Word Has It: Text Analytics for Topic Modeling of MODSIM Track Papers

Jay Gendron
United Services Automobile Association (USAA)
Chesapeake, Virginia
gerald.gendron@gmail.com

Gage Morgan
The Port of Virginia
Norfolk, Virginia
gtmorgan12@outlook.com

ABSTRACT

Big data and data science are part of mainstream society, as evidenced by the many articles and social media posts presenting artificial intelligence and autonomous vehicles in diverse contexts. The volume and persistence of these topics show their rise in public interest. MODSIM launched the Analytics and Decision-Making Track in 2014 as recognition of the then burgeoning field of data science and the importance it would have on industry. This leadership decision continues today, resulting in a body of work over five years that is discernibly different from the other MODSIM tracks.

This paper provides a five-year retrospective with an analysis of the papers published under the MODSIM Analytics and Decision-Making Track. It uses text analytics to investigate the content of the track with MODSIM papers from 2014 to 2018. After presenting a background of data science and current trends, the research question provides a foundation for data processing and analysis. The paper offers additional detail on the analytic techniques appropriate for unstructured text data followed by the specific results of these techniques. In this way, the paper not only provides measures of analytic content within MODSIM but also introduces readers to a look at the workflow underlying text analytics – a technique gaining more popularity since 2018. This paper closes with key findings that analytics tracks focused on contemporary issues in industry are value added to the MODSIM brand and these tracks will likely continue to attract papers and attendees.

Keywords: big data, text analytics, machine learning, topic modeling, cluster analysis, MODSIM, data science, unstructured data, decision making, analytic workflow

ABOUT THE AUTHORS

Jay Gendron is a data scientist with USAA. He is a business leader and algorithmic creative on a quest to show how good questions and compelling visualizations make analytics accessible to decision makers. He is a machine learning engineer, an award-winning speaker, founder of Data Science Management, a speaker at George Mason University, and mentor to data science graduate students. Jay is the author of *Introduction to R for Business Intelligence* as well as book chapters published on various data science topics. For more information, please visit www.linkedin.com/in/jaygendron.

Gage Morgan is an undergraduate student of Business Information Systems and Data Analytics at Old Dominion University. He is a longshoreman at the Port of Virginia and autodidact whose exposure to data-driven innovation has given rise to a fervid interest in the fields of artificial intelligence, operations research, and game theory. He believes that mathematical modeling and simulation should be a catalyzing force behind any decision-making cohort. For more information, visit www.linkedin.com/in/gagemorgan.

Word Has It: Text Analytics for Topic Modeling of MODSIM Track Papers

Jay Gendron
United Services Automobile Association (USAA)
Chesapeake, Virginia
gerald.gendron@gmail.com

Gage Morgan
The Port of Virginia
Norfolk, Virginia
gtmorgan12@outlook.com

INTRODUCTION

> Data science is about the data as much as a book is about the words.
> *– Jay Gendron, George Mason University lecture, 2018*

A book imparts wisdom – data science divulges insights. A book is great when an author skillfully arranges ordinary words in an extraordinary way. Modern marvels of mobile technology can read your book to you, word for word. Abundant storage allows you to carry thousands of books and documents on your phone; and, readily available cloud resources means you can download books on a just-in-time basis. But, can computers discern the meaning of words to help you organize your collection? The answer is a resounding, "Yes." Data science allows computers to not only digest the words but also understand their meaning – their semantics. Machine learning lies at the heart of this paper to analyze words and documents in the collection of MODSIM papers. It considers the lasting results of a decision by MODSIM World 2014 Leadership to create an Analytics and Decision-Making Track.

This paper consists of six main sections: it opens with a summary of data science from the early days of data in business to current interest in artificial intelligence and deep learning. Section two presents the problem statement and research question focusing the analysis. Section three presents techniques available for text-based problems and the methodology used for this work. Section four presents the analysis pipeline to prepare, explore, and analyze text – allowing readers to consider how they might approach text-based data. Section five provides analytic results, discussion, and key findings from the topic modeling used to learn semantics from the words in a collection of documents. The sixth and final section presents three conclusions rendered from the analysis.

INTEREST IN DATA SCIENCE CONTINUES TO GROW

Since the advent of the digital computer, information collection has grown more sophisticated as companies focus on gathering granular insights from data. Industry leaders like Amazon and Google have deployed cybernetic infrastructures in the form of search engines, digital assistants, and smart devices to predict consumer behavior and customize marketing campaigns. Data science reveals patterns and meaning from data collected on a global scale. This is true for numerical data as well as an unstructured corpus of text. Despite origins in the 1970s (Batarseh, Gendron, Laufer, Madhavaram, & Kumar, 2018), the term "data science" broke onto the popular scene when Davenport and Patil (2012) published "Data Scientist: The Sexiest Job of the 21st Century" in *Harvard Business Review*. The term seemed to stimulate more interest than earlier quantitative disciplines like business intelligence, applied statistics, and analytics; and it expanded on significant technical works such as *The Fourth Discipline* that shared energizing essays on "data-intensive science" (Hey, Tansley, & Tolle, 2009).

Roles in the field of data science have risen by 650% over the last five years (LinkedIn, 2017), contributing to related jobs like "machine learning engineer" and "big data developer" that have become some of the fastest-growing jobs in the United States. LinkedIn (2018) also reports labor gaps exceeding 150,000 people skilled in data science. In addition to computational and statistical skills, a well-qualified data science professional:

> **D**evelops increasing domain expertise to ask better questions,
> **A**sks questions before collecting data,
> **T**inkers with data to generate hypotheses and drive modeling, and
> **A**rticulates actionable model results through data storytelling.

PROBLEM STATEMENT AND RESEARCH QUESTION

The growth and interest in data science led to a change in MODSIM technical offerings. The MODSIM World 2014 Committee Members decided to include two new tracks: Analytics and Decision Making as well as Entertainment, Sports, Media, & Visualization (later Visualization and Gamification). Both debuted in the 2014 conference.

Problem Statement

On the five-year anniversary since that decision, the authors wanted to know if it had a lasting impact on the MODSIM World Conference. Table 1 provides the distribution of 208 papers published in the *Proceedings of MODSIM World* (2019a) by year and track over the five-year period covering 2014 to 2018.

A typical track offers about 10 papers in any given year. The analytics and the science tracks display a balanced offering from year to year. The other two tracks show a relationship where the Training and Education Track (averaging 13 papers each year) absorbs papers from the Visualization and Gamification Track (averaging 8 papers each year). With this context in mind, the authors developed a research question to guide analysis.

Table 1. MODSIM Corpus. Count by Year and Track

Year	MODSIM Tracks				Total
	Analytics and Decision-Making	Science and Engineering	Training and Education	Visualization and Gamification	
2014	10	9	15	4	38
2015	10	12	10	12	44
2016	10	10	11	8	39
2017	11	9	17	7	44
2018	12	10	13	8	43
Total	53	50	66	39	208

Research Question

> *Is there a discernible difference in the essential qualities or characteristics of papers submitted to the MODSIM Analytics and Decision-Making Track, compared to the other three tracks?*

METHODOLOGY

Various methods are available to conduct text analytics. The choice of methods depends on the research question and the expected output. In this situation, the research question only required an understanding of the structure within text data. It did not require prediction or inference. Clustering and topic modeling satisfy this use case.

Clustering

Cluster analysis is one of many forms of unsupervised learning used to find the underlying structure of a dataset, such as a corpus of text documents. Cluster analysis is a family with dozens of different algorithms – each designed with a use case in mind. Figure 1 illustrates nine clustering algorithms responding to six different dataset structures.

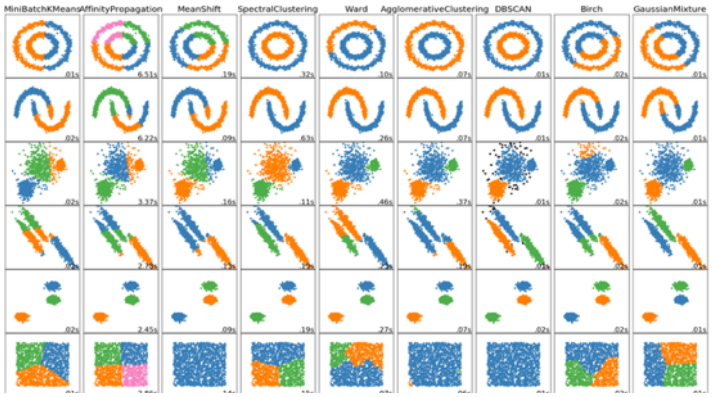


Figure 1. Comparison of Clustering Algorithms on Toy Datasets (Buitinck et al., 2013)

The MODSIM corpus use case possessed two dominant characteristics suitable for agglomerative clustering: numerous clusters and non-Euclidean distances (Buitinck et al., 2013). Clustering relies on computational methods to recast the text data into a numerical form. Distance measures result from pairwise calculations of those numbers.

Representing Text Numerically

Vector space allows for the numerical representation of text data. Each document is a collection of words. The set of unique words across all documents in the corpus forms the dimensions of the vector space. Consider a toy example with a corpus containing three documents, each having just one sentence:

- Document 1: The boy hit the green ball.
- Document 2: The ball hit the boy.
- Document 3: The boy has a green ball.

The seven unique words establish a word vector of the corpus in seven-dimension vector space. Regardless of vector-space dimensionality, the word vector calculations are matrix operations in two-dimensional space. Figure 2 shows two of the more common document vector numerical representations for text data: bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF).

Word Vector	the	boy	hit	green	ball	a	has
BOW for doc1	2	1	1	1	1	0	0
BOW for doc2	2	1	1	0	1	0	0
BOW for doc3	1	1	0	1	1	1	1

Word Vector	the	boy	hit	green	ball	a	has
TF for doc1	2/6	1/6	1/6	1/6	1/6	0	0
TF for doc2	2/5	1/5	1/5	0	1/5	0	0
TF for doc3	1/6	1/6	0	1/6	1/6	1/6	1/6
N / df_t	3/3	3/3	3/2	3/2	3/3	3/1	3/1
IDF for terms	0	0	0.4	0.4	0	1.1	1.1

Figure 2. Document Vector Representations. Bag of Words (left panel) and TF-IDF Components (right panel)

Much like it sounds, a bag-of-words is the count of each word based on their distribution across a document. This document-term matrix is a simple way to numerically represent the document vectors relative to the word vector. The right panel of Figure 2 shows the components of a TF-IDF formulation prior to calculating the document-term matrix shown in Figure 3. Each component shown in Equation 1 derives from Manning, Raghavan, and Schütze (2009).

Word Vector	the	boy	hit	green	ball	a	has
TF-IDF for doc1	0	0	0.07	0.07	0	0	0
TF-IDF for doc2	0	0	0.08	0	0	0	0
TF-IDF for doc3	0	0	0	0.07	0	0.18	0.18

Figure 3. TF-IDF Document-Term Matrix

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t = \frac{f_{t,d}}{\sum_{t' \in d} t'_d} \times \ln \frac{N}{df_t} \quad (1)$$

The TF-IDF formulation adjusts the numerical representation from the bag-of-words approach using two components: a) the $tf_{t,d}$ component, from normalizing the frequency a term appears in a document, $f_{t,d}$, by the total number of terms in the document, $\sum_{t' \in d} t'_d$; and b) the idf_t component, from taking the natural logarithm of the quantity resulting from dividing the total document count in the corpus, N, by the number of documents containing each term, df_t.

The resulting document-term matrix (DTM) provides a sense of the importance each word has in a corpus. In Figure 3, observe how TF-IDF removed the impact of three words – "the", "boy", and "ball" – because they appear in all documents. The normalized and scaled weights are larger for two words – "a" and "has" – because they are rare in the corpus – in fact, unique in this case – and rewarded for their value in information retrieval (Manning et al., 2009).

Measuring the Distance Between Pairs

A common distance metric for text-based data is cosine similarity. This technique does not suffer from the curse of dimensionality, as experienced with large text corpuses containing hundreds or thousands of unique words, and hence thousands of dimensions (Weiss, Indurkhya, Tong, & Darerau, 2005). Cosine similarity measures the angle in n-dimensions of two document projections, $\boldsymbol{A}$ and $\boldsymbol{B}$, in the word vector space. In this analysis, that is based on either the bag-of-words or the TF-IDF numerical representation.

Equation 2 provides the trigonometric calculation for calculating distance based on cosine similarity.

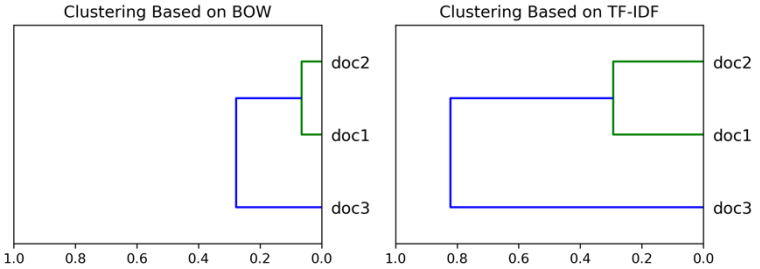
$$\boldsymbol{distance = 1 - similarity(d_A, d_B) = 1 - cos\theta = 1 - \frac{A \cdot B}{\|A\| \, \|B\|}} \quad (2)$$

Two document vectors are closer when their distance value decreases – and similarity increases – because the angle separating them gets smaller. They are identical when $\theta = 0$; $\{\cos(0) = 1\}$. For document vectors orthogonal to one another, then $\theta = 90$; $\{\cos(90) = 0\}$ and the documents are not similar at all. Using the value (1 – similarity) provides a distance for use in agglomerative clustering, where smaller distance indicates greater similarity (Batarseh, Nambiar, Gendron, & Yang, 2018). The resulting matrix of pairwise distance measures allows for visualization.

Visualizing Distance Results

Plotting a tree pattern, or "dendrogram", begins with the pair of items having the smallest distance value and joining them. The next closest pair creates a new branch or merges with an existing branch. This continues until all elements form a single cluster. "Agglomeration" is a process of joining individual elements from the "bottom-up" beginning with those having the smallest distance – or, in other words, the largest cosine similarity (Barhak, 2018).

Dendrograms in Figure 4 show visualizations of the three-document corpus toy example.



	Distance = 1 – Cosine Similarity	
Pair Compared	**BOW DTM**	**TF-IDF DTM**
doc1 vs. doc2	0.06	0.29
doc1 vs. doc3	0.28	0.82
doc2 vs. doc3	0.38	1.00

Figure 4. Dendrogram Using BOW (left panel), TF-IDF (center panel), and Distance Values (right panel)

Documents form the leaves at the start of the tree on the right-hand vertical axes. The horizontal axes in the left and center panels represent the distance values where the joins occurred, as shown tabularly in the right panel. Documents joining closer to the leaves are more similar than ones that join at larger distance values. Furthermore, the distance between joins helps reveal segmentation among the clusters. In cases where join heights are nearly the same, there is little distinguishing those clusters or individual documents – meaning they are very similar (Manning et al., 2009).

Interpreting the toy example results, *doc1* and *doc2* merge into clusters first based on distance values calculated with cosine similarity. These two documents are closest in terms of distance with values of 0.06 and 0.29 based on the BOW DTM and TF-IDF DTM, respectively. These two documents also show the highest similarity in the dendrograms. The next closest pair is *doc1* and *doc3,* and those joins creates a single cluster. The structure of the dendrograms shows *doc2* and *doc3* are the least similar. The properties of the TF-IDF technique focuses on words of interest. A review of the DTM in Figure 3 shows *doc2* and *doc3* share no words of interest. This makes the dot product $\boldsymbol{A \bullet B}$ is zero. A review of Equation 2 shows the distance for this pair equals 1.00, meaning they are not similar at all.

Yet, reviewing the content of *doc1* and *doc2* reveals they tell very different stories – one of triumph and the other of injury – yet they clustered early and tightly. This is because agglomerative clustering has limited ability to discern semantics. The technique works well when dealing with a corpus of distinct topics, for example finances, pro football, and Starbucks coffee. The technique is less powerful when dealing with documents made of similar topics and words.

Topic Modeling

During initial planning of this analysis workflow, the authors chose to use agglomerative clustering in exploratory data analysis based on an assumption that the MODSIM World corpus possessed differences in words among analytics and simulation tracks. That exploratory analysis revealed the underlying theme in most MODSIM World abstracts was the topic *simulation*. This is similar to having the words *boy* and *ball* in every document. Based on this exploration,

the authors looked at other techniques such as topic modeling to detect semantic differences. Topic modeling is a text mining tool that finds hidden, semantic structures within a corpus. It uses sophisticated statistical sampling and machine learning techniques like Gibbs Sampling and the Latent Dirichlet Allocation algorithm. The details are beyond the scope of this paper, but Graham, Weingart, and Milligan (2012) provide a useful description of the technique. Meanwhile, Figure 5 provides an illustration to depict the mechanism involved with topic modeling.

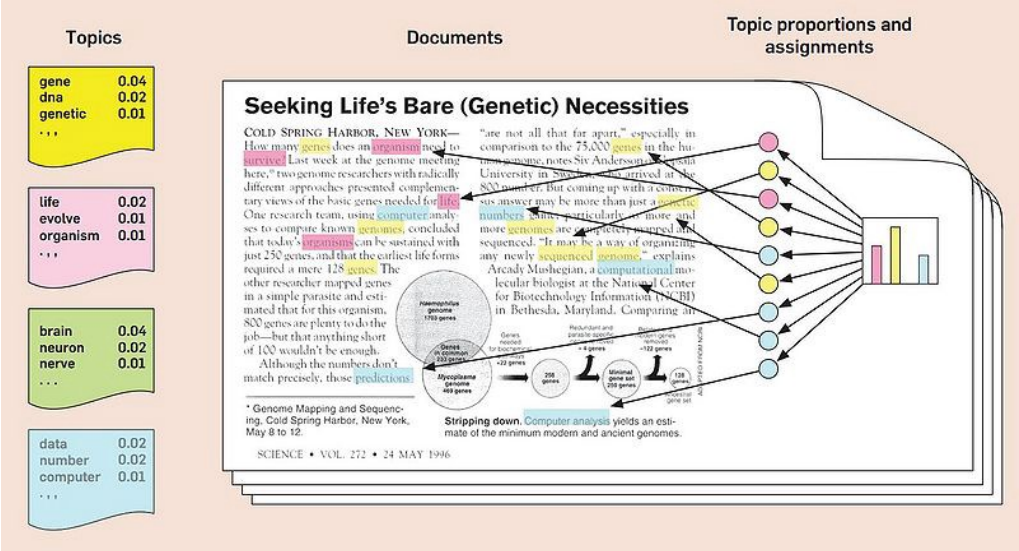


Figure 5. Graphical Depiction of the Topic Modeling Technique (Ghosh, 2018)

The document shown in the figure is one of a larger corpus of documents. Three topics emerged from topic modeling based on finding collections of words that appear together with some statistical significance and not due to random chance. Having found several topics based on word patterns, the proportions of those topics go into the various analyses to provide a numerical measure of topic strength by document. Those topic strengths output as a topic-document matrix. This is like the DTM seen in clustering, but the topic-document matrix values provide information useful for clustering or direct observation. This is because the topics imply meaning, unlike words in a DTM.

ANALYTIC WORKFLOW GUIDED BY THE DATA PIPELINE

Data science relies on a data pipeline to organize and execute workflow. It is composed of three major areas: data engineering, exploratory data analysis, and data modeling. There are many versions of data pipelines, but these three areas are consistent among versions. The analytic workflow begins with data engineering.

Data Engineering – Processing the Raw Data

Another term for data engineering is ETL – Extract, Transform, Load. This industry term captures the essence of getting data ready for analysis. This is a time-consuming process. Anthony Goldbloom, the CEO of Kaggle, noted in a personal interview, "Eighty percent of data science is cleaning data and the other twenty percent is complaining about cleaning data" (as cited in Gendron, 2016, p. 23).

Extract

This phase of the analysis did indeed consume over 80 percent of the total time expended. That statement is not surprising to a data science professional working in the field as this is typically the case. The raw data for this analysis consisted of the five years of conference papers in PDF format available on the MODSIM World website (MODSIM, 2019a). Many tools exist for batch downloading. This analysis used the Chrono Download Manager (Google, 2018). After downloading the papers as PDFs, the Python BeautifulSoup4 library enabled extraction of key paper metadata from HTML files also provided on the MODSIM website for each paper. The metadata provided the track, filename, author(s), title, and year as HTML tags. A dataset held all this information and aided in data transformation.

Transform

Transformation includes two functions: cleaning data to ensure quality and adapting raw data for purposes of analysis. Metadata on each paper allowed the conversion of PDF to text format. This included establishing a consistent file name that served as a document label. The text required cleaning because a small percentage of PDF extractions yielded corrupt text – either **"a d d i n g"** spaces between letters or **"combiningwordsinto"** long strings. Three transforms occurred after data cleaning: a) adding missing keywords such as "ABSTRACT" to aid text extraction; b) extracting just the abstract to a separate data file; and c) changing the file encoding to Unicode (UTF-8) in all instances.

Load

The transformed data loads into a file structure for use later in analysis. The file structure for the exploratory data analysis consisted of a root directory holding all 208 papers in a single folder. This changed after exploratory analysis to one folder with twenty documents containing the combined abstracts from each of four tracks across five years.

Exploratory Data Analysis – Discovering Relationships in the Data

Clean, transformed data moves to the next phase of the pipeline. John Tukey (1977) coined the term exploratory data analysis with a book of the same name. It is still true today that humans are very good at seeing patterns and trends emerge from raw data. Machines have become very good at this phase; however, the analyst must still internalize the underlying data structures through exploratory data analysis. In short, two clustering techniques used in this phase shaped the final modeling direction. The first clustering technique used a unigram model (only single words) with a cosine similarity distance measure on a TF-IDF matrix covering all 208 abstracts. The second technique used a bigram model (all single words and ordered word pairs) with a cosine similarity distance measure on a bag of words, also composed of the text from 208 abstracts as individual documents. The plots in Figure 6 show the exploratory results.

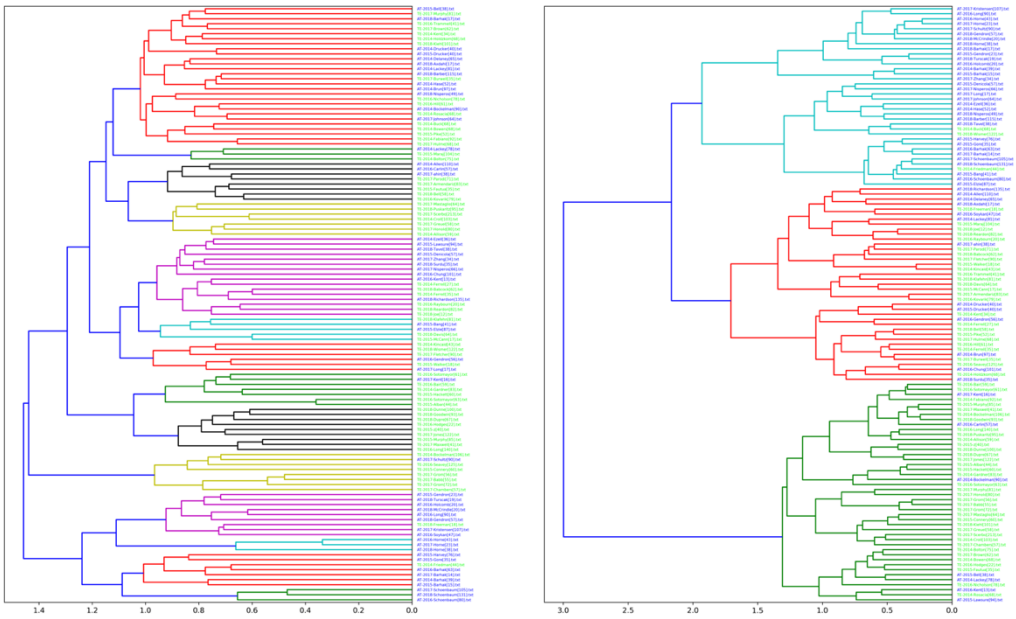


Figure 6. TF-IDF Unigram Model (left panel); Bag-of-Words Bigram Model (right panel)

The figure shows two of the many exploratory runs on the 208-document corpus resulting in dendrograms based on a cosine similarity distance. Because of the small size of the document labels on the right-hand axis, the authors coded the plots to display documents in the same MODSIM track to appear in the same color – black for Training and Education (TE) and light green for Analytics and Decision-Making (AT). The left panel shows the more sophisticated TF-IDF technique. It credits a document for words it contains, like the bag of words technique. It also penalizes a document for words it has in common across the entire corpus. This technique works quite well for information retrieval (Manning et al., 2009), but it proved too aggressive for the MODSIM corpus having great similarity in word content. The right panel shows a simple bag of words (the TF portion) and a bigram model that accounts not only for each word but also each consecutive word pair as they appear in the document (e.g. "the training event" becomes "the

training" and "training event"). Exploratory work showed three clusters: TE, AT-TE mix, and AT. This does show some discernible difference of the Analysis and Decision-Making track; however, other techniques exist that could show greater segmentation.

Data Modeling – Answering Questions with Insights

Exploratory data analysis indicated that the research question provided a useful hypothesis to model. It also showed the cluster analysis would not work on a small dataset composed of relatively homogenous documents. Authors turned to a technique called topic modeling based on the Latent Dirichlet Allocation (LDA) approach. Whereas the bag of words and TF-IDF methods account for words present in a document, topic modeling provides insights beyond the bag of words to a set of topics that lie latent within documents. Topics are a probability distribution over the words present in a document (Steyvers & Griffiths, 2007). Topic modeling can uncover semantic meaning in documents based on word relationships. The MALLET machine learning application for statistical natural language processing generated the topic models (McCallum, 2002). Steyvers and Griffiths (2007) note that one can derive topics from a corpus and use those to answer questions about document similarity. Modeling choices include the number of topics to model (in this analysis ranging from 6 to 24), the use of stop words (removal of common words like "at", "the", and "in"), using single words or including word sequence pairs (unigrams and bigrams), and a decision to combine all abstracts from each track-year combination into a single document (yielding 20 transformed documents for modeling).

TEXT-BASED ANALYSIS IDENTIFIES SEGMENTATION IN THE MODSIM CORPUS

Permuting the modeling choices generated numerous outputs. The results for a 12-topic LDA unigram model provided here include the topic contents, their relative importance, and a heatmap of topic distribution by document. Table 2 provides the 12 topics resulting from the LDA unigram model run against the 20 documents combined by track-year.

Table 2. Topics Resulting from 12-topic LDA Model Using Corpus of Track-Year Abstracts

Topic	Weight	Words Composing the Topic
0	0.02661	distributed school world students visualization brain virginia taking workload driving instructors stories mobile mishap land lg-raid governors academy efficacy pathway
1	0.02328	learning access air fidelity motion training bcs simulators thinking awsim norad interoperability mission pilots soldier e-learning fluids research experiences enders
2*	0.0865	disease data farming world scenario lrts war tutoring insights stress instructors spread impact animal food sharing results sky adoption health
3	0.05029	joint test real aircraft scenarios perceptual selection sets section ssat radio phase cognitive gunnery voice decision immersive office planning workforce
4	0.02259	research interoperability wave phase lvc m&s ionosphere case likelihood pcs key shm agents satcom language angles extreme state response modelling
5*	**0.54555†**	data model decision tool modeling population analysis techniques analytics planning approach dynamic acquisition engineering predictive algorithms strategies threats impact areas
6*	0.076	intelligence nuclear operators algorithm operations manikin pedestrian dhs enforcement law hapmed eye-tracking cultural selection text analysts creating user encounter plants
7*	0.03696	mts constraints farming optimization job resources objectives pilot mining disease statistics schedule proposed processes populations igtm simple recognition researchers scheduling
8	0.07104	visualization platform specific device real quality medical reality authors stem scheduling interaction sim underlying models ins defender typically autonomous sensors
9	**1.5262‡**	training simulation paper system systems data design environment learning development performance virtual technology process model research approach study provide based

10	0.03893	learner joint training authoring ar/av jtse jtt assessment force services planning speech efficacy database psychomotor terrain aviation modular operationally recognition
11	0.06091	aircraft lta material carriers engineering rules work vehicles motion networks space engine traffic physics palms amste air organizations modeling security

** topic is associated with analytic domain: 2 – data mining; 5 – predictive analytics; 6 – text analytics; 7 – health analytics*
† the second largest weight but the topic is associated with analytics and predictive algorithms (i.e., machine learning)
‡ the largest weighted but the topic is associated with core words found across all documents in the corpus (see Figure 7)

The model output designed for this analysis includes the topic, the weights showing the relative importance of each topic in the corpus relative to one another (Graham et al., 2012), and the 20 most probable words associated with each topic after model estimation (McCallum, 2002). Note that the topics are merely numeric markers – the model does not define the topics; it simply identifies them. One can read the words composing each topic and begin to see they point to a topic that lies latent within the corpus. Four topics annotated with a (*) are analytic in nature. The two largest weights are **bold** to indicate interesting findings – topic 5 has the second highest weight and is one of the four topics associated with analytics and decision-making; topic 9 has the highest weight – and by a large proportion – because of the abundance of those words in the corpus. However, as discussed in TF-IDF, the abundance of this topic means it is noise in the sense that these words appear in every track-year combined collection of abstracts (see Figure 7).

Figure 7 provides two topic-document matrices. The heatmap is a visual summary of the numerical values present in the cell of each matrix. Those values, in turn, indicate the proportion each topic is present in each document, from 0-56 percent – each row summing to 100 percent. Both panels contain the same values; however, the ordering of topics in the right panel is based on a clustering algorithm showing topic similarity. Steyvers and Griffiths summarize the relationship of words, topics, and documents as "two words are similar to the extent that they appear in the same topics, and two documents are similar to the extent that the same topics appear in those documents" (2007, p. 12).

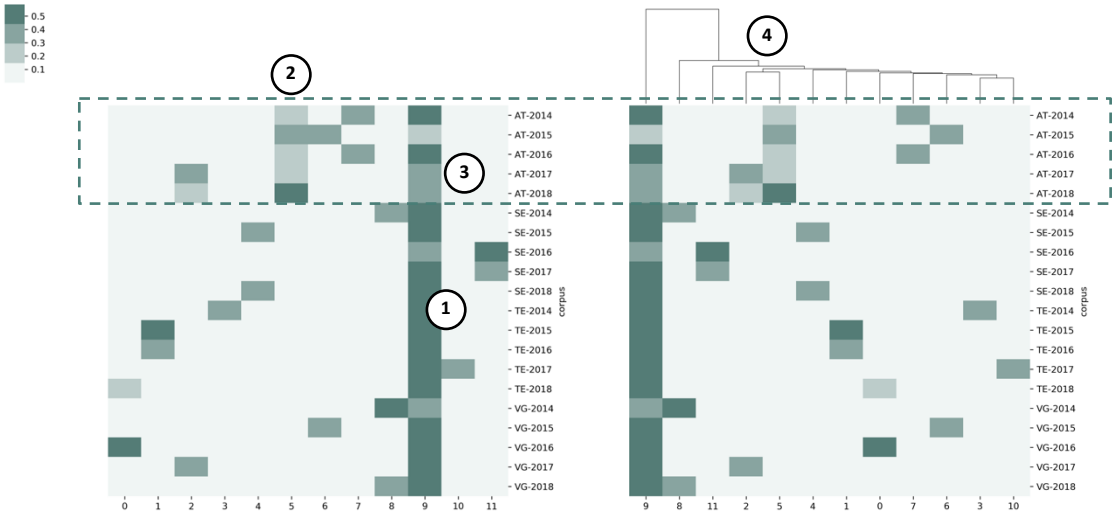


Figure 7. Topic-Document Matrices. Heatmaps Show the 12-topic LDA Unigram Model Using Corpus of Track-Year Combined Abstracts (left panel); Identical LDA Output Clustered by Topic (right panel)

Numerical tags in Figure 7 annotate key findings observed in the topic-document matrices:

1. Topic 9 consists of core words appearing in every track-year collection of abstracts. One expects such a result in a body of modeling and simulation papers, but it provides no distinguishing value in segmentation
2. Topic 5 is a significant topic based on weight (Table 2) and appears in each year since the Analytics and Decision-Making Track began. The trend is stable across years, with a spike in topic strength in 2018
3. Not only does topic 5 segment within the corpus, but it also reduces the proportion of topic 9 (core word) content. This indicates a more purely focused collection of papers in 2015, 2017, and 2018
4. No clusters show strong "join height" of the branches (nearly all join at the same height). Topic 9 does show a clear height separation (see Finding 1). Topics 2 and 5 are more similar based on the branch structure

CONCLUSIONS

Based on these key findings, three conclusions emerge with regard to the initiation of an Analytics and Decision-Making Track at MODSIM World 2014:

1. **There is a discernible difference in the characteristics of papers appearing in the MODSIM Analytics and Decision-Making track as compared to other tracks.** The analysis indicated that the Analytics and Decision-Making Track represents a segment of MODSIM papers differing from the other tracks. This segment, based on word usage and semantics, emerged computationally using text analytics techniques without labeling of papers as seen in supervised machine learning. Segments revealed themselves to a small degree in agglomerative clustering (due to the close relationship all the papers have regarding an underlying theme of modeling and simulation). Nonetheless, there were latent indicators due to semantics and word combinations that became clearer segments in topic modeling using LDA techniques.

2. **Leadership by the MODSIM Committee provided a new dimension to the MODSIM World 2014 participants.** MODSIM Committee recognition of the importance of data science resulted in a new track aligned to a growing area of industry interest – namely analytics in a business context. The decision to create a new MODSIM track continues to reap dividends as evidenced by a consistent volume of papers showing a slight increase in the last two years. A visual inspection of titles at the MODSIM World website (MODSIM, 2019a) also shows the content of the papers has steadily grown in analytic rigor over the last five years.

3. **Conference tracks focused on contemporary issues facing industry are value added to the MODSIM World brand and will likely continue to attract papers and attendees.** In the final analysis, it appears that conference leaders who stay abreast of industry trends and adapt their conference offerings to meet those trends earn rewards. It is noteworthy that the conference leadership has once again responded to industry trends by adapting the Visualization and Gamification Track to address augmented reality, virtual reality, and mixed reality (MODSIM, 2019b). This new track, entitled Cross Reality (XR), is a welcome addition to the conference by bringing an important area of work to participants.

ADDITIONAL RESEARCH

The results of the text-based analysis provided evidence of segmentation between those MODSIM papers published under the Analytics and Decision-Making Track versus the other three conference tracks. Having this proof of concept, additional analytic approaches may yield more refined results. For instance, data for the analysis included only the abstracts of each paper. A primary reason for this was the level of additional cleaning required to include the full text of each paper. Analytic results now support additional cleaning and the inclusion of full paper content.

Furthermore, this analysis focused on unsupervised techniques to find underlying structure and segmentation within the MODSIM corpus. Additional analysis using supervised machine learning techniques is also possible. This would require the hand-labeling of the MODSIM corpus into two or more categories (i.e., analytic and simulation, or the four track categories). Supervised learning techniques such as extreme gradient boosting or deep learning with neural networks could allow researchers to classify the MODSIM corpus. It could also allow for the inference of track placement of papers in future conferences.

REPRODUCIBILITY INFORMATION

Jupyter Notebook server version 5.6.0 on Anaconda 64 bit with Python 3.6.6, NumPy version 1.15.1, pandas version 0.23.4, scikit-learn version 0.19.2, BeautifulSoup version 4.6.3, requests version 2.19.1, matplotlib version 2.2.3, and seaborn version 0.9.0. Executed on a Mac with 8 GB memory, a 1.8 GHz Intel Core i5, and macOS High Sierra version 10.13.6. Code for ETL and analysis as well as the prepared load datasets used for analysis are available on GitHub at https://github.com/jgendron/org.modsim.topic.modeling.tracks.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the MODSIM Committee for their assistance in providing access to the published papers from MODSIM World 2014. A special thanks goes to Stefani Werner, the Deputy Program Chair for MODSIM World 2019, in leading the effort to post the 2014 papers on the MODSIM site.

REFERENCES

Barhak, J. (2018). "Visualization and pre-processing of intensive care unit data using python data science tools" in *Proceedings from MODSIM World 2018.* Retrieved from http://modsimworld.org/papers/2018/MODSIM_2018_Barhak.pdf

Batarseh, F., Gendron, J., Laufer, R., Madhavaram, M., & Kumar, A. (2018, November 20). A context-driven data visualization engine for improved citizen service and government performance. *Modeling and Using Context, 2,* 1-21. DOI: 10.21494/ISTE.OP.2018.0303

Batarseh, F. A., Nambiar, G., Gendron, G., & Yang, R. (2018). Geo-enabled text analytics through sentiment scoring and hierarchical clustering. *2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics)*, Hangzhou, 1-4. DOI: 10.1109/Agro-Geoinformatics.2018.8475993

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., … Varoquaux, G. (2013). "API design for machine learning software: experiences from the scikit-learn project" in ECML PKDD workshop: Languages for data mining and machine learning, 108-122. Retrieved from http://www.ecmlpkdd2013.org/wp-content/uploads/2013/09/lml2013_api_sklearn.pdf

Davenport, T. H., & Patil, D. J. (2012, October). Data scientist: The sexiest job of the 21st century. *Harvard Business Review, 90*(10). 70-76.

Gendron, J. (2016, August). *Introduction to R in business intelligence*. Birmingham, UK: Packt Publishing.

Ghosh, S. (2018, Mar 17). Topic modelling with latent dirichlet allocation (LDA) in pyspark [web log image]. Retrieved from https://medium.com/@connectwithghosh/topic-modelling-with-latent-dirichlet-allocation-lda-in-pyspark-2cb3ebd5678e

Google. (2018, July 11). Chrono Download Manager (Version 0.10.0) [Software]. Available from https://chrome.google.com/webstore/detail/chrono-download-manager/mciiogijehkdemklbdcbfkefimifhecn?hl=en

Graham, S. Weingart, S., & Milligan, I. (2012, September 2). Getting started with topic modeling and MALLET. *The Programming Historian, 1,* Retrieved from https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Hey, T., Tansley, S. & Tolle, K. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research. Retrieved from: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf

LinkedIn. (2017, December 7). LinkedIn's 2017 U.S. emerging jobs report [Web log]. Retrieved from https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report

LinkedIn. (2018, August 10). LinkedIn Workforce Report [Web log]. Retrieved from https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval*. Cambridge, England: Cambridge University Press.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit (Version 2.0.8) [Software]. Available from http://mallet.cs.umass.edu

MODSIM. (2019a). MODSIM world conference papers. Retrieved from http://modsimworld.org/conference-papers

MODSIM. (2019b). MODSIM world tracks. Retrieved from http://modsimworld.org/tracks

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic model. In Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.), *Latent semantic analysis: A road to meaning.* Mahwah, NJ: Laurence Erlbaum.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, PA: Addison-Wesley.

Weiss, S. M., Indurkhya, N., Tong, Z., & Darerau, F. J. (2005). *Text mining. Predictive methods for analyzing unstructured information.* New York, NY: Springer.