# An Analysis of Topic Modelling for Legislative Texts

James O' Neill
Insight Centre for Data Analytics
IDA Business Park
Galway, Ireland
james.oneill@insight-centre.org

Cécile Robin
Insight Centre for Data Analytics
IDA Business Park
Galway, Ireland
cecile.robin@insight-centre.org

Leona O' Brien
Governance, Risk and Compliance Technology Centre
University College Cork
Cork, Ireland
leona.obrien@ucc.ie

Paul Buitelaar
Insight Centre for Data Analytics
IDA Business Park
Galway, Ireland
paul.buitelaar@insight-centre.org

## ABSTRACT

The uprise of legislative documents within the past decade has risen dramatically, making it difficult for law practitioners to attend to legislation such as Statutory Instrument orders and Acts. This work focuses on the use of topic models for summarizing and visualizing British legislation, with a view toward easier browsing and identification of salient legal topics and their respective set of topic specific terms. We provide an initial qualitative evaluation from a legal expert on how the models have performed, by ranking them for each jurisdiction according to topic coherency and relevance.

## 1 INTRODUCTION

The legal domain is experiencing a major shift towards automated tools that can perform tasks that are becoming increasingly difficult for legal practitioners to carry out, due to the rate of change in the legal domain. Regulatory change (RC) is a notable area that has gained more attention in recent years due to the difficulties in compliance. In order to build automated solutions for compliance and verification, automated knowledge acquisition is an imperative for related tasks. An initial step towards such a system requires an overview/summarization of the core topics within the domain, in order to identify salient terms within the topics that are potentially associated with compliance across various documents. Many approaches in legal systems require metadata from an XML schema to carry out analysis such as topic modelling. This paper analyzes the use of topic models to do this automatically from raw text. We start with a background to the models used for testing.

## 2 TOPIC MODELLING

### 2.1 Dimensionality Reduction Approaches

A basic approach to modelling topics is to view a corpus as a set of term frequencies (tf) where the weight for each term is also dependent on the inverse document frequency (idf) ( e.g "and" occurs many times in a document, therefore its weight is low). Formally, $f_{t,d} * \log \frac{N}{n_t}$ where $N$ represents the number of documents and $n_t$ is the number of documents term $t$ appears in. From a term-document matrix $M$, dimensionality reduction techniques are often used to reduce all terms to a set of concepts, which can be interpreted as approximations of "topics" in a given corpus. The matrix factorization techniques we discuss include Singular Value Decomposition (SVD) and Non-Negative Matrix Factorization (NMF).

*2.1.1 Non-Negative Matrix Factorization.* NMF is specifically for factorizing matrices with non-negative values, hence why it is particularly suitable for term-document matrices. Since $M$ is represented as non-negative values, features are composed of additive computations resulting in a part based representation (as opposed to subtracting values which would not lead to parts-based factored representation) [6]. The objective of NMF is to find an approximation of matrix $M$ by factorizing it into $W(r\ x\ k)$ and $H(k\ x\ c)$ such that $M \approx WH$ and $k$ have lower rankthan $M$. The reconstruction error is minimized according to that shown in Equation 1 [11, 12].

$$\frac{1}{2}|M - WH|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(M_{ij} - WH_{ij})^2 \tag{1}$$

Also described by Lee and Seung [11], the multiplicative update algorithm is used for updating both $W$ and $H$. Both update rules are outlined in Equation 2. The objective ensures the minimization is constrained to $W$ and $H$ being positive and that the distance $D$ between both is positive.

$$\begin{aligned} H_{\alpha,\mu} &:= H_{\alpha,\mu}\frac{(W^T M)_{\alpha,\mu}}{(W^T WH)_{\alpha,\mu}}, \\ W_{i,\alpha} &:= W_{i,\alpha}\frac{(MH^T)_{\alpha,i}}{(WHH^T)_{\alpha,i}} \end{aligned} \tag{2}$$

In this work, instead of using gradient descent to minimize the sum of squared (euclidean) distance (SSD) between $M$ and $WH$, we use the Coordinate Descent solver. Lin et al. [13] describe the process that builds upon the multiplicative update algorithm by applying Alternating Non-negative Least Squares (ANLS) using projected gradient descent which is a parameter estimator with lower-bounded constraints. Although, NMF is widely used for topic modelFling [21], it is sometimes known to produce non-meaningful topics, particularly if a term-document matrix is relatively sparse. Therefore, the identification of both rare and non-distinct terms is an important step

In: Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017), June 16, 2017, London, UK.

Published at http://ceur-ws.org

to consider for removal before factorization. Furthermore, NMF can be prone to local minima.

*2.1.2 Singular Value Decomposition.* SVD decomposes a matrix into three parts as shown in Equation (3) in order to find a lower rank[1] approximation of the term-document matrix. Consider $M$ to be a tf-idf matrix representation of the corpus, where $U$ diagonalizes $MM^T$ and $u_i$ represents the corresponding eigenvector. Similarly $V^*$ diagonalizes $M^TM$ and $v_i$ represents $M^TM$ eigenvectors. The diagonal values of $\Sigma$ are ordered singular values[2].

$$M = U\Sigma V^* \tag{3}$$

SVD on a term-document matrix is also referred to as Latent Semantic Analysis (LSA), as the lower ranked matrix $M$ is said to represent a latent semantic space. In information retrieval, it is referred to as Latent Semantic Indexing (LSI), where SVD is used to index documents by representing documents (document-document) and terms (document-term where terms are query terms) in vector space where the elements in the vector correspond to the degree that a term or document has to a given topic. The similarity between a query and a given set of documents can then be determined using a term-topic-matrix [18]. This is particularly helpful for distinguishing polysemous and synonymous terms.

## 2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation was first introduced by Blei et al. [2] and has since been a state of the art (SoTA) topic model, showing to have more expressiveness over probabilistic LSA (pLSA) [3]. LDA builds a Bayesian generative model using Dirichlet priors for topic mixtures (an assumed prior probability for each topic distribution, Dirichlet is a set of categorical distributions in this sense), in contrast to pLSA that can be considered to use uniform prior distribution for the topic mixtures. Further extensions since then have been made to improve and adapt this model in a continuous space setting. In this sense, continuous word embeddings are used. Categorical distributions are replaced with multivariate Gaussian distributions, meaning that Gaussian LDA has the capability of handling out of vocabulary words on unseen text [8]. The probability of word $w$ is dependent on a topic $k$ in $z$ which is dependent on probability of a document $\theta_d$ that is drawn from a Dirichlet prior $\alpha$. Likewise, a word $w$ is also dependent on the probability $\phi$ that a word $w$ is in topic $k$.

The LDA generative process is described by Blei [3]. For each document, a parameter $\theta_d$ is chosen from a Dirichlet prior distribution, then for each word in $d$ a topic category is chosen according to the Dirichlet. A word $w$ is generated afterwards, given the topic $z_w$ and $\beta$.

The aforementioned Gaussian LDA represents these words as continuous embedded vectors instead of discrete co-occurrence counts, replacing the categorical distributions for $z_n$ and $w_n$ with Gaussians.

The saliency of terms within a topic is considered by [7] and formulated in Equation 4. A distinctive word $w$ is a word that has a higher log-likelihood of being in a topic $K$ compared to a random word. Hence, if a word $w$ occurs in many topics it is non-informative, resulting in lower saliency. More informative topic-specific terms and less general terms are desired by the legal practitioners, hence we use this saliency measure in our analysis.

$$S(w) = P(w) \sum_T P(K|w) \log \frac{P(K|w)}{P(K)} \tag{4}$$

Sievert and Shirley [19] describe the relevance measure, also shown in 5, where $\phi_{k(w)}$ is the probability of $w$ for topic $k$ and $p(w)$ is the probability of observing $w$ in corpus $D$. In this work, $\lambda$ is can be chosen between 0-1. We set $\lambda$ according to term relevance judgments made by a legal practitioner, prior to the final analysis of each topic model.

$$r(w, k|\lambda) = \lambda \log(\phi_{k(w)}) + (1 - \lambda) \log(\frac{\phi_{k(w)}}{p(w)}) \tag{5}$$

## 2.3 Saffron

*Saffron* is a software tool[3] that can construct a model-free topic hierarchy. It extracts terms related to the domain of expertise, establish semantic relations between them, and constructs a taxonomy out of it. *Saffron* also deals with multiword expressions, which can improve topic coherency as phrases are often necessary for better readability and understanding.

*Saffron* builds the topic hierarchy of a corpus by first capturing the expertise domain through a model represented as single-word list. The latter is extracted using feature selection during a term and linguistic pattern extraction phase. It uses constraints such as limiting to contentful parts-of-speech, to single words (in order to target a more generic level) and to terms distributed across at least a 1/4 of the corpus (for the specificity to the area of expertise). Topic coherency, which is a main issue for statistically driven models in order for Subject Matter Experts (SMEs) to reply upon them, is tackled here by using semantic relatedness to filter the candidate words. It is interpreted here as a domain coherency measure using Pointwise Mutual Information (PMI) (see [4] for more details). The domain model is then used as a base to measure the coherence of the topics within the domain in the next phase.

After extracting candidate terms following a standard multi-word term extraction technique (see [4]), the first step involves searching for words from the domain model in the immediate context of those candidates. This allows to determine a term's coherence within the domain. This is achieved again through PMI calculation, by using top level terms to extract intermediate level terms.

To create the pruned graph which represents the taxonomy, the strength of relationship between two research terms is measured, defined as $I_{ij} = D_{ij}/(D_i \times D_j)$ where $D_i$ is the number of articles that mention the term Ti in our corpus, $D_j$ is number of articles that mention the term $T_j$, and $D_{ij}$ is the number of documents where both terms appear. Edges are added in the graph for all the pairs that appear together in at least three documents, threshold fixed based on the results of previous studies and tests (see [4] for more details). *Saffron* also uses a generality measure to direct edges from generic concepts to more specific ones. This results in a dense, noisy directed graph that is further trimmed using a specific branching algorithm which was successfully applied for the construction of domain taxonomies in [14]. This yields a tree structure where the

[1] The rank of a matrix is the number of linearly independent column vectors in a matrix (e.g document-term matrix), which can be used to reconstruct all column vectors.
[2] singular values are the square root of the eigenvalue
[3] see here - http://saffron.insight-centre.org/

root is the most generic term and the leaves are the most specific terms.

## 3 RELATED WORK

Wiltshire et al. [20] introduced a large scale machine learning systems that incorporates the use of hierarchical topic construction after the extraction of terms, legal phrases and case cites. Their system allows for a ranking and classification of topics given a legal concept as input according to a scoring criterion. George et al. [10] provide a legal system for ranking documents according to their similarity to legal cases by finding similarity between documents in the latent topic space and query terms. They then use human assistance to provide annotate documents that are relevant to the query in a semi-supervised fashion. In contrast, our work is fully unsupervised with no human assistance during the topic modelling process. LDA has been used extensively on natural language texts such as social media texts [16], publication texts, newspapers etc. and typically not in formal settings such as their use on legal texts.

Raghuveer and Kumar [17] use LDA to cluster Indian legal judgements and use cosine similarity as the distance measure between documents for clustering. However, their evaluation does not present the prior knowledge of a legal expert to determine if the clusters coincide with legal knowledge within the domain.

O' Neill et al. [15] have identified salient legal statements (in contrast to salient topics) by extracting deontic modalities from using a small number of labeled samples to train a recurrent neural network.

Ahmed and Xing [1] use dynamic HDP to track topic over time, documents can be exchanged however the ordering is intact. They also use longitudinal NIPS papers to track emerging topics and decaying topics (this is worth noting, particularly for tracking changing topics around compliance issues).

The use of the aforementioned *Saffron* has been previously demonstrated through a wide range of projects from several domains and for different tasks. In [5], Bordea used *Saffron*'s topic extractor to analyze legal documents arising around the financial crisis in 2008. She mapped the problem as an expert finding task, which aims at ranking people that have knowledge about a given topic. In that particular context, the task allowed the identification of individuals involved in defining the response of the U.S. government to the financial crisis by searching for a topic of interest. In [4], Saffron was used as a tool to detect the presence of different disciplines within the field of Web Science. By running it on over 10 years of Web Science conference series documents, it resulted on a discovery of 4 communities (Communication, Computer Science, Psychology, and Sociology), and trends over time and types of paper. *Saffron* was also used in a demo for an Irish bookshop website[4] to extract topics from book descriptions/reviews and then classify them accordingly. It was also used to link the books for the creation of a multi-level browsing application for book navigation.

## 4 METHODOLOGY

This section outlines the steps towards creating each topic model and their configurations used for analysis. We start with a brief introduction to the corpora used and preprocessing steps common to all topic models. United Kingdom legislative texts were used for topic modelling[5]. The corpus contains 41,518 documents between 2000 - 2016. However, for practical purposes the analysis is carried out on the year 2016, only to lessen the reading burden on the legal practitioner. The legislative types consist of the following: 304 *Northern Ireland Statutory Rules*, 838 *UK Statutory Instruments*, 132 *Welsh Statutory Instruments* and 317 *Scottish Statutory Instruments*.

### 4.1 Text Preprocessing

Corpus specific regular expressions (RE) are used to clean legal domain syntax (e.g bracketed alphanumerics), followed by tokenization and lemmatization using the WordNet lemmatizer [9]. The structure usually contains nested expressions e.g (ii) followed by (a) and (b) subsections. This syntax is removed using the regular expressions along with other standard RE for identifying references and alphanumeric expressions e.g "*Regulation EC No. 1370/2007 means Regulation 1370/2007 ...*". Redundant stopwords are removed from the corpora for word frequency $f < 2$. This is carried out under the supervision of a subject expert by analysing a subsample of terms which are considered for removal. We assume that terms with high frequency are not specific to a particular topic e.g 'the','of' etc. Also, rare terms that occur infrequently are not representative of a single topic since they do not appear enough to infer that it is salient for a topic. Each corpus (corpus per jurisdiction) is then converted to a term-document matrix where weights are placed on each word using the aforementioned tf-idf weighting scheme. Furthermore, 30 terms for all models except *Saffron* are listed for SME for ranking. For *Saffron* we rely on a visualization of the term hierarchy for a domain expert to judge.

### 4.2 Ranking Criterion and Model Configurations

In order for a legal practitioner to assess the models in a fair manner, a set of guidelines are presented for the ranking of the models. An important aspect to ranking is the pretuning of the term relevance parameter $\lambda$, which chooses the top 30 terms that are presented for each topic within the jurisdiction accordingly. We also assess a number of parameter setting for NMF, LSA, LDA and HDP before finally choosing the final 10 set of topics which the legal expert makes their final judgment. Since the term-document matrix is quite sparse (evident from 1), NMF is initialized using Non-Negative Singular Value Decomposition (NNSVD). The Coordinate Descent solver is used for minimizing the reconstruction error as mentioned in section 2.1.1. The number of components is set to $n_k = 10$. LSI uses standard SVD which does not require much tuning only to choose the number of singular values, also set $n_k = 10$. For LDA we choose low relevance $\lambda = 0.25$ to highlight topic specific terms.

## 5 RESULTS

In this section we analyse the topics retrieved for each approach, and an SME evaluated the topics for the regulations. Figure 1 simply compares the effects of dictionary size once infrequent terms are increasingly removed. It is evident that after removing terms that occur less than twice, the corpus' size dramatically decreases, meaning that a significant number of terms are too specific to a particular document. We remove these terms for subsequent analysis.

[4] see http://kennys.insight-centre.org/

[5] Retrieved from: http://www.legislation.gov.uk/

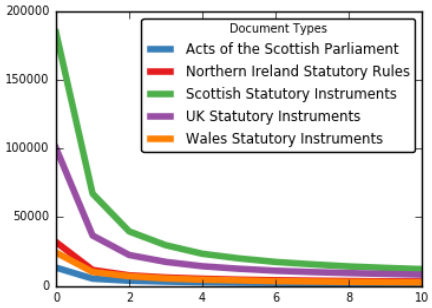


**Figure 1: Rare-word Removal For Each Corpus**

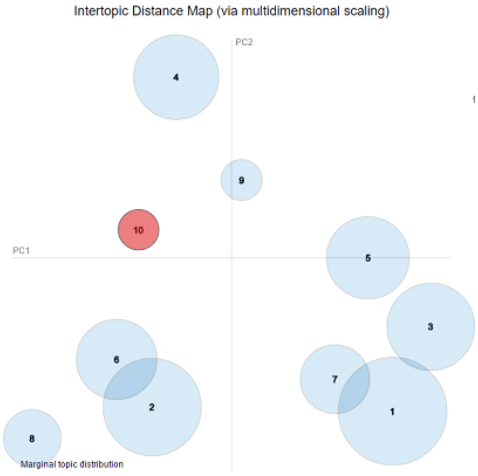


**Figure 2: LDA topics for Northern Ireland Statutory Rules projected to 2 principal components using multi-dimensional scaling (MDS)**

***Latent Dirichlet Allocation Visualization***. For the visualization of LDA topics, we use the *pyLDAvis* [19] visualization tool. A multidimensional scaling projects the $t$ dimensional space to a 2 dimensions as shown in figure 2. Ten topics for Northern Ireland Statutory Rules (NISR) are presented with the relevance metric set $\lambda = 0.25$ (which decides the term-topic specificity). This is done under the supervision of a legal practitioner to ensure that $\lambda$ is tuned to a correct specificity and that topics are also coherent, before a final evaluation.

Some terms such as *biomass*, *biomaterial*, *bioliquid*, *fossil* and *fuel* show a clear and distinct topic and are quite topic specific given $\lambda = 0.25$, shown by red bars which indicate the term frequency with the given topic as opposed to the blue bar that indicate the term frequency among the whole corpus.

***Saffron***. In *Saffron*'s results, a cluster is located around the extracted topic of *department of justice*, and *support allowance* which derives the whole taxonomy for the Northern Ireland Statutory Rules. This topic is thus the primary node of the 2016 corpus. In Figure 4, we zoom in a subset of this graph (and thus subdomains) which includes *housing benefit*, *income support*, *social security*, *personal_independence payment*. They all are semantically related to the mother node *support allowance*, but tackling

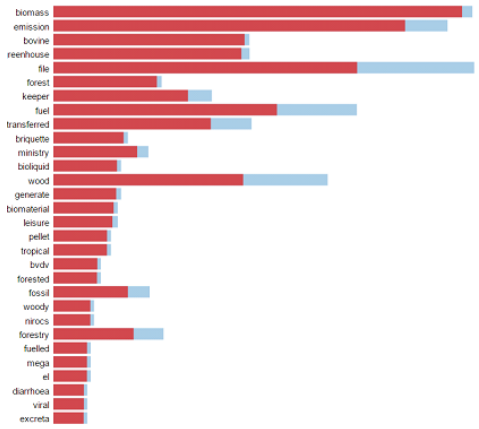


**Figure 3: Latent Dirichlet Allocation terms for topic 10 of Northern Ireland Statutory Rules**

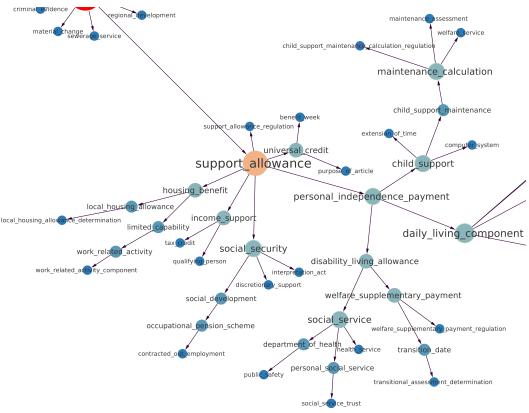


**Figure 4: Support Allowance topic within Northern Ireland Statutory Rules**

different aspects of it. We can see the advantage of the hierarchical structure of the graph, with semantically related topics going from the more generic to the more specialized ones. We can this way identify a waterfall structure from the *housing benefit* branch, logically followed by the more specific *local housing allowance*, and then *local housing allowance determination*. Another quite clear example can be observed from the *child support* branch, related to the *personal independence_payment* node. From *child support*, the directed edge links to *child support maintenance*, then *maintenance calculation*, and finally the three topics *child_support_maintenance_calculation_regulation*, *welfare service* and *maintenance assessment*. The *police service* node is at the root of a taxonomy that includes children nodes *northern_ireland_reserve* ⇒ *notice_of_appeal* ⇒ *written_representation,avoiding service* ⇒ *reasonable_amount_of_duty_time*. This example summary allows a legal practitioner to identify topics surrounding certain legal issues, or for simply summarizing a complete jurisdiction. Zooming in on a subset of the hierarchical tree, we highlight a topic with coherent

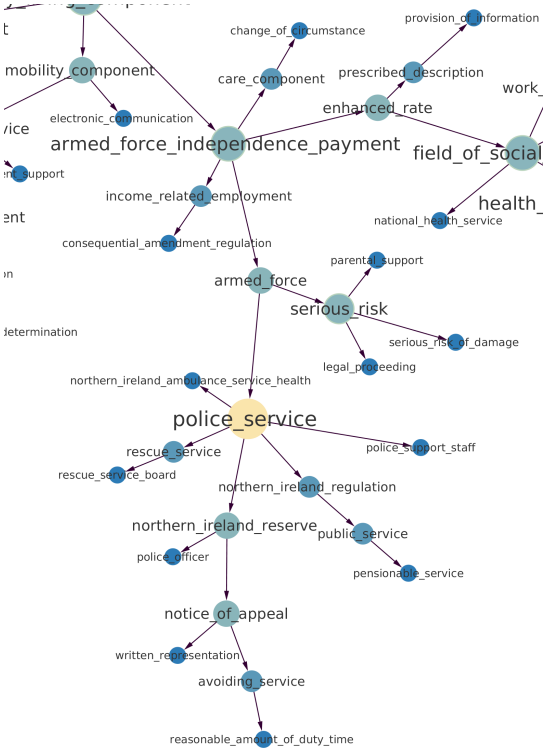


**Figure 5: Police Service topic within Northern Ireland Statutory Rules**

| Rank | NISR | SSI | UKSI | WSI |
|---|---|---|---|---|
| 1 | Saffron | Saffron | Saffron | Saffron |
| 2 | LDA | LDA | LDA | LDA |
| 3 | HDP | NMF | HLDP/LSI | HLDP/LSI |
| 4 | LSA | LSI | HLDP/LSI | HLDP/LSI |
| 5 | NMF | HLDP | NMF | NMF |

**Table 1: Subject Matter Expert Ranking of Topic Models**

multi-word expressions summarizing an area within the Northern Ireland Statutory Rules in Figure 5.

***Ranking***. Table 1 shows the results of SME ranking after assessing each topic model for each jurisdiction. *Saffron* overall is favored for all jurisdictions, considering it is the only model that performs multi-word expression topic extraction and weighting of descriptive noun terms/phrases. We conjecture that the appeal of a hierarchical structure and multi-word noun expressions has influenced the interpretation of the salient terms in the domain, making it easier for legal practitioners to identify important and coherent legal topics.

We emphasize at this point that single word topic models and multi-word hierarchical models are not directly comparable for this reasons outlined however, they are included in table 1 to highlight the importance of longer expressions that are linked in a taxonomy, providing more clarity on what the emerging topics are.

## 6 CONCLUSION

This work has presented a fully automated approach for identifying topics in regulations that assist in easier tracking of important domain terms that correspond to compliance related issues. After evaluation *Saffron* has been consistently ranked as the most favourable of all models, as the aforementioned vocabulary pruning and usage of multi-word expressions has played a fundamental role in topic coherency. Standard LDA has performed the best of all single term models, particularly when top terms are chosen according to their topic specificity. HDP has inferred a similar number of topics as that of LDA according to an analysis of the log-likelihood curve and the legal practitioners judgment. This work is an early indication as to how legal practitioners can identify salient and coherent topics using automatic topic modelling tools.

## REFERENCES


[1] Amr Ahmed and Eric P. Xing. 2012. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. *CoRR* abs/1203.3463 (2012). http://arxiv.org/abs/1203.3463

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[4] Georgeta Bordea. 2013. *Domain adaptive extraction of topical hierarchies for Expertise Mining*. Ph.D. Dissertation.

[5] Georgeta Bordea, Kartik Asooja, Paul Buitelaar, and Leona OâĂŹBrien. 2014. Gaining insights into the Global Financial Crisis using Saffron. *NLP Unshared Task in PoliInformatics* (2014).

[6] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. 2008. Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 63–72.

[7] Jason Chuang, Christopher D Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 74–77.

[8] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for Topic Models with Word Embeddings.. In *ACL (1)*. 795–804.

[9] Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

[10] Clint Pazhayidam George, Sahil Puri, Daisy Zhe Wang, Joseph N Wilson, and William F Hamilton. 2014. SMART Electronic Legal Discovery Via Topic Modeling.. In *FLAIRS Conference*.

[11] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[12] Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*. 556–562.

[13] Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19, 10 (2007), 2756–2779.

[14] Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three (IJCAI'11)*. AAAI Press, 1872–1877. DOI:http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-313

[15] James O' Neill, Paul Buitelaar, Cecile Robin, and Leona O' Brien. 2017. Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM, 159–168.

[16] Marco Pennacchiotti and Siva Gurumurthy. 2011. Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 101–102.

[17] K Raghuveer. 2012. Legal documents clustering using latent dirichlet allocation. *IAES Int. J. Artif. Intell.* 2, 1 (2012), 34–37.

[18] Barbara Rosario. 2000. Latent semantic indexing: An overview. *Techn. rep. INFOSYS* 240 (2000).

[19] Carson Sievert and Kenneth E Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 63–70.

[20] James S Wiltshire Jr, John T Morelock, Timothy L Humphrey, X Allan Lu, James M Peck, and Salahuddin Ahmed. 2002. System and method for classifying legal concepts using legal topic scheme. (Dec. 31 2002). US Patent 6,502,081.

[21] Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. 2013. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 749–757.