The latent topics underlying US battery storage bills in the 116th Congress

A Text Mining Project applying
Non-negative Matrix Factorisation
and Probabilistic Latent Semantic Analysis

Mirjam Grünholz

Chaletweg 5

4600 Olten

Matriculation number: 13-213-350

ETH Zürich

Master in Comparative and International Studies

Seminar: Methods of Unsupervised Machine Learning

Autumn 2019

Professor Dr. Marco Steenberg

4 January 2020

Table of Content

1. Introduction

In the changing world where renewable energies are become increasingly important and electric vehicles increasingly prominent, new challenges arise. The sun does not always shine, and the wind does not always blow. Renewables are often produced decentralised and locally. And electric cars still suffer from range and battery limitations. Despite of diversified renewable energy production technologies and rapid improvement in electric vehicle technology, battery storage remains one of the main challenges in the field. Efficient and smart policies would be a solution (Beuse, Schmidt, and Wood 2018). But before asking how smart policies are, it should be investigated what different battery storages policies deal with. On the case of the US Congress, the following research question will be answered: *What are the latent topics underlying the US battery bills of the 116th Congress?*

In order to answer this question, the text analysis methods of Non-negative Matrix Factorisation (NMF), and Probabilistic Latent Semantic Analysis (PLSA) are applied. The analysis is part of a larger data collection project investigating US battery storage bills over the last 20 years. This paper is a preliminary analysis applying NMF and PLSA on a set of 45 US bills of the 116th Congress which deal with batteries in the context of energy storage and electric vehicles. The results, particularly the uncovered latent topics and the critical investigation of the methods, are thought to motivate the further steps in the larger data collection project.

2. Data

The texts chosen for this analysis are US battery storage bills from the 116th Congress (2019-2020). The data was retrieved from the Congress.gov homepage (2019) which provides information on the US Congress and its legislation since the 93th Congress (1973-1974). 37 bills were found under the search terms "'energy storage' + battery" and 29 bills were covered by the search terms "'electric vehicle' + battery". The 12 bills appearing in both categories were only counted once, resulting in 50 documents. 90 percent of those have between two and 261 pages, the other five between 650 and 1976 pages. An initial analysis has shown that this variation heavily biases the analysis, which is why the five longest texts were removed. The remaining corpus contains 45 bills.

Different methods of text analysis are tested on a unique sample of US bills and their applicability to a new case is critically investigated. The data has been collected as part of a larger data collection project on US bills dealing with batteries in the context of energy storage and electric vehicles over the last 20 years. For the current paper, the 116th Congress is chosen as a case to run a preliminary analysis in order to uncover the latent topics underlying these bills. It is expected that nowadays it is more likely to

explicitly observe a variety of latent categories which cover distinct transport and energy storage topics, compared to a Congress 20 years ago. The fact that the 116th Congress is still running but has already introduced more related bills than any Congress before supports this approach. In a later step, the bills could be categorised according to the latent topics. Moreover, the analysis is likely to give evidence on the relevance of individual bills with respect to battery storage policies. Concludingly, this text analysis serves as a pre-processing and its results will drive the decision whether and how to expand the analysis to the entire corpus of bills.

In order to prepare the texts for the analysis, the PDFs were loaded into the software R, version 3.5.2, using the *pdftools* library. The *tm* package was used to remove a list of terms which were likely to add unnecessary noise to the subsequent analysis. This list included double and triple letter (i.e. "aa", "bbb"…), roman numbers (i.e. "iv", "v", "ix"), brackets and paragraph signs "§" which are distinctive for bills. The terms "bill", "bills", "section", "sec." and "act" were removed as well as some words which appeared on every downloaded page (i.e. "frm", fmt", "jkt"). The remaining terms were transformed to lower case, numbers, and punctuation were removed. Stop words, such as common prepositions, pronouns, articles or conjunctions, were deleted (Aggarwal 2018b). Multiple white space characters were collapsed to a single blank (Feinerer 2019). The words were reduced to their stems by using Porter's stemming algorithm integrated in the *tm* library which removes the common morphological endings from English words conflating related words with the same root (Aggarwal 2018a; Porter 2019). The cleaned texts were transformed into a document-term-matrix with 45 texts. Lastly, sparse terms were removed on a threshold of 97 percent, resulting in a matrix with 3'988 instead of 8'633 terms and a sparsity reduction form 92 percent to 86 percent.

3. Methodology

Non-negative Matrix Factorisation (NMF) and Probabilistic Latent Semantic Analysis (PLSA) are applied to the text corpus with the aim of finding a representative and distinctive number of latent topics representing the US battery-storage bills of the 116th Congress.

These methods require the text to be represented as a bag-or-words, whereby the semantic meaning of the sentences is lost. Instead, the frequencies of words per bill are represented in a document-term matrix with n documents indicating the rows and d different terms indicating the columns. The cells show the frequencies of the terms. The matrix is a non-negative, sparse and highly dimensional representation of the collection of bills. The aim is to reduce the original data matrix to a factorised form by reducing the original number of terms to fewer, unobserved, or latent, topics. Dimensionality reduction is possible because the rows in the document-term matrix are highly correlated.

The original data matrix *D* can be represented as the product of two matrices with considerably smaller entries compared to the original matrix. A common representation of the *n x d* document-term matrix *D* is the following:

$$D \approx UV^T$$

where *U* is an *n x k* matrix containing in its row the *k* number of semantic topics with $k << min\{n, d\}$. The content of the matrix shows how strong a bill is associated with each topic *k*. *V* is a *d x k* matrix of terms and latent topics. The original matrix *D* is an approximately linear transformation of the matrices *U* and the transpose of *V*.

Alternatively, the matrix *D* can be represented as the product of three matrices. Multiplying each column of *U* and *V* by their L_2-norm results in a *n x k* matrix Q with documents as rows and latent topics as columns, and in a *k x d* matrix P^T of topics by terms. The products of the L_2-norms of the columns of U and V result in the *r*th diagonal element of the third matrix Σ (Aggarwal 2018c).

$$D \approx Q\Sigma P^T$$

Different constraints on *U* and *V* result in different text mining models. NMF and PLSA are typical methods relying on these dimensionality reductions. They are constrained by nonnegativity (Aggarwal 2018a, 2018c).

3.1 Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) models directly observable word counts in a document-term matrix from a set of hidden variables or latent topics. Semantically related words are grouped, but at the same time the context of the word is maintained. It relies on a part-based representation permitting multiple latent topics to represent one document (Lee and Seung 1999). Concludingly, every cell of the matrices *U* and *V* contains information on the topics each document and every word belong to (Aggarwal 2018c).

NMF aims at minimising the Frobenius norm $\|D - UV^T\|_F^2$ subject to nonnegativity constraints on U and V. Intuitively, the equation $D \approx UV^T$ is minimised in an iterative process based on the minimisation of the Kullback-Leibler (KL) divergence loss function, resulting in a converged, local optimum of the matrix factorisation (Lee and Seung 2001). In R, the *svs* library with KL specification is used.

NMF are highly interpretable and account for the meaning of words in different contexts. Because semantic topics are often related, they relatively accurately cluster topics. A disadvantage of NMF is that it is difficult to fold in new document into the original data matrix.

3.2 Term-frequency Inverse-document-frequency

Depending on the text mining method, some normalisation of the document-term matrix can improve the results. For other methods, especially those based on probabilistic methods, the generative assumption implies that raw term frequencies are better suited (Aggarwal 2018b).

The term-frequency inverse document frequency (tf-idf) normalisation accounts for the phenomenon that the most frequently used terms in a text also have a high occurrence in all the other texts. Contrarily, other terms are distinctive for one specific document. Firstly, the inverse document frequency id_i for every term i is calculated:

$$id_i = log\left(\frac{\#\ total\ documents}{\#\ documents\ containing\ the\ word}\right)$$

The normalisation is completed by multiplying the term frequency x_i with the id_i (Aggarwal 2018b; Gandhi 2018). In this analysis, NMF will be applied on the raw term frequencies as well as on the tf-idf normalised data.

3.3 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is represented as a normalised factorisation splitting the document-term matrix D into three matrices $D \propto Q\Sigma P^T$ (similar to equation introduced in section 3 but with proportionality due to the probability-centric scaling). Q is a $n\ x\ k$ matrix and P is $d\ x\ k$ matrix. Their columns sum to 1 and are interpreted as probabilities. Σ is a $k\ x\ k$ diagonal matrix containing prior probabilities and entries summing to 1.

D with respect to Q, Σ and P is optimised with maximum likelihood. The three matrices are the result of a generative process: The frequencies in the document-term matrix are assumed to be sequentially increased, generated by latent topics. The number of iterations depends on the number of tokens in the corpus (Aggarwal 2018c). A set of latent components η_r are selected with probability Σ_{rr}. Given η_r, Q_{ir} displays the probability of generating document X_i. Given η_r, P_{jr} is the probability of using a specific term T_j. The optimisation process assumes conditional independence between the processes generating documents and topics (Steenbergen 2018).

To overcome the problem of conditioning on the unknown latent topics η, the expectation-maximisation (EM) algorithm is used. Firstly, the algorithm estimates the posteriori probabilities of each document-term pair $P_r(\eta|X_i, T_j)$ (E-step). Secondly, the parameters Q, Σ and P are estimated using conditional

probabilities. These steps are repeated to convergence (Aggarwal 2018c). In R, this optimisation is solved using the *svs* library.

PLSA shares the advantages and disadvantages of NMF discussed above. However, compared to the factorisation model NMF, PLSA is a probabilistic model. This leads to additional advantages: PLSA is intuitive and with its probabilistic representation easy to interpret, and since no assumptions on the topic distribution in a document are made the model can easily be applied to large document collections.

4. Results

4.1 Non-negative matrix factorisation

NMF was applied to the text corpus by retaining four up to seven topics with the goal of finding a representative and distinctive number of topics representing the US battery-storage bills of the 116th Congress. The procedure was repeated with the tf-idf normalised data.

Table 1 shows the topic distribution summarised from the V matrix for the original data (black) and for the tf-idf normalised data (blue) retaining four up to seven topics. Retaining six topics appears to create the most accurate and complete representation of categories. Retaining fewer topics misses out on important battery technology and application perspectives, whereas retaining more latent topics leads to too many too specific categories. The six topics retained cover: 1) land and water management, 2) energy program, 3) energy finance, 4) electric vehicles grants, 5) energy industry and emissions, and 6) US-china zero emission.

Recalculating NMF with tf-idf normalised data leads to a more accurate description of the different categories due to the mechanisms that document specific words receive more weights. Moreover, NMF with normalised data appears to partly extract policy instruments. The second topic captures grid grants, the third solar R&D grants, the fourth grants for green transport, and the fifth carbon taxes.

Furthermore, the U matrix describes with which topic each document is associate most. When comparing the titles in Table 3 with the topics represented in Table 2, the close link between the topics and the titles becomes visible.

Table 1: NMF topic distribution (black = original document-term matrix, blue = tf-idf normalised document-term matrix)

4 Topics	**5 Topics**	**6 Topics**	**7 Topics**
1) land management 1) land water wild park management 2) energy/ electricity program 2) green transport (bus, train) 3) energy industry 3) solar PV grant 4) vehicle emissions 4) china explort vehicle zero emission	1) land management 1) land water wild park management 2) energy/ electricity program 2) broadband electricity grid program 3) energy finance 3) china export vehicle zero emission tax 4) electric vehicle technology 3) solar PV grant 5) Vehicle zero emissions 5) green transport (bus, train) and jobs	**1) land and water management** **1) land water wild park management** **2) energy program** **2) broadband + grid grant** **3) energy finance** **3) resident community + solar PV R&D grant** **4) electric vehicles grant** **4) bank grant for green transport (bus, train) and jobs** **5) energy industry + emissions** **5) emission tax on carbon and gas** **6) US-China zero emissions** **6) China trade exports huawai**	1) land and water management 1) land water wild park management 2) energy program 2) broadband + CO2 + health 3) electric vehicle development 5) green transport (bus, train) and jobs 4) energy credit + vehicle emission 4) solar PV grid 5) energy industry 5) R&D grant 6) US-China 6) China trade exports huawai/ telecom 7) public grant 7) zeroemission vehicles + wildfire + credit

Table 2: Most prevalent bills per topic resulting from NMF with six latent topics using the original document-term matrix

Topic	Title
1	S.47: John D. Dingell, Jr. Conservation, Management, and Recreation Act
2	H.R.2741: Leading Infrastructure for Tomorrow's America Act
3	H.R.3423: National Green Bank Act of 2019 H.R.5185: Green New Deal for Public Housing Act
4	H.R.2170: Vehicle Innovation Act of 2019 S.1085: Vehicle Innovation Act of 2019
5	H.R.4520: Modernizing America with Rebuilding to Kickstart the Economy of the Twenty-first Century with a Historic Infrastructure-Centered Expansion Act .1740: Flexible Grid Infrastructure Act of 2019 S.1742: Distributed Energy Demonstration Act of 2019
6	H.R.4863: United States Export Finance Agency Act of 2019 H.R.704: Fair Trade with China Enforcement Act

4.2 Probabilistic Latent Semantic Analysis

PLSA was conducted retaining five and six latent topics. Retaining six topics resulted in more distinctive and representative due to a category covering energy industry development which was added. Five out of 6 topics are clearly related to energy: 2) Energy program, 3) Energy industry development program, 4) General electric vehicle development + energy storage, 5) US-China electric vehicle + energy storage, and 6) vehicle emissions. The probabilities associated with the six topics range from 36 percent for the first topic, over 19, 14, 11, 9 down to 1 percent for the sixth topic. Table 3 shows the 15 most frequent terms per latent topic.

Table 3: Topics and most frequent terms for PLSA with six latent topics

Topic	The 15 most frequent terms
1) Land and water management	**land** shall secretari nation **manag** public state law act area feder usc date unit **river**
2) Energy program	**energy** shall secretari jun **program** state sep verdat project sec ehih dcprod pbinn **assist provid**
3) Energy industry development program	**grant** hous **program** act public verdat sep elig **energi** shall entity secretari **development** provid **industri**
4) Electric vehicle development + energy storage	**Energi electr technolog** sep **vehicl** verdat includ **develop** secretari shall **storag** pamtmann act use bfkhbprod
5) US-China energy	**state unit** shall **bank china** agenc sep verdat **energi** act **people' republic** may project sec
6) vehicle emission	**vehicle year** shall erdat sep state **emiss** product ehih **code** may act **zeroemiss credit gas**

The topics of Table 3 compared with the document titles of Table 4 represents the close link between latent topics and titles.

Table 4: Most prevalent bills per topic resulting from PLSA with six latent topics

Topic	Title
1	S.47: John D. Dingell, Jr. Conservation, Management, and Recreation Act
2	H.R.2741: Leading Infrastructure for Tomorrow's America Act S.1740: Flexible Grid Infrastructure Act of 2019 S.1742: Distributed Energy Demonstration Act of 2019
3	H.R.5185: Green New Deal for Public Housing Act H.R.4061: Blue Collar and Green Collar Jobs Development Act of 2019
4	S.1740: Flexible Grid Infrastructure Act of 2019 S.1085: Vehicle Innovation Act of 2019
5	H.R.4863: United States Export Finance Agency Act of 2019 H.R.3423: H.R.3423: National Green Bank Act of 2019 H.R. Fair Trade with China Enforcement Act
6	H.R.4520: Modernizing America with Rebuilding to Kickstart the Economy of the Twenty-first Century with a Historic Infrastructure-Centered Expansion Act H.R.2764: Zero-Emission Vehicles Act of 2019

The topics uncovered by PLSA are partly the same as in NMF, although in a different order and thus differently prevalent. The main differences are that NMF includes a topic related to energy finance, whereas PLSA includes a topic related to vehicle emissions, which is part of the industry topic in NMF. Moreover, the topics related to the US-China relationship is concentrated on energy for PLSA, and on trade exports and emissions for NMF.

Since NMF and PLSA bases on an iterative process of minimising the Frobenius norm and applying the expectation-maximisation (EM) algorithm, respectively, one need pay attention to always set the same starting value. A changed seed changes the most frequently used values per topic and may even change give the topics different meanings. Moreover, the variation of the document size influences the analysis. Especially without normalising the data, it is very likely that a comparably long document is captured by a distinctive topic (cf. Table 2). These caveats need to be considered when expanding the analysis to a larger corpus of texts.

5. Conclusion

The present text analysis aimed at uncovering the latent topics underlying the US bills of the 116th Congress which deal with batteries in the context of energy storage and electric vehicles. The analysis is part of a larger data collection project investigating US battery storage bills over the last 20 years.

Non-negative Matrix Factorisation (NMF), applied on the original data and on tf-idf normalised data, and Probabilistic Latent Semantic Analysis (PLSA) have been applied. For all three analyses, retaining six latent topics results in the best representation of distinctive topics. NMF uncovers the topics 1) land and water management, 2) energy program, 3) energy finance, 4) electric vehicles grants, 5) energy industry and emissions, and 6) US-china zero emission. The normalised variant is more specific and appears to uncover different policy instruments. PLSA returned similar topics like NMF, however, displayed a topic related to vehicle emissions instead of energy finance. Moreover, the topics are ordered differently and are thus differently prevalent.

Further analyses could apply the same methods on different subsets of the corpus of battery storage bills and compare the shift in topics over the last 20 years. PLSA has certain advantages over NMF, such as the automatic normalisation and the easy interpretability in probabilities. Nonetheless, if the entire corpus covering the last 20 years should be analysed at once, PLSA is computationally too expensive and NMF should be chosen. Concludingly, since the pre-processing has shown that the three methods present very similar results, the aim of the further research determines the choice of method.

Bibliography

Aggarwal, Charu C. 2018a. “Chapter 1: Machine Learning for Text: An Introduction.” In *Machine Learning for Text*, 1–16. New York: Springer.

———. 2018b. “Machine Learning for Text Chapter 2: Text Preparation and Similarity Computation.” In *Machine Learning for Text*, 17–30. New York: Springer.

———. 2018c. “Machine Learning for Text Chapter 3: Matrix Factorization and Topic Modeling.” In *Machine Learning for Text*, 31–71. New York: Springer.

Beuse, Martin, Tobias S. Schmidt, and Vanessa Wood. 2018. “A ‘Technology-Smart’ Battery Policy Strategy for Europe.” *Science* 361 (6407): 1075–77. https://doi.org/10.1126/science.aau2516.

“Congress Library” . 2019. https://www.congress.gov/quick-search/legislation?wordsPhrases=&include=on&wordVariants=on&congresses%5B%5D=all&legislationNumbers=&legislativeAction=106&sponsor=on&cosponsor=on&representative=&senator=&houseCommittee%5B%5D=hssy00&searchResultViewType=com.

Feinerer, Ingo. 2019. “StripWhitespace.” 2019. https://www.rdocumentation.org/packages/tm/versions/0.7-7/topics/stripWhitespace.

Gandhi, Arun. 2018. “Topic Modeling with LSA, PLSA, LDA & Lda2Vec.” 2018. https://nanonets.com/blog/topic-modeling-with-lsa-plsa-lda-lda2vec/.

Lee, Daniel D., and Sebastian H. Seung. 1999. “Learning the Parts of Objects by Non-Negative Matrix Factorization.” *Nature* 401: 788–91. www.nature.com.

Lee, Daniel D, and H Sebastian Seung. 2001. “Algorithms for Non-Negative Matrix Factorization.” *Advances in Neural Information Processing Systems* 13: 556–562.

Porter, Martin. 2019. “The Porter Stemming Algorithm.” 2019. https://tartarus.org/martin/PorterStemmer/.

Steenbergen, Marco R. 2018. “Slides: PLSA AND LDAt.”