For example, topic 8 and 19 referring to supply chain risk and supplier relationships, shows a decrease after 2009 while topic 23, referring to demand forecasting and topic 1 referring to lean and organizational change shows a steady increase over time. This suggests that through overarching economic shifts, industry research in the operations and manufacturing space may have fundamentally shifted.

0.8 Opportunities for Further Research

The analysis presented in this report could be further improved in multiple ways. Manual annotation of theses pre 2006 so they are labeled with applicable industry would improve the evaluation of the representation of clusters and topics across industries. SPECTER embeddings of the theses could be used to address multiple natural language processing tasks such as a Question Answering service to complement capabilities available from AWS Kendra. Further analyses of the embedded theses could be explored to assess whether topics semantically similar to automation are being explored disproportionately in different time periods and industry sectors. Lastly, evaluation of the correlation between industry research and company performance particularly in a post-Great Recession period could inform the impact of particular areas of research which may prepare companies better for significant economic shifts due to the Great Recession. This information could be particularly helpful in predicting industry research which may also prepare companies to perform well in a post-COVID economy.

References

1. Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. S. Specter: Document-level representation learning using citation-informed transformers (2020). 2004.07180.

Topic	Terms
0	work , lean , team, change, organization, management, job, improvement, employee, company
1	model, datum, variable, forecast, numb, input, year, prediction, distribution, predict
2	machine, line, tool, operator, work, time, plant, material, production, run
3	process, control, datum, step, product, variation, quality, batch, measurement, experiment
4	assembly, design, plant, line, vehicle, production, build, body, engine, tool
5	time, lead, part, engine, schedule, inventory, day, material, aircraft, datum
6	inventory, stock, demand, level, lead, material, supply, order, item, safety
7	cost, model, labor, product, saving, decision, estimate, camera, numb, complexity
8	technology, manufacture, company, site, risk, development, equipment, production, communication, industry
9	product, cost, market, development, manufacture, complexity, platform, company, sale, component
10	item, order, pick, fulfillment, amazon, ship, process, center, sort, pack
11	capacity, datum, production, model, factory, area, plan, figure, constraint, equipment
12	order, time, customer, ship, delivery, process, factory, day, schedule, fulfillment
13	chapter, cycle, manufacture, summary, tool, plant, technology, system, table, throughput
14	wafer, tool, test, yield, lot, fab, equipment, process, intel, semiconductor
15	customer, service, order, call, datum, business, intel, delivery, information, support
16	company, engine, figure, service, price, cost, tier, industry, business
17	supply, chain, inventory, model, demand, time, lead, stock, service, stage
18	development, group, team, portfolio, management, resource, organization, task, work
19	supplier, source, cost, risk, supply, company, price, relationship, component, chain
20	cid, figure, wafer, lot, material, yield, risk, plan, inspection, module
21	time, cycle, figure, factory, manufacture, improvement, lead, chapter, reduction, table
22	production, section, schedule, demand, plant, system, figure, order, assembly, vehicle
23	forecast, demand, sale, week, plan, model, error, inventory, supply, accuracy
24	figure, process, manufacture, system, flow, map, lean, chain, production, improvement
25	test, failure, wafer, device, quality, defect, result, board, perform, unit
26	patient, bed, discharge, hospital, nurse, transfer, day, icu, wait, unit
27	quality, paint, cost, problem, appendix, plant, improvement, line, analysis, thickness
28	design, engineer, team, development, assembly, component, program, process, system, requirement
29	part, 18k3, fwc, machine, cell, manufacture, cellular, 737x, capacity, plane

Table 1. Topic Modeling Terms

Industry	Companies
Aero	Spirit AeroSystems Inc., United Technologies Corporation, Raytheon-Space and Airborne Systems, The Boeing Company, United Technologies Corporation - Pratt & Whitney, Raytheon Company, Bell Helicopter, United Technologies Corporation - Sikorsky Aircraft, Raytheon - Integrated Defense Systems, United Technologies Corporation - Aerospace Systems, General Dynamics Corporation, Lincoln Laboratory, Northrop Grumman Corporation, GE Aviation, Raytheon - Intelligence and Information Systems , NASA
Healthcare	Broad Institute, Amgen Inc., Genzyme Corporation, Novartis Pharma AG, Johnson & Johnson, Boston Scientific, Novartis Institutes for Biomedical Research, Inc., Pfizer, Sanofi, Novartis Vaccine and Diagnostics, Massachusetts General Hospital, Beth Israel Deaconess Medical Center, Quest Diagnostics, Genentech, Novartis Biologics, Danaher, PerkinElmer, Inc.
Hi-tech	Honeywell International Incorporated, Intel Corporation, Cisco Systems Incorporated, SanDisk Corporation, Cisco/Flextronics, Hewlett-Packard Company, Digital Equipment Corporation, Lucent Technologies Incorporated, Flextronics, Motorola Incorporated, Nokia Corporation, Polaroid Corporation, 3M Company, iRobot, Eastman Kodak Company, Dell, Inc., MASCHINENFABRIK REINHAUSEN GMBH, MASCHINENFABRIK REINHAUSEN, Teradyne Incorporated, ABB, ABB Limited, ATI Technologies, Inc., Axcelis Technologies Incorporated, American Industrial Partners
Auto	Ford Motor Company, Harley-Davidson Motor Company, General Motors Corporation, Goodyear Tire and Rubber Company, Caterpillar, American Axle and Manufacturing, Valeo SA
Attire	Nike, Li & Fung, Zara (Inditex, S.A.)
Utilities	National Grid, Pacific Gas & Electric, Schlumberger, Siemens Power Transmission and Distribution, Verizon
Consumer	Amazon, Kimberly Clark Corp., The Procter & Gamble Company, Pepsi Bottling Group, Pepsi Bottling Company, C&S Wholesale Grocers, Amazon.com Incorporated
Raw Materials	H.C. Starck GmbH, Aluminum Company of America, BMHC

Table 2. Manual curation mapping companies to industry groups

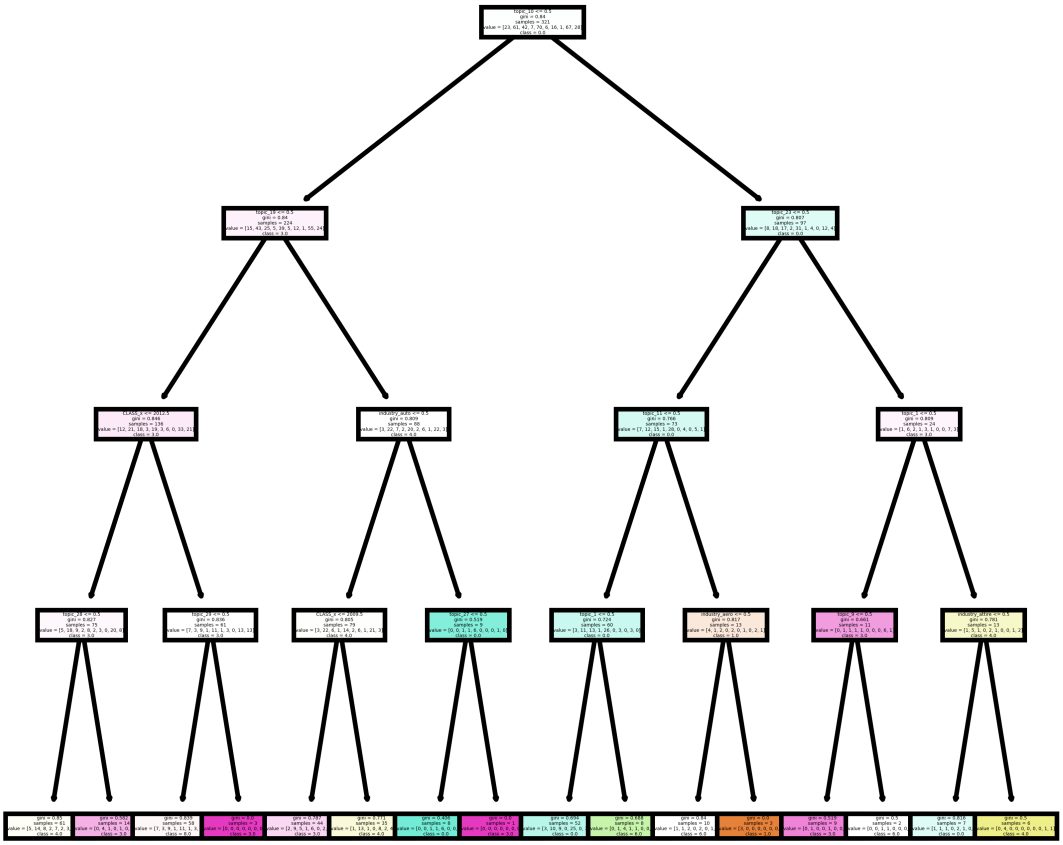

Figure 1. Decision Tree Explainer Results

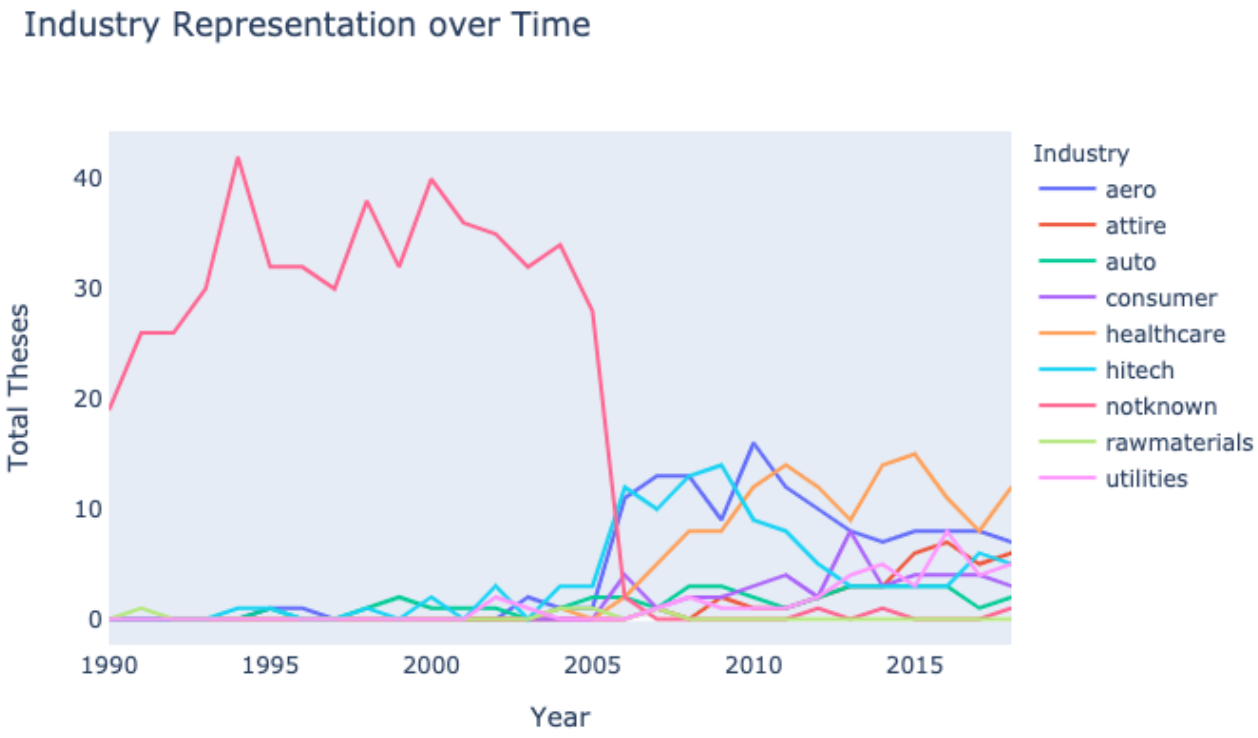

Figure 2. Theses by Industry over Years

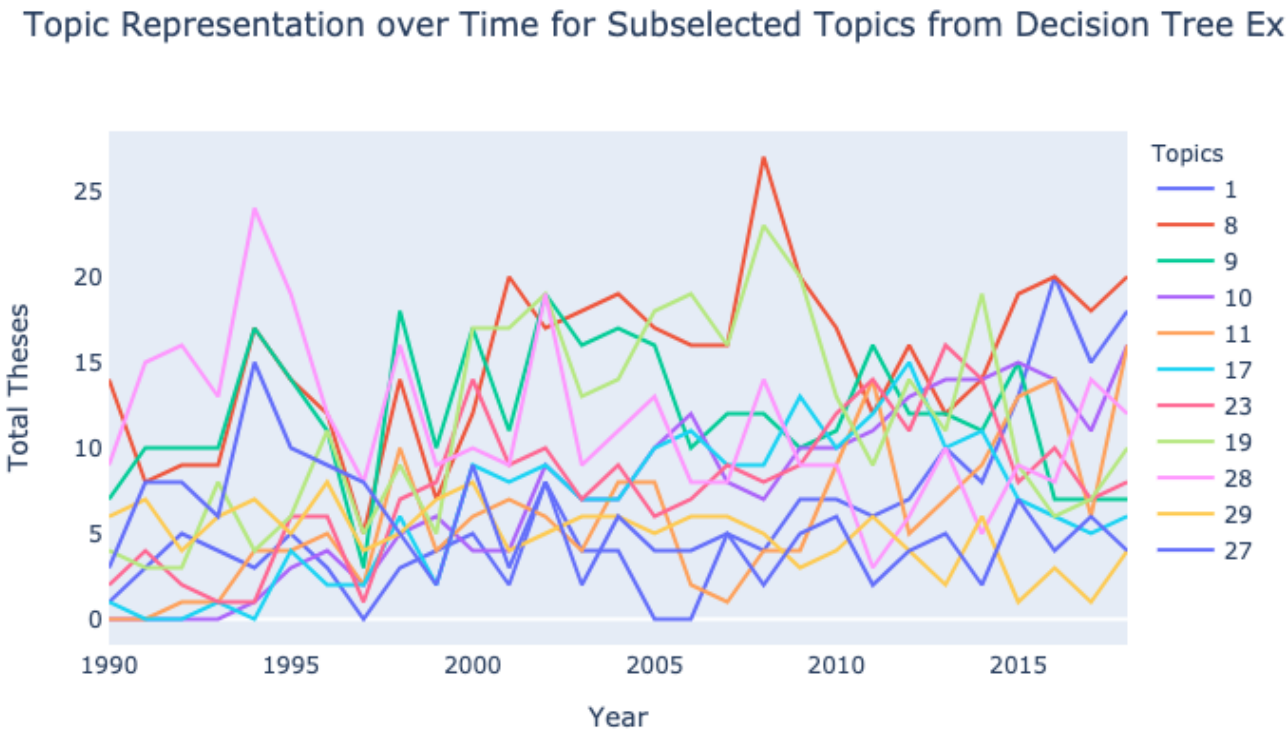

Figure 3. Sub Selected Topics over Years

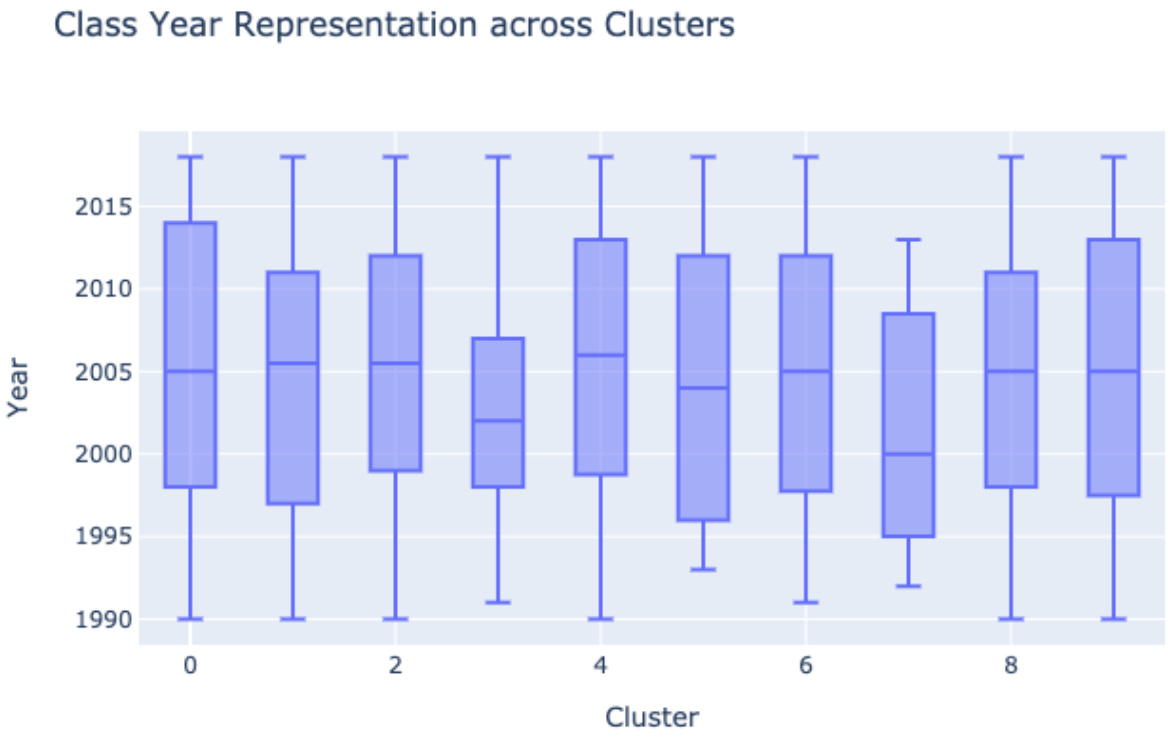

Figure 4. Average Class Year by Clusters

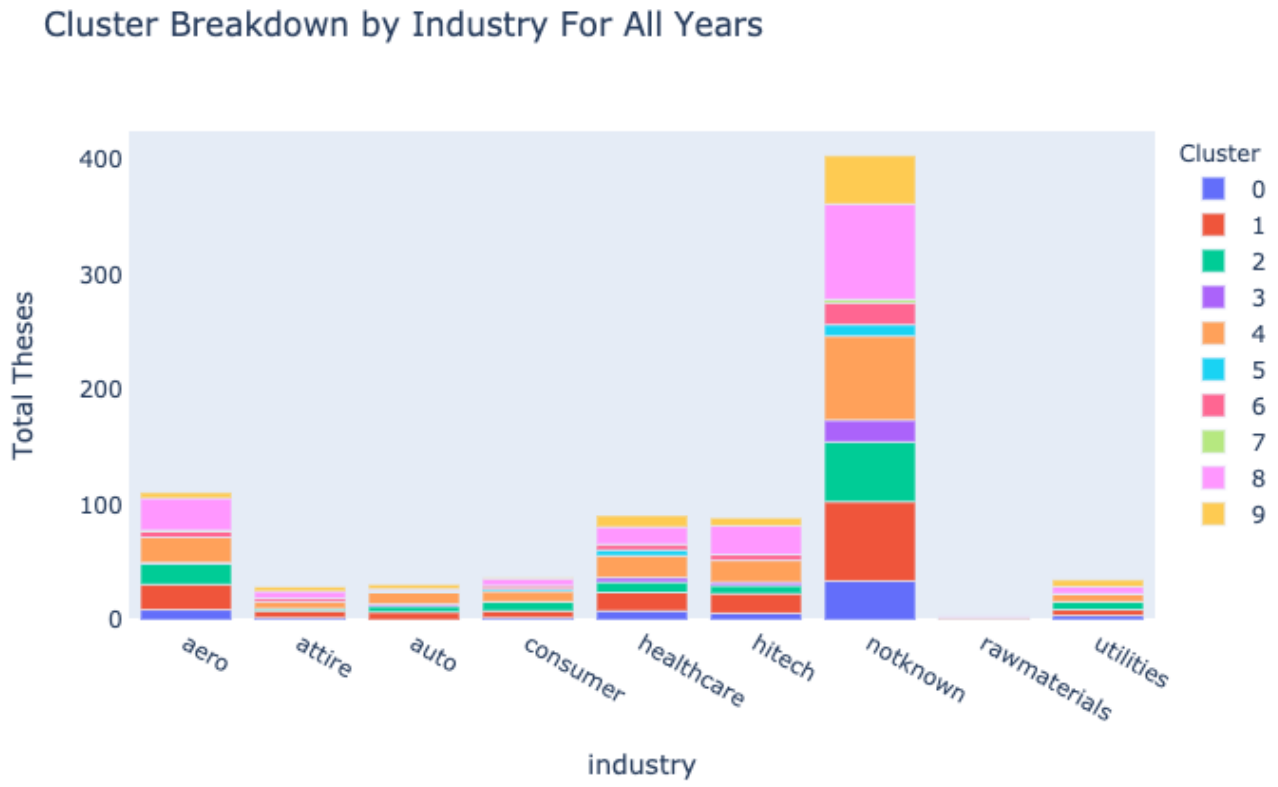

Figure 5. Cluster Breakdown by Industry over All Years

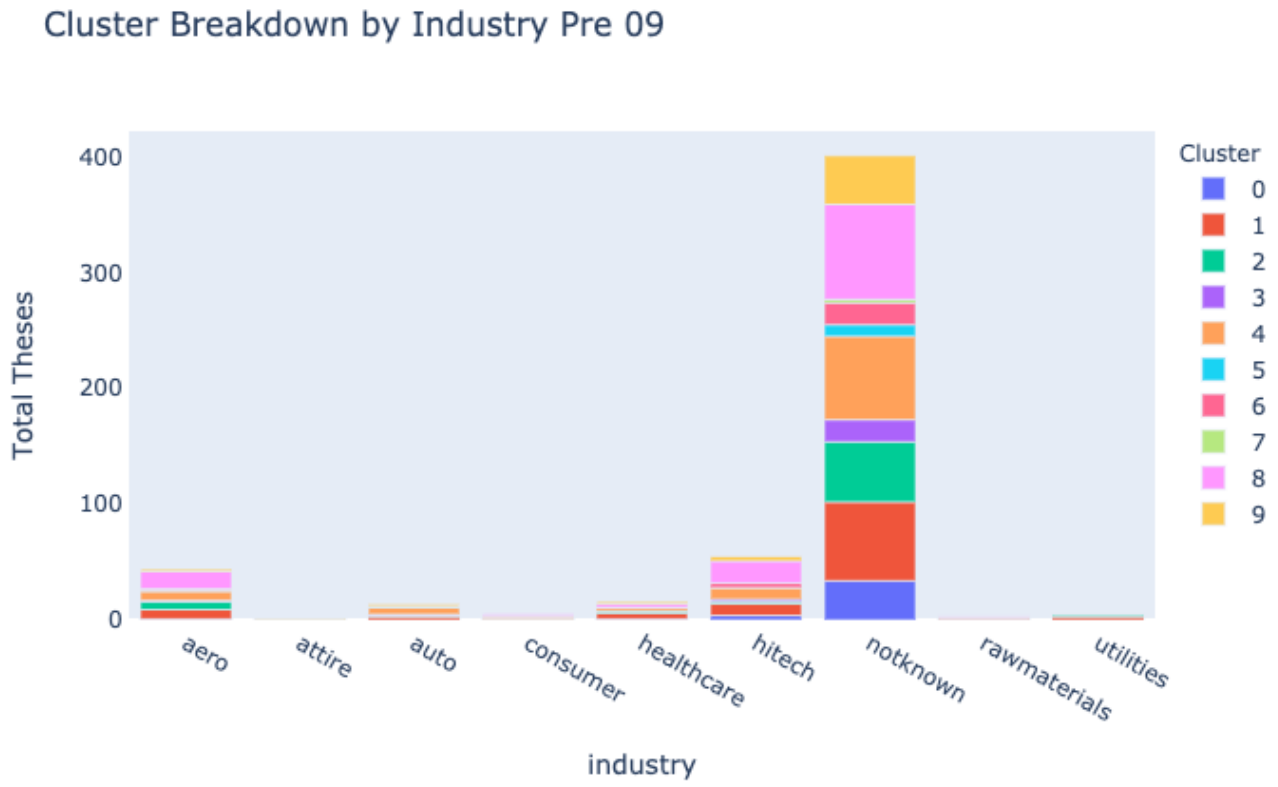

Figure 6. Cluster Breakdown by Industry pre 2009

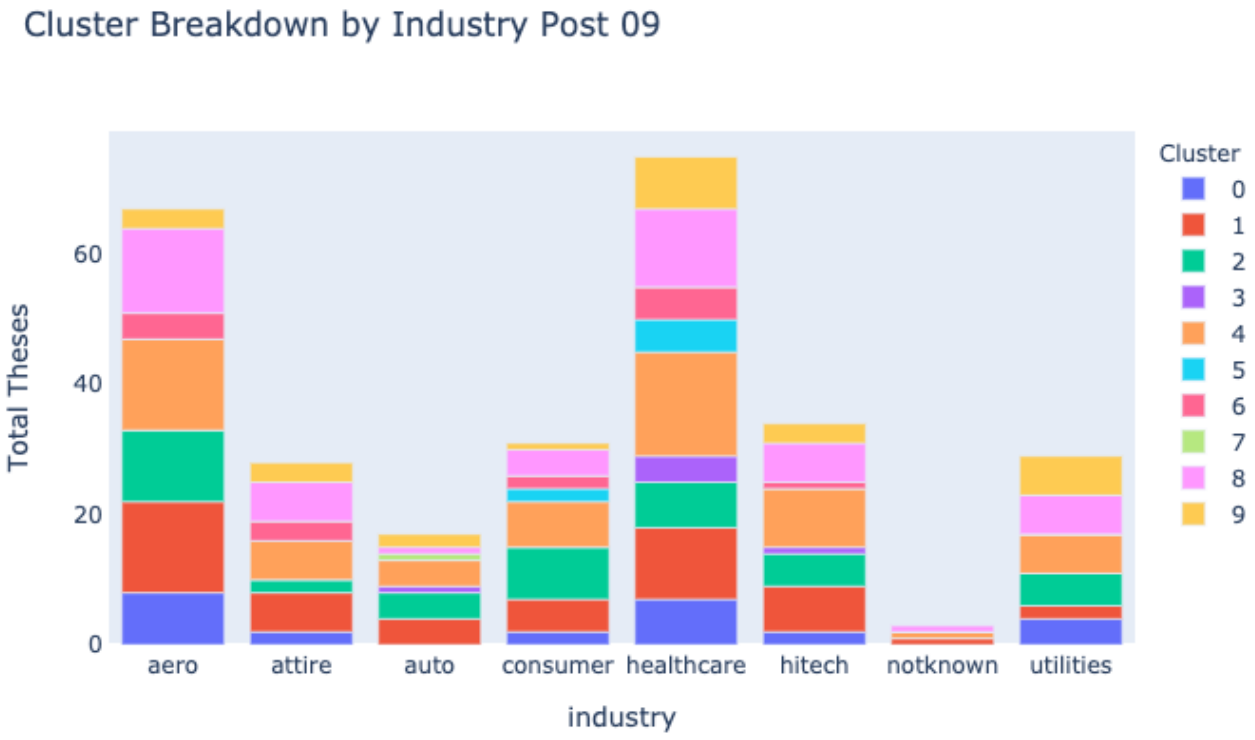

Figure 7. Cluster Breakdown by Industry post 2009

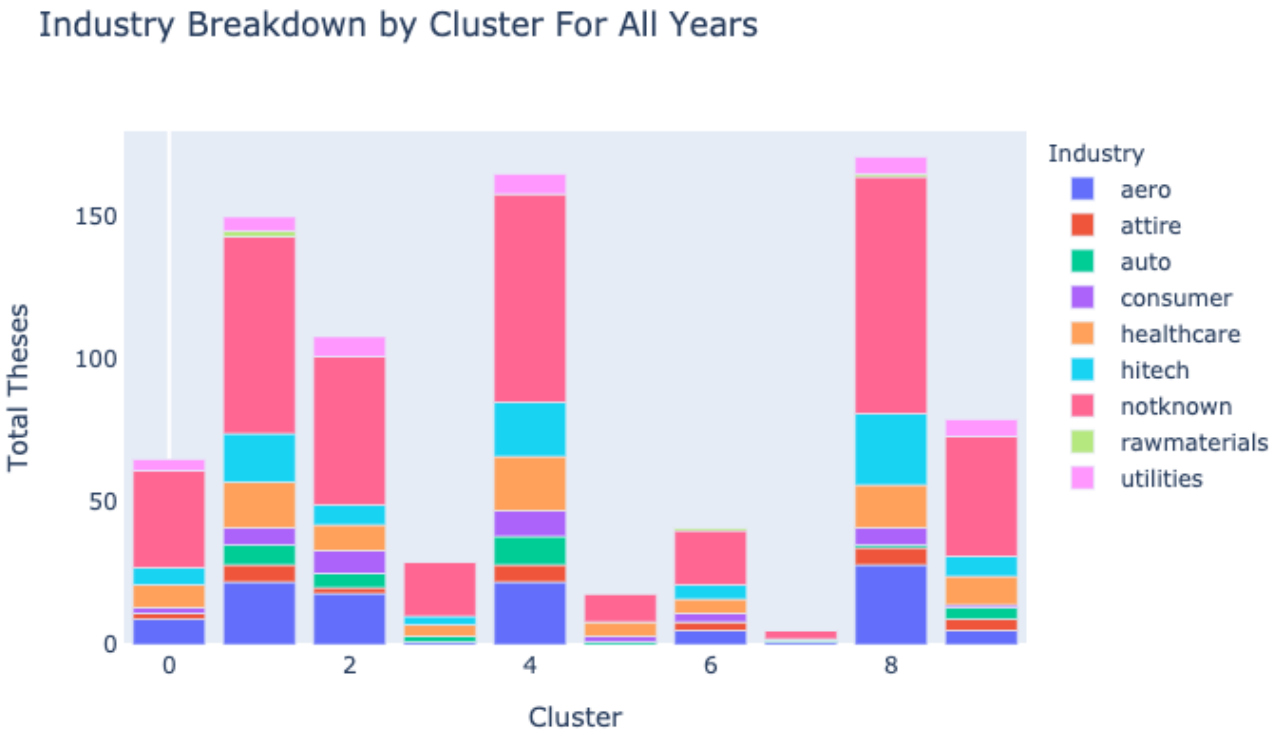

Figure 8. Industry Breakdown by Cluster for All Years

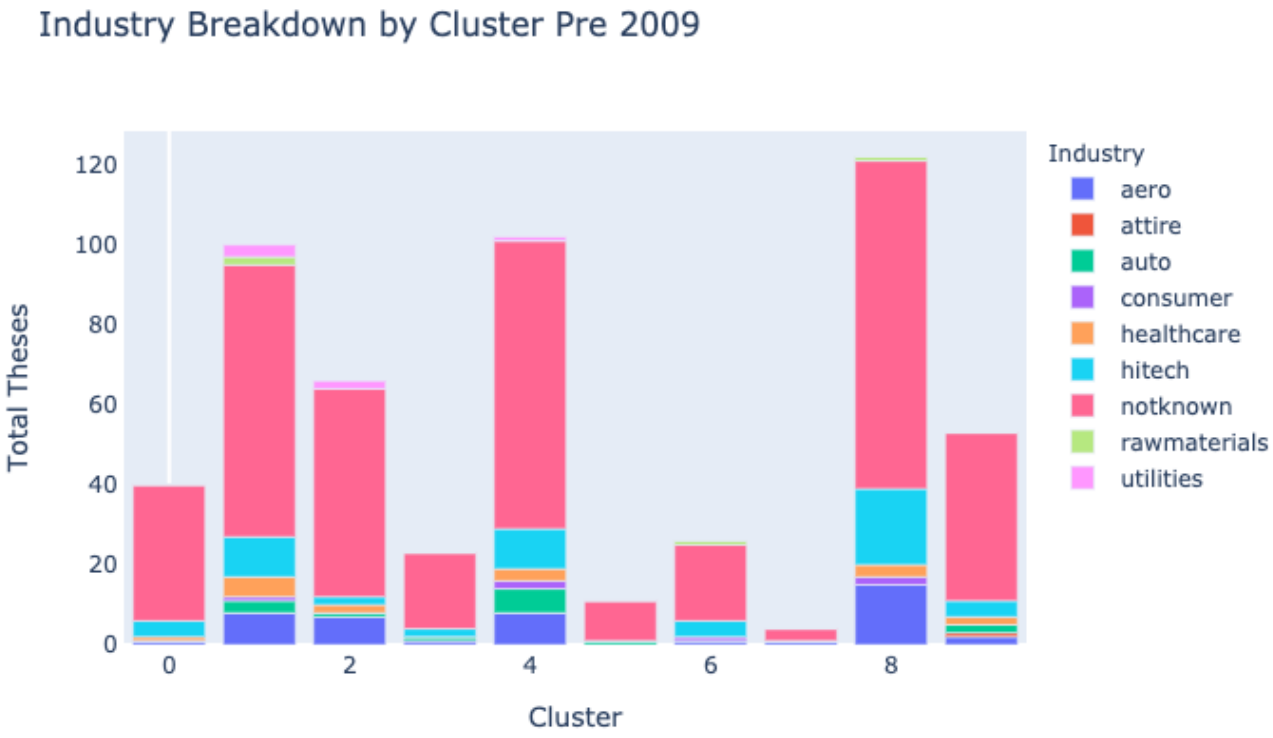

Figure 9. Industry Breakdown by Cluster pre 2009

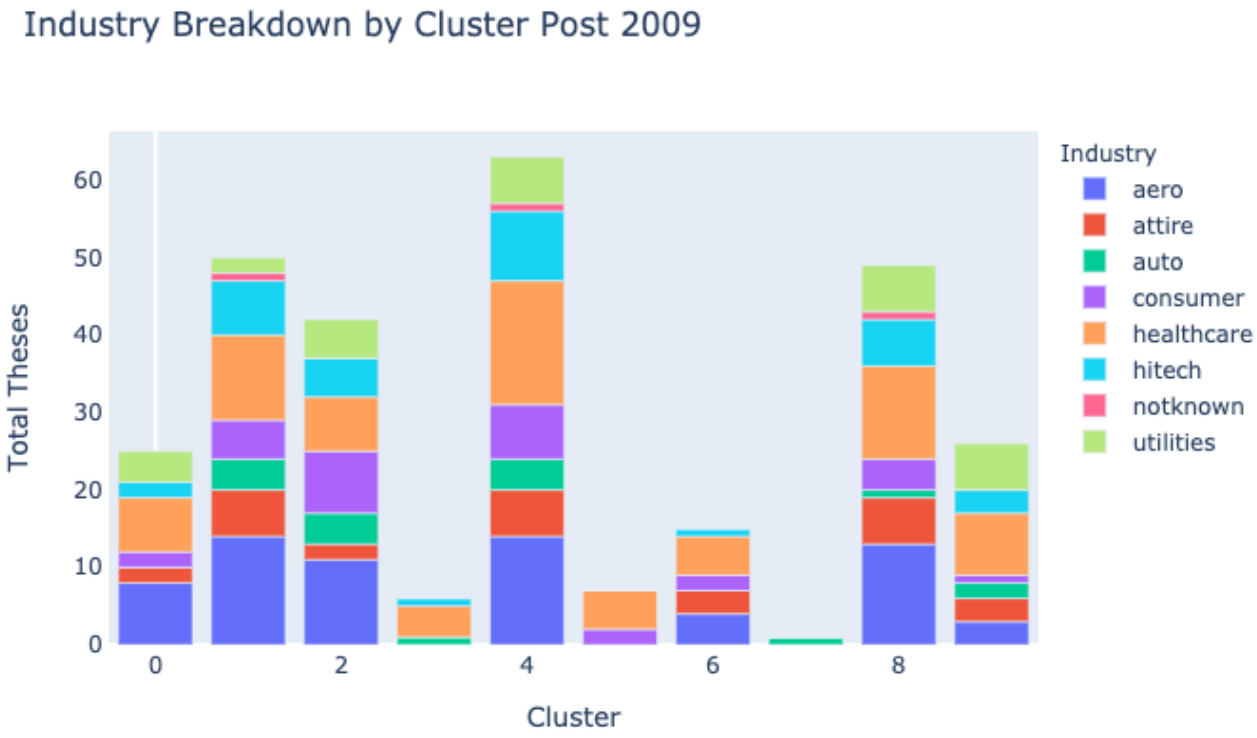

Figure 10. Industry Breakdown by Cluster post 2009

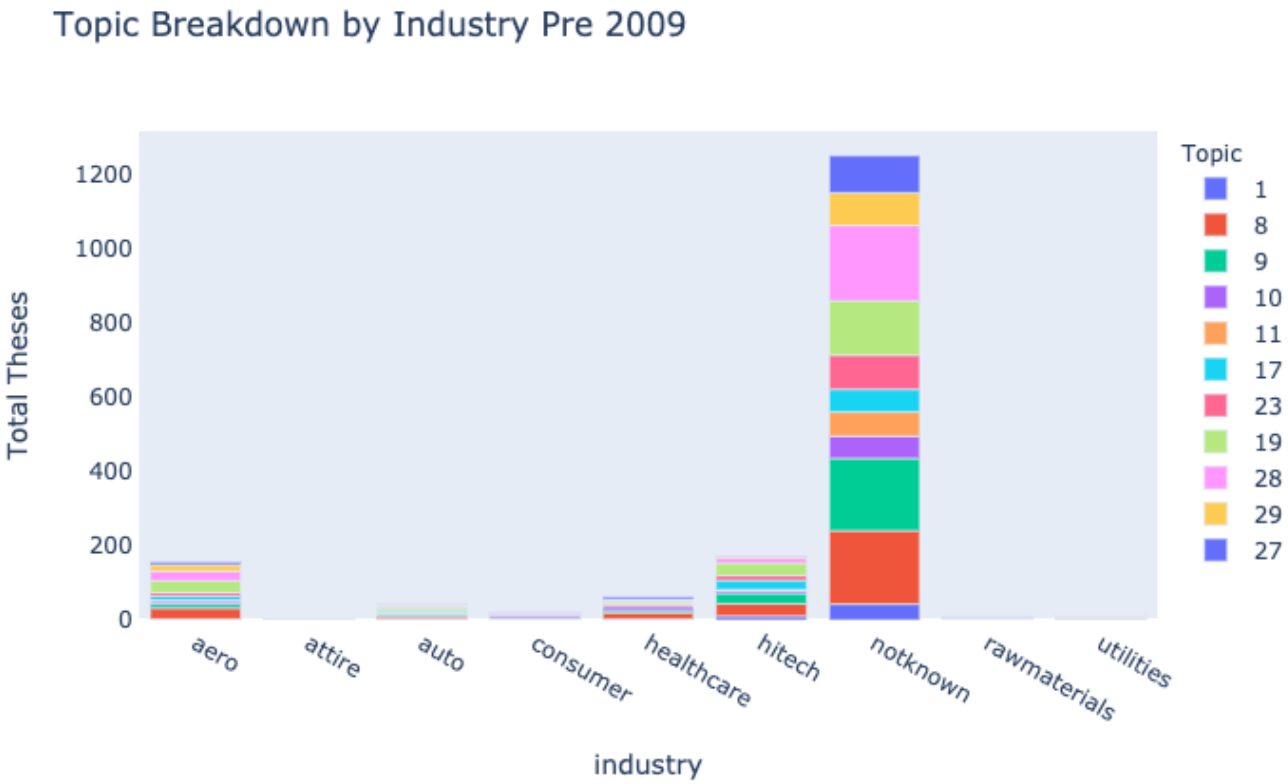

Figure 11. Topic Breakdown by Industry pre 2009

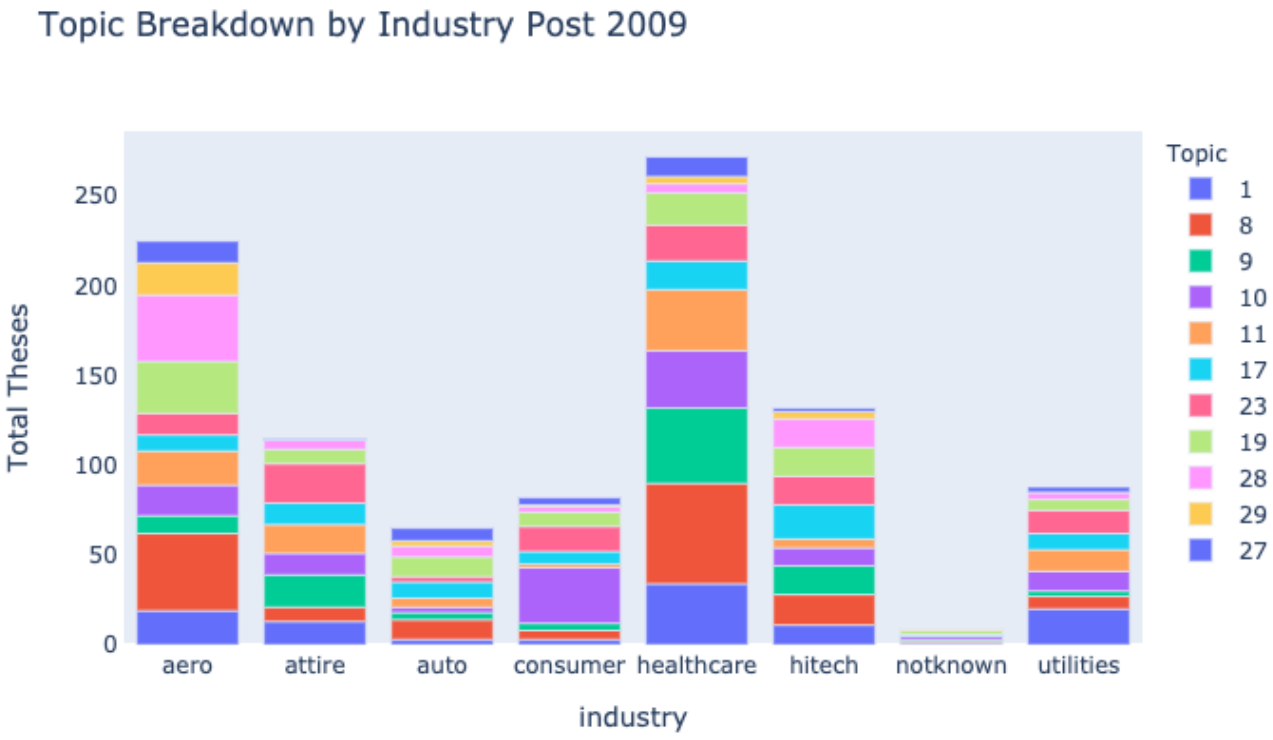

Figure 12. Topic Breakdown by Industry post 2009