UNIVERSITY COLLEGE LONDON

MASTERS THESIS

Dynamic Topic Modeling of PATSTAT Patents Using LDA

Author:
Christopher MARTIN

Supervisors:
Dr. John Shawe TAYLOR
& Christopher GRAINGER

This thesis is submitted in fulfillment of the requirements for the degree of Masters of Science

Machine Learning
UCL Dept. of Computer Science

It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

September 4, 2016

Declaration of Authorship

I, Christopher MARTIN, declare that this thesis titled, “Dynamic Topic Modeling of PATSTAT Patents Using LDA” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

UNIVERSITY COLLEGE LONDON

Abstract

Dr. John Shawe Taylor
UCL Dept. of Computer Science

Masters of Science

Dynamic Topic Modeling of PATSTAT Patents Using LDA

by Christopher MARTIN

In this paper we evaluate the performance of a time varying family of LDA based topic models, namely the Dynamic Topic Model (DTM), and Dynamic Influence Model (DIM). These models are meant to capture both the underlying semantic structure of a document collection and the evolution of that structure in time. Consequently, the DTM and DIM have proven useful for illustrating changes in the use of language regarding specialized subject matter. We compare these dynamic models to traditional topic models such as LDA as a benchmark, and explore their efficacy not only at creating semantically coherent topics, but at tasks such as document classification, and clustering. Finally, we present results on over 18 years of patent data from the PATSTAT database across 5 classes of patents demonstrating interpretable trends, improved document classification and clustering, as well as enhanced topic coherence.

Acknowledgements

I owe the completion of this project to the many people who have helped along the way, either directly or indirectly. To my project supervisors Christopher Grainger and Prof. John Shawe-Taylor for their continued guidance throughout the project, to my colleauges for their advice and discussion, and to my family for their moral support, I express my sincere gratitude, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Researchers today are faced with a deluge of data. As we continue to digitize and aggregate our collective knowledge we produce ever increasing archives of information. The sheer volume and variety of forms this information may take - text, images, audio, video, social connections etc. - makes it difficult and in most cases impossible to parse manually.

This driving factor of data growth has given rise to internet giants such as Google, Yahoo, and Baidu that help us access and browse pre-indexed swathes of information. However in order to go beyond mere keyword searches, or link analysis, and break into the realm of understanding each document, requires a new approach to data exploration.

A powerful set of computational tools referred to as probabilistic topic models have emerged to meet this challenge. Aimed to discover and annotate large archives of documents with thematic information, topic models identify patterns that reflect the underlying topics which combined to form those documents.

Naturally, it is rare that we would know beforehand exactly what topics a given document contains, and thus topic modeling constitutes an unsupervised task. As a result, topic modeling algorithms are designed to work without prior knowledge of the topic distribution of a given document — that is, the topics are derived from the texts themselves. This makes the organization, summarization and annotation of text corpora possible at an inhuman scale. Consequently, topic models are useful in a variety of settings and have successfully been applied to web archives, news articles (Newman et al., 2006), and academic literature (Steyvers et al., 2004) to elicit insight. In this paper, we focus our experiments on extracting topics from patent data with the hopes identifying meaningful trends in renewable energy technologies.

1.2 Primer on Latent Dirichilet allocation

Fortunately, the intuition behind LDA topic models is relatively straight forward. To understand how the algorithm infers the topics in relation to documents we first define what constitutes a topic. A topic is a distribution over a fixed vocabulary, where each word has an assigned probability of occurrence. Subsequently, we can take the view that each document is likely a product of one or more topics, a cocktail of themes as it were with different proportions of each ingredient.

Take for example the following document sampled from the August 2014 EPO Worldwide Patent Statistical Database (**PATSTAT**). The patent abstract contained in Figure 1.1 relates to a mechanism for stopping a water wheel. We have taken the liberty of highlighting a selection of words from a few of this document's prominent topics. Words like "pressure", "liquid", and "flow" belong to the **fluids/water** topic and are colored blue. While words relating to the **mechanisms** by which this fluid is directed such as "chamber", "valve", and "guide" are colored red. Finally, words such as "transmission", "speed", and "operated" belong to the topic associated with **signals** and are colored green.

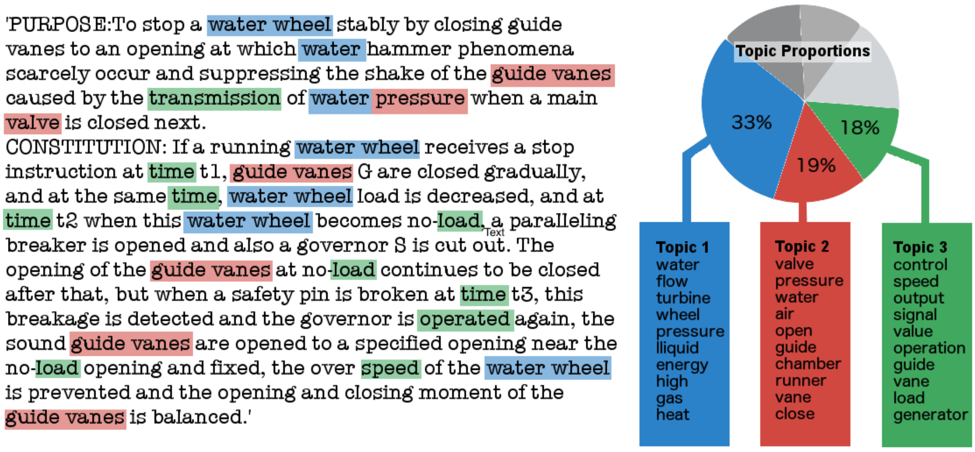

'PURPOSE:To stop a water wheel stably by closing guide vanes to an opening at which water hammer phenomena scarcely occur and suppressing the shake of the guide vanes caused by the transmission of water pressure when a main valve is closed next.
CONSTITUTION: If a running water wheel receives a stop instruction at time t1, guide vanes G are closed gradually, and at the same time, water wheel load is decreased, and at time t2 when this water wheel becomes no-load, a paralleling breaker is opened and also a governor S is cut out. The opening of the guide vanes at no-load continues to be closed after that, but when a safety pin is broken at time t3, this breakage is detected and the governor is operated again, the sound guide vanes are opened to a specified opening near the no-load opening and fixed, the over speed of the water wheel is prevented and the opening and closing moment of the guide vanes is balanced.'

FIGURE 1.1: Topic proportions in a sample patent abstract.

The topics in the previous example were formed not over a single document but over a collection. The grey sections of the pie chart above represent the topics that this patent does not contain strong elements of. This is a key characteristic of LDA topic models, each document has a unique topic 'fingerprint' as a result of a generative process. That process for generating a document word by word is as follows. First we decide, sampling from the distribution of topics, which topic our first word will belong to. Then we sample from that chosen topic's distribution to decide what the word itself will be. This process is then simply repeated for each word, and while it works it has the following assumptions worth noting:

- Documents can manifest multiple topics (however typically not many)
- Each document is assumed to be the product of a generative process.

- Generative process starts with a topic, i.e. a distribution over a fixed vocabulary.
- Assumes a fixed number of topics

Latent Dirichilet allocation falls into a family of machine learning algorithms called **hidden variable models**. In this family of models, the user customarily "posits a hidden structure in the observed data, and then learns that structure using posterior probabilistic inference" (Blei and Lafferty, 2009). For LDA specifically, the documents are the observed data, the topics and document topic proportions are hidden.

More formally, we may define this process mathematically as a joint distribution over our hidden variables and our observed variables. Specifically, we define the distribution over vocabulary as β, the topic proportions for document d θ_d, the topic assignment for a word in a document $z_{d,n}$ and of course the observed words themselves $w_{d,n}$.

$$\begin{aligned} &P(\beta_{1:K}, \theta_{1:D}, z_{1:D}, W_{1:D}) \\ &= \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \{ \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, z_{d,n}) \} \end{aligned} \tag{1.1}$$

In Eq 1.1 we see a few dependencies worth noting. Firstly that the topic we assign to a word $z_{d,n}$ depends on the distribution of topics of its document θ_d. Additionally, that the identity of the word itself is dependent on not only the topic we assigned to generate it $z_{d,n}$, but also the vocabulary distributions of each topic $\beta_{1:K}$. Equivalently we can express the dependencies between these variables as a graphical model, illustrated in figure 1.2.

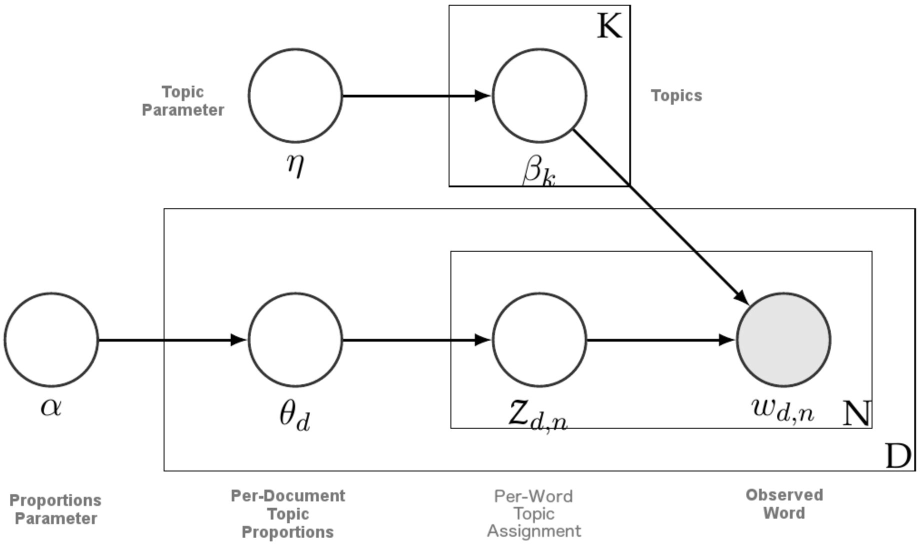


FIGURE 1.2: Graphical model for LDA

So how do we actually obtain our estimates of the hidden parameters? We need to calculate the conditional distribution of our hidden parameters (the topic structure), and the observed words i.e. the posterior distribution described in Eq. 1.2. However the denominator makes this calculation computationally infeasible due to the number of combinations our hidden parameters could take.

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|W_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|W_{1:D})}{p(W_{1:D})} \tag{1.2}$$

To move past this, most solutions use either sampling or variational based methods to perform approximate inference and obtain estimates of the hidden parameters. Variational methods allow us to translate the original problem to one of optimization and take advantage of the many optimization techniques available. This in turn allows us to make extensions that are often faster, scale better or allow for different forms of input such as streaming documents.

1.3 Adding a temporal component

One such extension, and the extension we explore in this study, is to relax the implicit assumption of LDA that the order of the documents doesn't matter. By incorporating the order of the documents to the model, a topic is no longer simply a distribution over words but now becomes a *sequence* of distributions over words. This is the jump that allows us not only to identify a theme, as with static LDA, but also track how it progresses in time, giving us the Dynamic Topic Model (**DTM**).

The DTM offers several advantages over traditional LDA including improved predictive performance (Blei and Lafferty, 2006). Primarily though, it facilitates a greater understanding of how each topic developed, and how the ideas therein formed and matured. With it, we can inspect trends of word usage to uncover a richer and more detailed hidden structure. For instance figure 1.3 contains a sample theme from a sub-collection of hydroelectric patents and the progression of word prevalences within it over time.

1.4 Why patents?

Patent data is specifically interesting in this context because of the role patents play in company formation, job growth, economic development, and novel invention. Their history tells a story of technological progression. In an attempt to maintain a competitive edge, many companies large and small spend a considerable amount of energy researching this history to identify technical trends relevant to their industry.

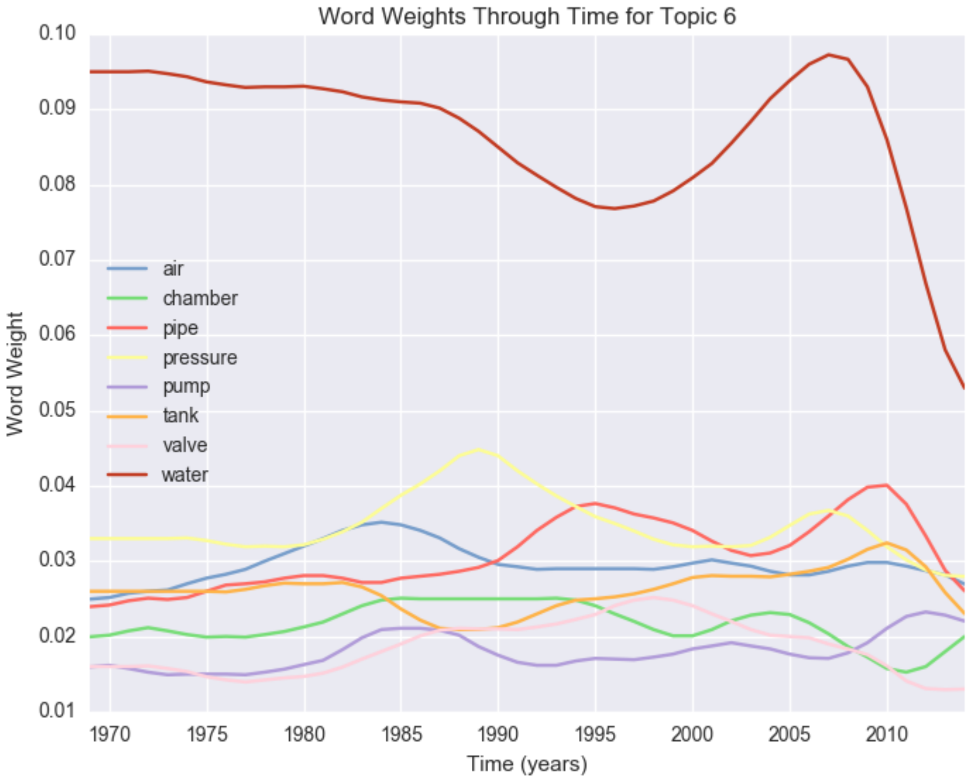


FIGURE 1.3: Distribution over words in a sample hydro-electric topic over time

Dynamic topic models have the potential to aid this research by enabling us to track the evolution of innovation through language use in patent abstracts. In this paper we look at a number of elements, including the evolution of technological themes and their proportions, the origination and development of language, as well as document influences. Furthermore, the patent corpus and associated International Patent Classification (**IPC**) labels provide a platform for the comparison of various topic modeling algorithms.

1.5 Overview of Experiments

At the time of writing this, surprisingly little has been published exploring the effectiveness of both the DTM and the DIM. Much research has evaluated model quality solely based on the likelihood of held-out predictions, however likelihood does not always translate to semantically meaningful topics (Chang et al., 2009). Additionally, the predictive performance of these models is adversely affected by longer time horizons due to an "increase in the rate of specialization in scientific language" (Blei and Lafferty, 2006). Acknowledging the room to explore alternative methods of model evaluation, we implemented the experiments listed below.

Historical Topic Trend Validation

The simplest, but also the most hands-on, method of evaluating the quality of topics produced by the DTM and DIM is simply to validate the inferred topic trends against known industry history. For instance, if in the topic of water purification systems we observe a rise in the usage of words "2D materials" and "lattice membranes" around 2005, we might substantiate this by pointing out graphene's isolation the previous year in 2004.

1.5.1 Topic Coherence

In light of research suggesting that likelihoods and perplexity don't always correlate with human judgement on the interpretibility of topics (Blei and Lafferty, 2006) we borrow several methods of topic coherence suggested by (Rosner et al., 2014). We evaluated model topic coherences using namely C_v, C_npmi,C_uci, and U_mass. Using C_v, the metric most correlated with human judgement, DTM achieved the highest with score with .59 compared to static LDA with .50. For complete results see Table 4.1 in Chapter 4.

1.5.2 Classification

We wished to evaluate the proficiency of the word vectors unsupervisedly generated by the DTM and DIM at forming an effective feature space for document categorization. To do this we made use of the IPC labels of patents as broad class labels for text content. The resulting topic vectors should then help identify which class a document belongs to. Naturally we tested the efficacy of each model's vector space at correctly classifying the IPC label of documents when fed to a range of classification algorithms. Peak classification performance of the DTM based classifiers was F1 = .64, while LDA was yielded F1 = .58 Text classification results are given in section Table 4.3 of Chapter 4.

1.5.3 Clustering

Another method we used to evaluate the quality of the resulting document vectors was by their ability to cluster the documents. In order to determine which models yielded vector spaces of the corpus that most effectively defined separations in the data relative to the ground truth CPC labels we used the following metrics: the adjusted rand index, normalized mutual score info, homogeneity, completeness and the V-measure which we cover in section 2.1. Indeed we found that the DTM's vector space tended to outperform that of LDA at clustering with a peak NMI score of .21 compared to .17. For more detailed results, refer to Table 4.6 of chapter 4.

1.6 Thesis Outline

The overall structure of this paper is as follows. In **chapter** 2 we review the literature surrounding the applications and various evaluation methods for topic models, and also give a detailed account of both the DTM and DIM. Then in **Chapter** 3 we cover the experimental set up and considerations in data preparation. The results of our experiments are subsequently presented in **Chapter** 4, and finally we conclude in **Chapter** 5 with a discussion of results and suggestions for future study.

Chapter 2

Background Information and Theory

2.1 Literature Review

In this section we begin by reviewing a few of the tasks common in topic modeling. Then we describe a handful of the ways the quality of topic models, LDA in particular, are commonly tested. Finally we discuss the specific tasks of document classification and clustering in the context of topic modeling, as well as the corresponding methods for evaluating model performance at these tasks.

2.1.1 Applications of Topic Modeling

The most popular application of topic models is simply summarizing large text collections by mining the topics. This is a task LDA is particularly suited for (Griffiths and Steyvers, 2004; Mei, Shen, and Zhai, 2007). The original LDA paper however (Blei, Ng, and Jordan, 2003) gave promising results on document classification as well. Since then LDA has been used with success not only for document classification, but also for clustering and information retrieval (Wei and Croft, 2006; Nagwani, 2015). This is due to the strength of the topic vectors LDA models provide, which tend to correlate strongly with human judgement.

2.1.2 Ensuring Model Quality

Perplexity Testing

In order to ensure the strength of these topic vectors researchers employ a handful methods to evaluate the topic models. While the most intuitive method is simply to have humans judge the coherence of each topic, this becomes prohibitively time consuming and expensive for large data sets. One commonly used method of automating this process is by evaluating the topic model on a held out set of testing documents and obtaining the log-likelihood perplexity of the unseen documents (Blei, Ng, and Jordan, 2003; Wallach et al., 2009). A higher likelihood on unseen documents, and a lower perplexity score indicates a better model. However this method of evaluating topic model performance has several issues. Firstly, it has been shown that predictive likelihood, or equivalently perplexity, is not always correlated with human judgement, and in some cases is even slightly anti correlated (Chang et al., 2009). Secondly this method of evaluation only acts as a general measure of the entire model. What about the quality of the individual topics?

Coherence Testing

Fortunately several methods of evaluating the coherence of individual topics from topic models exist. For a topic t we define the **Umass** coherence as a sum of the pairwise scores of that topic's top words $W_t = \{w1, ...w_n\}$.

$$\begin{aligned}\text{Umass Coherence } c(t, W_t) &= \sum_{w_i, w_j \in W_t} \text{score}(w_i, w_j) \\ &= \sum_{w_i, w_j \in W_t} log \frac{d(w_i, w_j) + \epsilon}{d(w_i)} \qquad (2.1)\end{aligned}$$

Where $d(w_i)$ is the number of documents containing the word w_i and $d(w_i, w_j)$ is the number of documents containing both word w_i and w_j. The ϵ in the numerator is simply to smooth the counts and is typically set to a minimal value such as 1 or .01. Intuitively then, a topic is good if its words cooccur often (Mimno et al., 2011).

The **UCI** measure introduced by (Newman, Bonilla, and Buntine, 2011), operates in the same manner as Umass but with the pointwise mutual information as a scoring function instead, given in eq 2.2.

$$\begin{aligned}\text{UCI Coherence } = c(t, W_t) &= \sum_{w_i, w_j \in W_t} \text{score}(w_i, w_j) \\ &= \sum_{w_1, w_2 \in W_t} log \frac{p(w_i, w_j)}{p(w_i) p(w_j)} \qquad (2.2)\end{aligned}$$

Where $p(w_i)$ is the probability of seeing word w_i in a random document and $p(w_i, w_j)$ is the probability of seeing both word w_i and word w_j together in a random document. It should be noted that obtaining these probabilities requires empirically estimating them from an external dataset.

Two more noteworthy measures of topic coherence, in addition to those outlined above, were developed by Roder, Both and Hinneberg in their study titled "Exploring the Space of Topic Coherence Measures" (Röder, Both, and Hinneberg, 2015). These measures were the **C_v** and **C_npmi** measures which demonstrated a substantial correlation with human judgement. For brevity we do not replicate their derivations here, but the interested reader will find a detailed description of each in (Röder, Both, and Hinneburg, 2015)

2.1.3 Document Classification

Though the topics produced by topic models are useful in their own right for the qualitative analysis of documents, they are also useful quantitatively when trying to classify documents. For instance a large news organization may want to automatically sort its thousands of articles into the categories "politics", "natural disasters" and "sports". To do this they might use a topic model to get a vector of topic proportions for each document to use as features for a classification algorithm. This process is referred to as document vectorization.

While baseline methods for document vectorization exist, such as the Term Frequency Inverse Document Frequency (tf-idf), LDA has been shown to outperform them in certain scenarios. For instance when less training data is available LDA boasted a shorter training time and higher classification accuracy (Li and Zhang, 2010). Additionally when tested against other baseline methods for document vectorization such as the unigram model or probabilistic latent semantic analysis (PLSA), LDA again proved consistenlty more accurate at document classification tasks (Lu, Mei, and Zhai, 2011).

Accuracy

When it comes time to evaluate a model's classification performance there are several approaches, the most intuitive of which is accuracy. Accuracy is defined as

$$acc = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{2.3}$$

Where Tp are our true positives, Tn are our true negatives, Fp are our false positives, and Fn are our false negatives. It should be noted that normal values of accuracy for classification tasks depend highly on the data at hand. Noisy data or a large number of classes can both artificially drive accuracy scores down.

Precision

But what if we want to know, out of the total number of guesses for a particular class, what fraction were correct? For this, researchers typically use the Precision, defined as

$$p = \frac{Tp}{Tp + Fp} \tag{2.4}$$

Recall

Conversely, if we wish to know out of the total number of cases we could have guessed correctly, what fraction we *did* guess correctly then Recall is typically used. Recall is defined as

$$r = \frac{Tp}{Tp + Fn} \tag{2.5}$$

F1 Score

The F1 score is a way of combining the above two metrics Precision and Recall into one wholistic measure. Conceptually it is the harmonic mean of the Precision and Recall where we assign even weights to each. The F1 score is defined as

$$\begin{aligned} F_1 &= \frac{1}{\frac{1}{2}\left(\frac{1}{p} + \frac{1}{r}\right.} \\ &= \frac{2pr}{p + r} \end{aligned} \tag{2.6}$$

2.1.4 Document Clustering

Another well established task for topic models is document clustering. LDA has been used to successfully cluster a range of documents such as news articles and legal judgements (Lu, Mei, and Zhai, 2011; Xie and Xing, 2013; Kumar and Raghuveer, 2013). As opposed to classification where we want to assign an explicit label to each document, with clustering we wish to evaluate how well the resulting document topic vectors separate the documents into a meaningful structure.

Adjusted Rand Index

One way of accomplishing this is by using the **Adjusted Rand Index**, a score which measures the similarity of two sets of class labels; namely the true labels U and those predicted by a clustering algorithm V. We may calculate the raw (unadjusted) Rand index following equation 2.7 (Hubert and Arabie, 1985).

$$RI = \frac{a+b}{U_2^{n_{\text{samples}}}} \tag{2.7}$$

Where a is the number of pairs of elements in U belonging to the same class, and in V belonging to the same class. Conversely b is the number of pairs of elements in U belonging to different classes, and in V belonging to different classes. Finally, $U_2^{n_{\text{samples}}}$ is the total number of possible pairs in the dataset. In order to ensure that random labelings receive a score of zero we define the Adjusted Rand Index as

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{2.8}$$

Values of the Adjusted Rand Index range from -1 to 1, with a score of indicating a perfect match between the predicted and true class labels. Though the ARI requires the ground truth labels, it does have a unique beneficial trait. Namely that random labelings receive scores near 0, such that we may gauge how close our predicted labels are to a random guess.

Normalized Mutual Info

The Normalized Mutual Info is another method of evaluating clustering performance that has been successfully applied in the context of topic modelling (Xu, Liu, and Gong, 2003; Cai et al., 2008). It again assumes we have two sets of labels U and V, over N objects. We define the entropy of a label set U in equation 2.9, where $P(i) = |U_i|/N$ is the probability that a random object from U falls into class U_i.

$$H(U) = \sum_{i=1}^{|U|} P(i)log(P(i)) \tag{2.9}$$

The mutual information (MI) between U and V can be expressed as

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i,j)log(\frac{P(i,j)}{P(i)P(j)}) \tag{2.10}$$

With these two components we can write the normalized mutual information as proposed in (Vinh, Epps, and Bailey, 2009).

$$NMI(U,V) = \frac{MI(U,V)}{\sqrt{H(U)H(V)}} \tag{2.11}$$

Similar to the ARI, random labelings under the NMI criterion receive a score close to 0, though values for NMI range from 0 to 1. Additionally, the NMI has the advantage of being unbiased in its assumptions of cluster structure, and has no preference towards either isotropic or "folded" cluster shapes.

Homogeneity

Homogeneity is a clustering metric that measures how purely each cluster contains a single class. It requires the cluster labels K and class labels C and is defined as

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{2.12}$$

Where H(C | K) is the conditional entropy of the classes given the cluster assignments, and H(C) is the entropy of the classes. We express these as

$$\begin{aligned} H(C|K) &= -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{n_{c,k}}{n} log(\frac{n_{c,k}}{n_k}) \\ H(C) &= -\sum_{c=1}^{|C|} \frac{n_c}{n} log(\frac{n_c}{n}) \end{aligned} \tag{2.13}$$

Where n, n_c, n_k represent the total number of samples, the number of samples belonging to class c and to cluster k respectively. $n_{c,k}$ then represents the number of samples from class c assigned to cluster k.

One benefit of this metric is that it remains agnostic as to the shape of clusters, they need not belong to a certain distribution. Additionally, the homogeneity is nicely set between 0 and 1 with higher scores being more desirable. The down side however is as mentioned before, homogeneity requires the ground truth labels which in practice can be quite limiting.

Completeness

Completeness is a clustering metric complementary to Homogeneity outlined above. In fact if one reverses the class and cluster labels they would obtain the completeness. Following the same definitions as with Homogeneity, completeness becomes

$$c = 1 - \frac{H(K|C)}{H(K)} \tag{2.14}$$

Conceptually though the completeness answers a different question, namely how many members of a single class reside in the same cluster. This has much the same benefits and drawback as Homogeneity: it requires the ground truth labels, but allows for various cluster shapes and yields a score bounded between 0 and 1.

Silhouette Coefficient

One clustering measure that operates without the need for true labels is the Silhouette Coefficient. The Silhouette Coefficient score evaluates how dense and disparate each of the clusters are. This is characterized by clustered points close together with centroids further apart which we measure as

$$s = \frac{b - a}{max(a, b)} \tag{2.15}$$

Where a is the mean distance between a sample and all other points in the same *class*, and b is the mean distance between a sample and all other points in the next nearest *cluster*. The score is bounded from -1 to 1, again with higher scores corresponding to improved clustering.

Aside from not requiring the true class labels, one advantage of using the Silhoutte Coefficient is that its definition of a 'good' cluster is somewhat intuitive. Points within a cluster should ideally be closer together, and the clusters themselves should remain as distinct as possible. However one downside of the Silhoutte Coefficient score is its sensitivity to the convexity of clusters, i.e. using methods such as DBSCAN can distort the value of the score. (Though methods such as K-means are acceptable.)

2.2 DTM Model Overview

This section briefly outlines the dynamic topic model (**DTM**), following closely the original derivation found in (Blei and Lafferty, 2006). As this is intended as more of a summary, we recommend the reader examine the original paper for a complete exposition of the mechanics of the DTM.

In our primer on LDA (in section 1.2) we outlined the conceptual basis for static LDA topic models. Namely, that topics consist of a distribution over a fixed vocabulary and are determined by a set of hyper-parameters β. Additionally each document is represented as a combination of topics, with proportions controlled by their corresponding set of hyper-parameters α. Roughly speaking, the goal of the Dynamic Topic Model (**DTM**) is to account for the drift in topics over time by chaining together a series of static LDA models. This is accomplished by tying the hyper-parameters $\alpha_{t-1}, \beta_{t-1}$ at time step $t-1$, to the hyper-parameters α_t, β_t at time step t. The result is a model that allows us to track how our topics evolve at each time step.

2.2.1 Chaining models together

The question then, is how do we tie our hyper-parameters together? Well, regularly with static LDA we would simply use a Dirichilet distribution to model our uncertainties in word distributions (hence the name). Unfortunately, the Dirichilet distribution does not lend itself to sequential modeling, which eliminates this option. Instead, we make a state-space model that evolves with Gaussian noise to chain together the natural parameters of each topic $\beta_{z,t}$ such that each topic "evolves" from the last.

$$\beta_{z,t}|\beta_{z,t-1} \sim \mathcal{N}(\beta_{t-1,z}, \sigma^2 I) \tag{2.16}$$

Similarly, with static LDA we would also pull our document specific topic proportions θ from the Dirichilet distribution. For the same reason as above this is no longer an option. So to express our uncertainty over our topic proportions, we use a logistic normal with mean α. Then we chain our topic proportions together using the same trick as we did above with word distributions, (by using Gaussian noise). This yields the graphical model in figure 2.1.

$$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I) \tag{2.17}$$

2.2.2 Variational approximate inference

Because we have used the Gaussian distribution to model the progression of our parameters, inference becomes intractable due to the non-conjugacy of Gaussian and multinomial models. To get around this we take the same variational approach to approximate inference as before in section 1.2 with static LDA. Taking a variational approach has the advantage of allowing

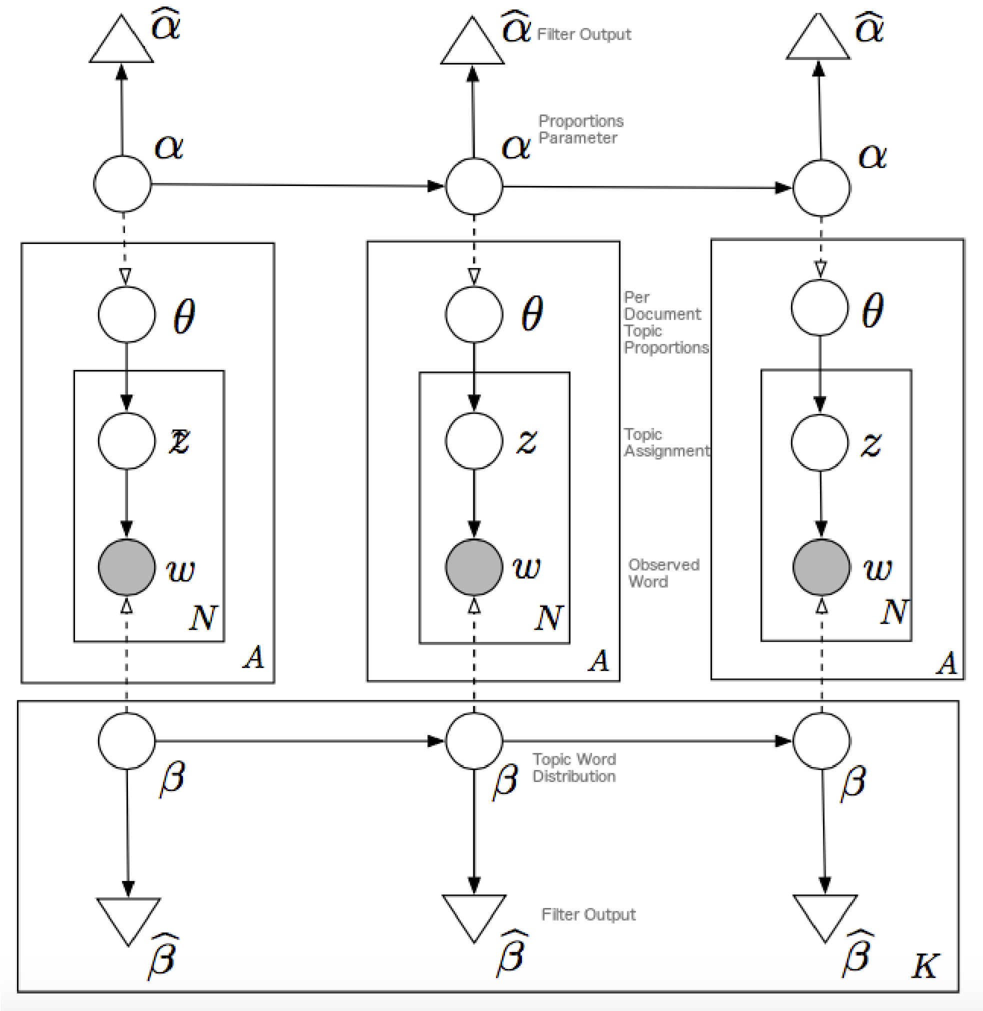


FIGURE 2.1: Graphical model for DTM showing a series of chained static LDA models. The triangles represent the Kalman filter estimates of the hyper-parameters

us to handle larger document sets compared to Gibbs sampling which becomes computationally difficult at large corpus sizes.

The general strategy of variational approximate inference is to use a carefully tuned 'approximate' distribution as a substitute for the true posterior. To tune this approximate distribution we minimize the KL divergence between our estimated and true posterior. Finally we may then use this approximated posterior distribution to perform inference.

We begin by creating a collection of variational parameters we will optimize over our latent variables. Our latent variables are the topics $\beta_{t,k}$, topic proportions $\theta_{t,d}$, and topic indicators $Z_{t,d,n}$. While we have variational parameters for each topic (consisting of a sequence of multinomial parameters), and for each document (the latent topic proportions). The resulting posterior, again following the notation of (Blei and Lafferty, 2006) is given by equation 2.18.

$$\prod_{k=1}^{K} q(\beta_{k,1}, ..., \beta_{k,T} | \hat{\beta}_{k,1}, ..., \hat{\beta}_{k,T}) \times$$
$$\prod_{t=1}^{T} \Big(\prod_{d=1}^{D_t} q(\theta_{t,d}) | \gamma_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n} | \phi_{t,d,n}) \Big) \tag{2.18}$$

Now we are ready to tune our approximate posterior, and specifically the variational observations $\{\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,T}\}$ according to the KL divergence between the estimated and the true posterior. Note that here, each topic proportions vector $\gamma_{t,d}$ receives a corresponding free Dirichilet parameter, while each topic indicator $z_{t,d,n}$ receives a corresponding free multinomial parameter $\phi_{t,d,n}$. To optimize the document topic proportion vectors we subsequently employ gradient ascent, however this is not necessary for the document level parameter updates as they simply have a closed form.

Finally, we may track our variational parameters $\hat{\beta}$ and $\hat{\alpha}$ between time slices using either the Kalman filter or wavelet regression. For brevity we will not replicate the mechanics of these methods here and encourage the interested reader to refer to the derivations provided in detail in (Blei and Lafferty, 2006).

2.3 DIM Model Overview

In this section we provide an overview of the Dynamic Influence Model (**DIM**) following closely the derivation found in (Gerrish and Blei, 2010). For brevity, we do not replicate all technical aspects of the model here and encourage the interested reader to examine the original paper for a complete exposition of the mechanics of the DIM.

2.3.1 DIM Purpose

Aside from generating thematic topics, the unique purpose of the DIM is to provide estimates of document influences on each of those topics. The novelty of the DIM lies in the fact that this influence measure is inferred solely from language use. This makes the DIM particularly useful when traditional bibliometrics such as citations are unavailable, and in such cases inferred document influences could act as a proxy as the two are known to correlate (Gerrish and Blei, 2010). The quantitative estimate of a document's influence that DIM provides is valuable as it allows for more informed decisions about publishing and funding, and can help us assign 'value' to documents in an automated manner.

2.3.2 Encoding Influence

The DIM itself is a close relative to the DTM, as outlined above in section 2.2; it too consists of a series of individual LDA models connected via a Markov chain of term distribution parameters β. Naively we may describe the drift of this distribution as a stationary autoregressive process with transition variance σ^2, given by equation 2.19.

$$\beta_{t+1}|\beta_t \sim \mathcal{N}(\beta_t, \sigma^2 I) \tag{2.19}$$

This in essence is the simplest form of the DTM, where we compute the posterior distribution of our sequence of topics $\beta_{1:T}$ conditioned on the observed documents, and represent our corpus as a smooth trajectory of word frequencies.

However, with the DIM we wish to encode document influences into this model of topic drift. To do this we need a method of describing the probability of a word dependent on a set of natural parameters β_t. We write this probability as the softmax transformation of the unconstrained vector.

$$p(w|\beta_t) \propto exp(\beta_{t,w})$$

Additionally, we assign to each document d a normally distributed scalar influence score ℓ_d that describes the influence of that document on a given topic. Where a higher influence score indicates that a document's words had a larger impact on the drift of this topic. Finally, we modify our naive Markov chain of term distributions from equation 2.19 to relate the next epoch's set of natural parameters β_{t+1} to the probability of words under the *current* set of natural parameters β_t, and the influence score of each document ℓ_d.

$$\beta_{t+1}|\beta_t, (w, \ell)_{t,1:D} \sim \mathcal{N}(\beta_t + exp(-\beta_t) \sum_d w_{d,t}\ell_{t,d}, \sigma^2 I) \tag{2.20}$$

In this way, the words of a high influence document will have a higher expected probability in the following time step. In effect, each topic's natural parameters are "nudged" by the words of each document in proportion to the influence score of that document. This is a core principle of the DIM, that not all documents exert the same force on the direction of a topic, some documents have a larger impact, a larger *influence* on a given topic than others. As a result, an article whose words can help explain the way the word frequencies change will have a high posterior influence score.

Something is still missing though, equation 2.20 only expresses the progression of natural parameters for a single topic. For multiple topics we generalize this expression to equation 2.21, defining $z_{d,n,k}$ as the indicator that the nth word in document d is assigned to topic k.

$$\beta_{k,t+1}|\beta_{k,t},(w,\ell,z)_{t,1:D} \sim \mathcal{N}(\beta_{k,t} + exp(-\beta_{k,t})\sum_d \ell_{d,k} \sum_n w_{d,n}z_{d,n,k}, \sigma^2 I) \tag{2.21}$$

Where to accommodate multiple topics we have kept the influence score of a document outside the summation over the possible indicator assignments. This yields the graphical model illustrated in figure 2.2. If we compare this to the graphical model for the DTM present in figure 2.1 we can observe how the DIM differs.

Particularly, we no longer chain our document topic-proportions α between time steps; instead each epoch's corresponding LDA receives the same set. Additionally the topic term-distribution hyper parameters of a given timestep β_t depend not only the previous timestep's params β_{t-1} but also the indicators z, terms w, and influences ℓ of the past timestep. Together these differences constitute a significant enough change to create a noticable difference in results between the DIM and DTM, which we report in Chapter 4.

2.3.3 Inference with DIM

Conducting inference with the DIM is largely similar to the DTM. Again, computing the exact posterior is intractable so we turn to variational methods to approximate it with a simpler distribution over latent variables. Here our latent variables are the sequence of topics and the per-document influence values, conditioned on an observed corpus.

Our simple approximation of the true posterior is given by equation 2.22, which we call our *variational distribution*. To specify this variational distribution we introduce a set of free variational parameters which we fit to minimize the KL divergence between the variational distribution and the true posterior.

$$\begin{gathered} q(\beta,\ell,Z,\theta|\widetilde{\beta},\widetilde{\ell},\phi,\gamma) = \prod_{k=1}^{k} q(\beta_{k,1:T}|\widetilde{\beta}_{K,1:T}) \\ \prod_{t=1}^{T}\prod_{d=1}^{D_t} q(\theta_{t,d}|\gamma_{t,d})q(\ell_d|\widetilde{\ell}_d)\prod_{n=1}^{N_{t,d}} q(Z_{t,d,n}|\phi_{t,d,n}) \end{gathered} \tag{2.22}$$

The first such variables we introduce include the word assignments z_n

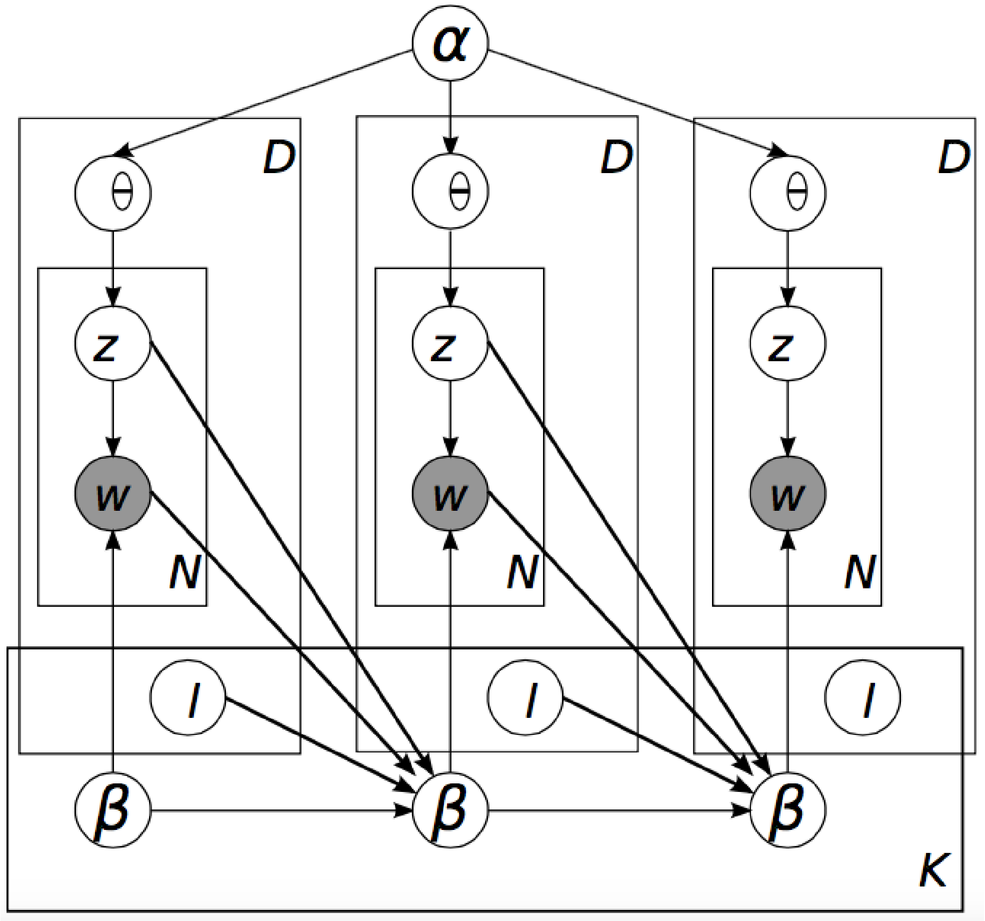


FIGURE 2.2: Graphical model of the DIM. Note the increased number of dependencies for the term distribution parameters β compared to that of the DTM.

and topic proportions θ_d which are governed by multinomial parameters ϕ_d and Dirichilet parameters γ_d. Additionally, in order to govern the variational distribution for topic trajectories $\{\beta_{k,1}, ..., \beta_{k,T}\}$, described by a linear gaussian chain, we specify the set of parameters $\{\widetilde{\beta}_{k,1}, ..., \widetilde{\beta}_{k,T}\}$ to act as the "variational observations" of the chain. Finally, in order to describe the variational distribution of the document influence value $\ell_{d,k}$ as a Gaussian we introduce $\widetilde{\ell}_{d,k}$ representing the mean, and σ_ℓ^2 representing the fixed variance.

Finaly, as stated above, we obtain estimates for these variational parameters by fitting our variational distribution, minimizing the KL divergence via gradient descent. It is here that we end our brief summary of the DIM. For more details such as parameter updates, and initial empirical results, we direct the interested reader to the full original exposition found in (Gerrish and Blei, 2010).

Chapter 3

Experimental Set Up

In the interest of reproducibility we include in this section some of the considerations specific to our data set. Additionally we cover the data preprocessing steps taken prior to our experiments and describe our process for model tuning. Finally, we outline the procedure taken for the coherence testing, classification and clustering experiments.

3.1 Data Considerations

The data we used for our experiments comes from the August 2014 EPO Worldwide Patent Statistical Database (**PATSTAT**). According to the EPO, this data set contains over 90 million patent documents from leading industrialized and developing countries, with some documents as early as the nineteenth century. Thankfully, the expanse of this data is met with an equal level of documentation[1].

Due to time and computation limits, we ran experiments only on a subset of this data rather than the whole corpus. In order to break the PATSTAT into manageably sized subclasses we made use of Cooperative Patent Classification (**CPC**) labels. The CPC labels are a hierarchical system used globally to annotate patents according to the area of technology to which they relate. Predominantly we carried out our research on patents under the umbrella of **Y02E 10/20**, the subclass of patents relating to **hydro energy**. This subclass itself consisted of 5 smaller subsets of documents with more specific labels which we treated as true labels during our classification experiment, namely "Hydro energy", "Conventional", "Turbines and wheel", "Other parts", and "Stream and damless".

An additional consideration in filtering our data was the International Patent Documentation Centre **(INPADOC) patent family**. A patent family is a collection of patents filed in various countries to protect the same invention. We ensure that only one member of each patent family participates in the study so as not to ′double count′ the same invention. Furthermore, only patents filed in English are considered. Figure 3.1 contains the schema for the PostgreSQL database we constructed to hold relevant patent data.

[1] For more information the official catalogue can be found at http://goo.gl/LRQWnu

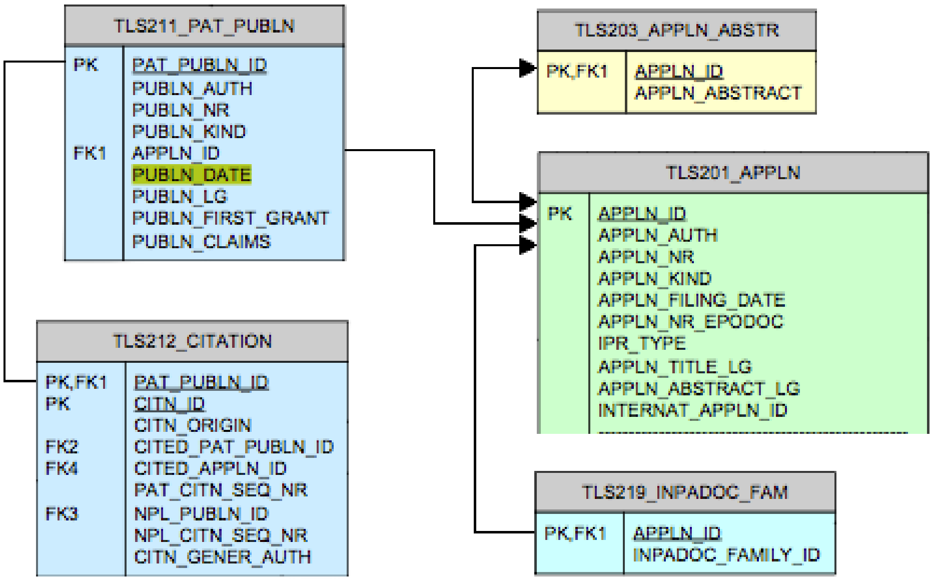
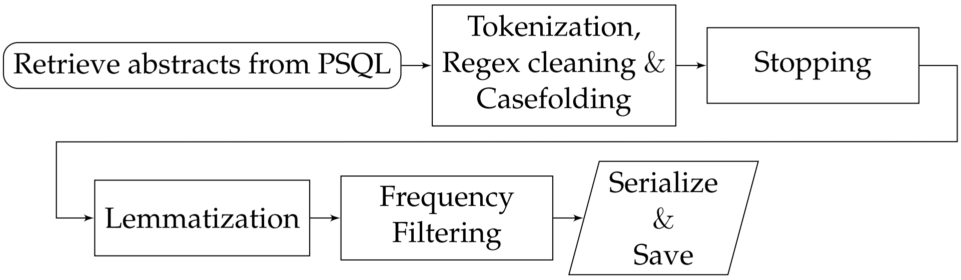


FIGURE 3.1: Schema of PostgreSQL database containing relevant tables

3.2 Data Pre-Processing

The phrase "garbage in garbage out" is commonly used in machine learning to emphasize the importance of the data pre-processing step. We begin our pre-processing by retrieving abstracts from the aforementioned PSQL database and applying case folding such that words like "Turbine" at the beginning of a sentence will match the word "turbine" elsewhere. We then remove unnecessary symbols and punctuation via regex and apply stopping using the English **NLTK** stopwords list. This effectively removes common conjunctions and operational words from our corpus that don't actually contribute to the topics, such as "but, if, the, and" etc.


Retrieve abstracts from PSQL
Tokenization, Regex cleaning & Casefolding
Stopping
Lemmatization
Frequency Filtering
Serialize & Save

We then further distill the abstracts by applying lemmatization. This step is meant to reduce inflectional forms of a word such as "operates, operating, operational" to a base form "operate". After lemmatization, we remove any words occurring less than 25 times total, matching the implementation found in (Blei and Lafferty, 2006). Finally, we save the resulting corpus, totaling over 6,400 documents, and serialize the dictionary of vocabulary for subsequent access.

3.3 Tuning models

The hyperparameters we select for our models can have a large influence on the topics that they infer. For this reason, prior to running any experiments, it is important to tune the hyperparameters of our topic models. The parameters we tuned specifically, were the number of topics K and the maximum number of iterations M_{iter}. To do this we defined a range of values for these parameters and sampled parameter sets from this space uniformly. At each parameter set we evaluated the **umass**, **uci**, **npmi**, and **cv** topic coherence of the resulting model.

In order to inspect the relationship our parameters had with topic coherence we visualized our results. Figure 3.2 displays the c_uci coherences of the DTM as a function of K and M_{iter}. From this plot we can see that increasing the number of topics K beyond a certain point fails to improve topic coherence with similar results for the number of iterations.

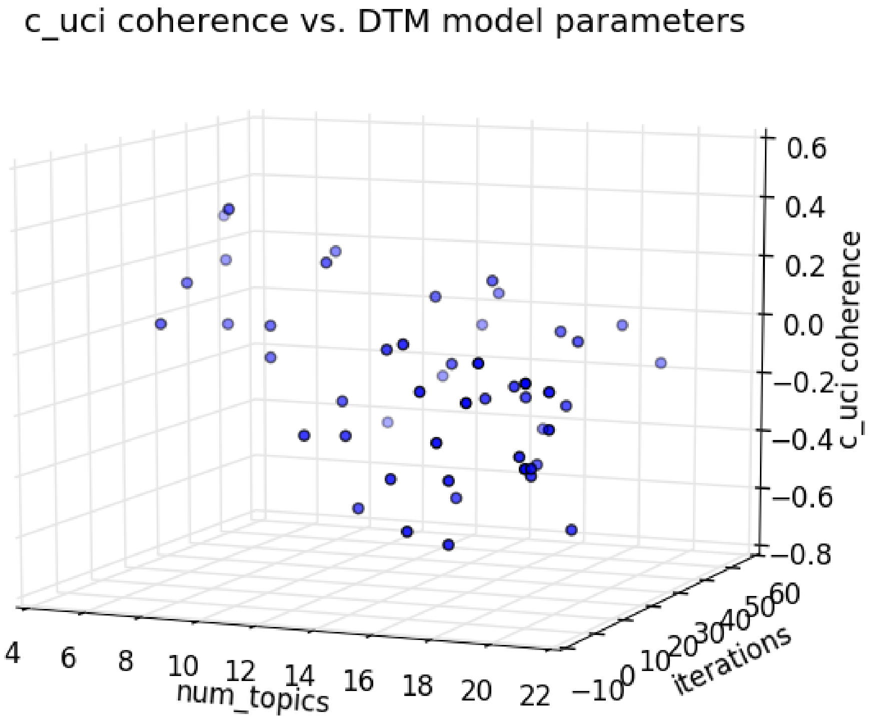


FIGURE 3.2: c_uci coherence scores for the DTM as a function of K and M_{iter}

Table 3.1 contains the best coherence scores obtained by each model for each measure after sampling our constrained parameter space, with higher scores being more desirable. Despite tuning the models on a smaller set of documents it is interesting to note that a trend emerges, namely that the DTM tends to outperform both the DIM and the static LDA model in terms of topic coherence. However we reserve judgement until obtaining the final coherence testing results listed in section 4.1.

TABLE 3.1: The peak coherences achieved for various models and parameter choices.

Model	c_v	c_uci	c_npmi	u_mass
LDA	.4871	.1455	.0694	-0.9813
DTM	.5999	.4098	.1287	-1.4496
DIM	.5789	.1145	.1175	-1.6033

Though the optimal parameter sets suggested by each coherence measure did not greatly differ from one another, ultimately we selected the parameters recommended by the c_v measure. We preferred the judgement of the c_v measure as it is known to have the strongest correlation with human judgement (Röder, Both, and Hinneburg, 2015). The optimal parameter set found for the LDA model under the c_v measure was K = 13, M_{iter} = 40, while for the DIM it was K = 5, M_{iter} = 32, and for the DTM it was K = 17, M_{iter} = 36. These were the parameters we used for the full scale coherence testing reported in section 4.2.

3.4 Experimental Procedures

3.4.1 Coherence Testing

The coherence testing experiment follows much the same procedure as above for hyperparameter tuning, with the difference being a full data set, and defined model parameters. We begin by training the LDA model, DTM and DIM over the same set of input documents belonging to the **Y02E 10/20** hydro energy CPC label, ranging from the year 1997 to 2015. Each model was trained under the optimal parameters found via the hyperparameter tuning process from section 3.3.

Following model training, we retrieved all topics from the LDA model, DTM and the DIM, and assessed their semantic value via the **c_v**, **c_uci**, **c_npmi**, and **U_mass** coherence metrics. However because the DTM and DIM both output a *series* of topics rather than a single set, the topic coherences were tested at each time step. Results for this experiment can be found in section 4.2.

3.4.2 Classification

To asses the efficacy of each model at creating feature vectors useful for document classification we begin by taking the same models as above, (the LDA model, DTM and DIM, trained on the hydro electric patents) and extract the document topic vectors for each document.

We then use 80% of these unsupervisedly learned feature vectors to train a collection of classifiers, and tested on the remaining 20% using the CPC subclasses for each document as true labels. We cross validated each classifier to ensure more robust estimates of performance, and selected the results of the classifier with the highest cross validated F1 score for each topic model. To further analyze the performance of each topic model's peak classifier we made use of the classification metrics outlined in section 2.1.3 namely **accuracy**, **precision**, **recall**, and **F1 score**. Results for this experiment can be found in section 4.3.

3.4.3 Clustering

Lastly, we conduct a clustering experiment to determine how well each model creates internally consistent groups of documents that are distinct from one another. We begin with the same set of models and data as with the other experiments, namely the LDA model, DTM and DIM trained on hydro electric PATSTAT patents. As with the classification experiment, we again extract the document topic vectors for each document.

However for this experiment, rather than beginning by handing the vectors to a series of classifiers, we begin by visualizing the space they formed using **TSNE**, **PCA**, and **MDS** embeddings to inspect the clusters by eye. After visually inspecting each model's document vectors we performed K-means clustering with $K = 5$, (the number of true topics in the corpus). To evaluate clustering performance we feed the labels estimated by K-means for each set of document vectors, (alongside the respective true labels), to each of the clustering metrics outlined in section 2.1.4. These metrics include the **adjusted rand score**, **normalized mutual info**, **homogeneity**, **completeness** and **V-measure**. Results for this experiment can be found in section 4.4.

Chapter 4

Experimental Results

In this chapter we begin with a qualitative analysis of the posterior topic distributions generated by the DTM and inspect the historical significance and context of patents with particularly high inferred influence from the DIM. More quantitatively, we then present results from evaluating the coherence of each model's topics. Following the results of coherence testing we show the results of document classification via the document topic vectors produced by each of the three models. We conclude this chapter by presenting the performance of K-means clustering on each of the document vector spaces as measured by the metrics described in section 2.1.4.

4.1 Qualitative analysis of topics

The experimental results show that both the DTM and DIM successfully identify latent topic structures consistent with known industry history. As an example we inspect the topic generated by the DTM most closely associated with the "Stream and Damless" hydro power CPC patent subclass (those with the label **YO2E 10/28**). Figure 4.1 illustrates the mean topic vector for documents of this CPC subclass across epochs, primarily dominated by topic 4. If we inspect the top words from the inferred posterior distribution of this topic, shown in figure 4.2, we see that while central words such as "water" and "power" maintain a high likelihood, words such as "float" experience an increase in likelihood as time progresses.

To understand why this might be the case we reviewed the known history of stream and damless hydro technology, and found that the increase of the word "float" is of particular interest as it parallels closely the development of the wave and tidal energy industry both in the UK and globally. Vertical cross flow turbines such as the Gorlov turbine, patented in 2001, contributed to a rise in the popularity of floating barges for tidal power, a technology that represents the current norm of tidal current development (Khan and Bhuyan, 2009). Closely following this the European Marine Energy Centre (**EMEC**) was founded in 2003 and since then has supported the deployment of more wave and tidal energy devices than at any other single site in the world. This substantiates the linguistic trends we observe in Topic 4 and offers an explanation as to the rise in the estimated posterior mean number of occurrences of words such as "power" and "float" after

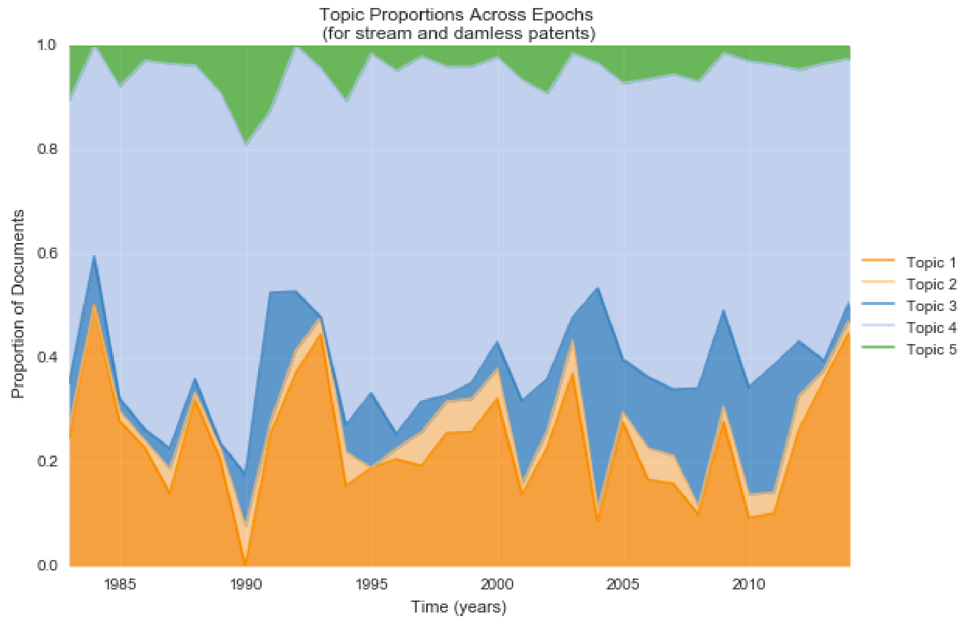


FIGURE 4.1: Proportion of topics at each epoch for patents relating to stream and damless hydro electric technology

2003.

Additionally, inspecting the patents in the "Stream and Damless" subclass that were estimated by the DIM as having a particularly high influence to Topic 4 reveals a similar narrative. During the same period of growth for barge based systems as discussed above, beginning around the year 2000, we see the DIM infer linguistic importance to related patents. For instance, a buoyancy pump in 2000, an energy generating method and device for utilizing buoyancy in 2006 and a cross flow hydraulic turbine in 2007 all received the largest influence scores of their epoch, a metric shown to correlate with forward citation rates (Gerrish and Blei, 2010). In this way we are able, *without* knowledge of citation rates, i.e. solely through analyzing language use, to identify individual patents likely to have had an influence on the terminology of the field in which they were published.

4.2 Coherence Testing

The aggregated topic coherence results of the tuned models are shown in Table 4.1 for the DTM, DIM and LDA models respectively.

When testing the coherence of each model's topics, the DTM proved superior in all coherence measures except Umass, the measure that correlates the least with human judgement by a considerable amount. Not only did the DTM regularly obtain the highest coherence scores, it did so by a large margin obtaining nearly double the c_uci score obtained by static LDA. We attribute this performance to the unique structure of the DTM. Rather than

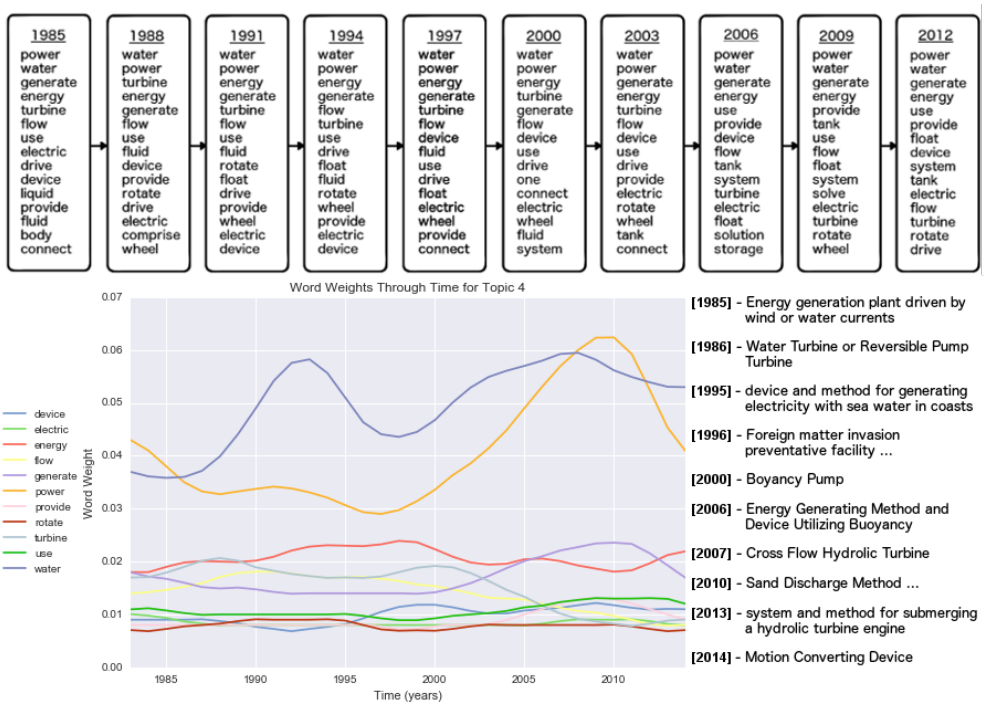


FIGURE 4.2: (top) The top fifteen words from the inferred posterior distribution of Topic 4 at three year lags. (bottom) The posterior estimate of the frequency as a function of year of words from Topic 4. (right) Patents of linguistic influence to topic across epochs as estimated by the DIM.

enforcing broad sweeping "one size fits all" topics across the whole corpus as with traditional LDA, with the DTM each epoch receives in effect its own LDA model. Each of these individual LDA models inherits a set of variational parameters α and β, from the previous time step that have been perturbed by Gaussian noise and control its document topic-proportions and topic word-distributions respectively. The result is a richer posterior compared to static LDA that allows for a 'tighter fit' to the true posterior. This naturally takes longer to train but is demonstrably worth the performance increase.

Though the DIM received the highest U_mass score, it is interesting to note that for the most part the DIM obtained coherence scores similar to, but consistently smaller than, that of its counterpart the DTM. This is again due to the structure of the model. The DIM borrows a similar Markov chain structure of word distributions in order to capture drifts in probabilities over the course of the document collection, but with one critical difference. While the topic word-distribution natural parameters β are passed along at each time step the DIM does not chain its document topic-proportion natural parameters α. Thus each epoch's LDA operates under the same α parameters which weakens its ability to superresolve temporal topic trends. This explains the increased performance of the DIM over traditional LDA, and reduced performance compared to the DTM.

TABLE 4.1: The coherence values attained by each model

Model	c_v	c_uci	c_npmi	u_mass
LDA	.5021	.1904	.0750	-1.6283
DIM	.5581	.2865	.1011	**-1.3478**
DTM	**.5980**	**.4373**	**.1213**	-1.6688

Results are analogous when looking at coherences across epochs, illustrated in figure 4.3, as opposed to the aggregated coherence scores. As the static LDA produced only a single set of topics for the entire corpus its coherence scores remain constant. The DTM however provided unique topics across epochs and thus obtained a sequence of coherence scores. This sequence of scores contained minor fluctuations but again persistently remained above those of the static LDA model for all measures except Umass.

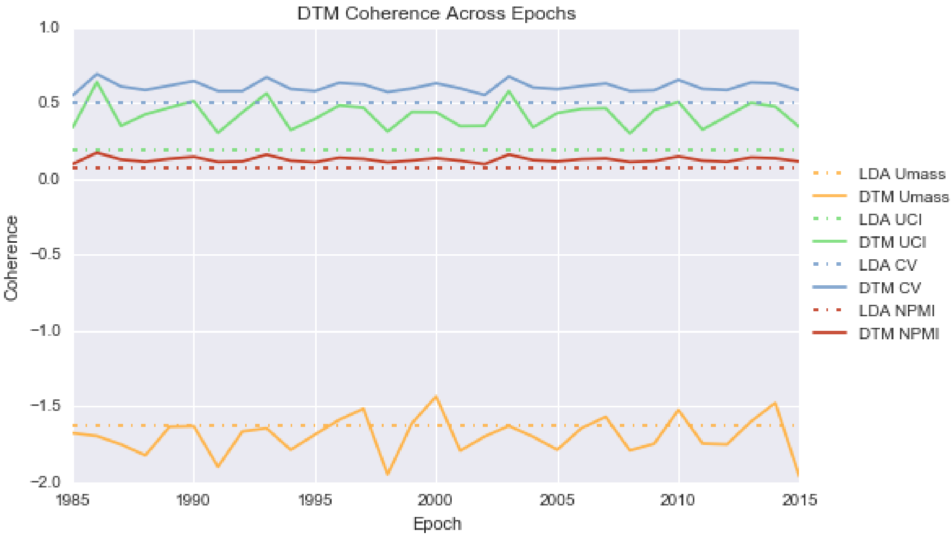


FIGURE 4.3: Coherence scores for the DTM and LDA across epochs

4.3 Classification Results

At classification the DTM again proved superior, with a final precision, recall and F1 score of .64, .64, and .64 respectively. However the DIM performed marginally worse than the static LDA model, a contrast to the DIM's improved topic coherences compared to static LDA. The DIM generated a final precision, recall, and F1 score of .58, .56, and .56 while the static LDA model yielded precision, recall and F1 scores of .58, .58, and .58.

Table 4.2 contains the cross validated F1 scores of each classifier trained on the document topic vectors provided by static LDA, the DTM and DIM. We used radial basis function support vector machines to generate the final classification results as this classifier yielded the highest cross validated F1 score across all three models.

The precision, recall, and F1 score for LDA, DIM and the DTM, broken down by patent class labels, can be found in tables 4.3, 4.4, and 4.5 respectively. These results are illustrated in their corresponding normalized confusion matrices found in figures 4.4, 4.5, and 4.6. Patents relating to "turbines and wheels" tended to be the most difficult to classify while "conventional" patents were generally the best classified. The precision and recall was reasonably balanced for the classifiers of all three models, with the DIM based classifier favoring precision slightly.

TABLE 4.2: Cross validated accuracy score for each classifier

Model	DTM CV score	DIM CV score	LDA CV Score
ASGD	0.621013	0.521755	0.556590
Adaboost	0.567315	0.504048	0.514788
Decision Tree	0.491707	0.546153	0.490798
Gaussian NB	0.582415	0.540346	0.530325
KNN	0.553341	0.472817	0.494431
LDA	0.618973	0.534562	0.551576
LR	0.620385	0.521948	0.555715
Lin. SVC	0.564651	0.447450	0.510240
Passive-Aggressive I	0.601775	0.506445	0.450347
Passive-Aggressive II	0.385359	0.378960	0.397031
Perceptron	0.458625	0.379983	0.434744
QDA	0.585174	0.453922	0.529446
RBF SVM	**0.637477**	**0.564730**	**0.578102**
Random Forest	0.524156	0.547191	0.472841
SGD	0.602344	0.514596	0.550329

TABLE 4.3: Classification results of LDA based classifier

Class Label	Precision	Recall	F1
Hydro energy	0.49	0.52	0.50
Conventional	0.62	0.65	0.64
Turbines and Wheels	0.45	0.44	0.44
Other Parts	0.61	0.57	0.59
Stream and Damless	0.63	0.62	0.62
Avg/Total	0.58	0.58	0.58

TABLE 4.4: Classification results of DIM based classifier

Class Label	Precision	Recall	F1
Hydro energy	0.38	0.44	0.41
Conventional	0.68	0.60	0.63
Turbines and Wheels	0.42	0.43	0.43
Other Parts	0.56	0.66	0.61
Stream and Damless	0.68	0.53	0.59
Avg/Total	0.58	0.56	0.56

TABLE 4.5: Classification results of DTM based classifier

Class Label	Precision	Recall	F1
Hydro energy	0.70	0.65	0.67
Conventional	0.69	0.69	0.69
Turbines and Wheels	0.49	0.50	0.49
Other Parts	0.63	0.62	0.63
Stream and Damless	0.66	0.69	0.67
Avg/Total	0.64	0.64	0.64

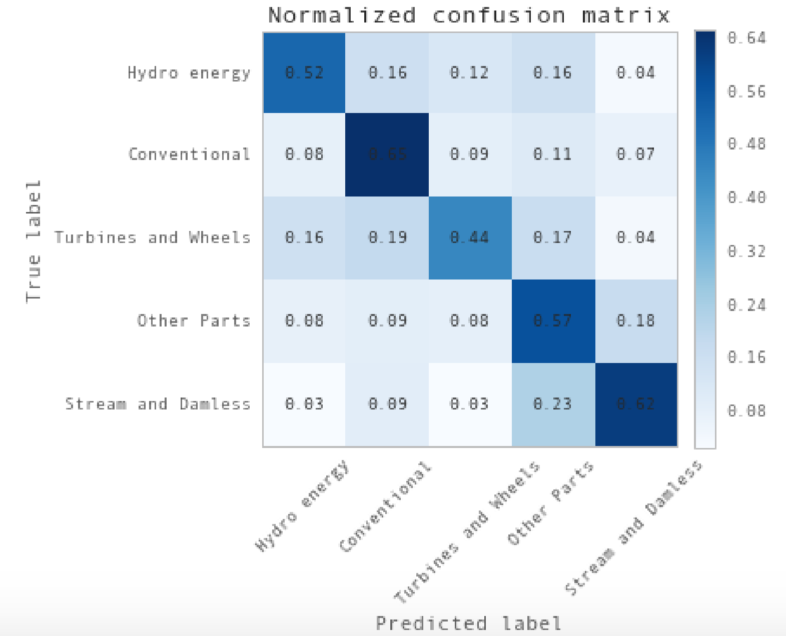


FIGURE 4.4: Confusion matrix of LDA classification results

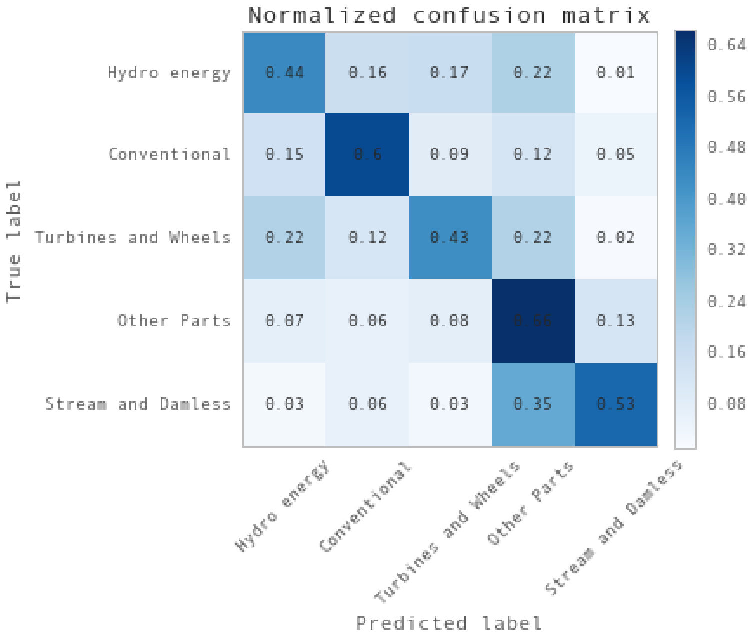


FIGURE 4.5: Confusion matrix of DIM classification results

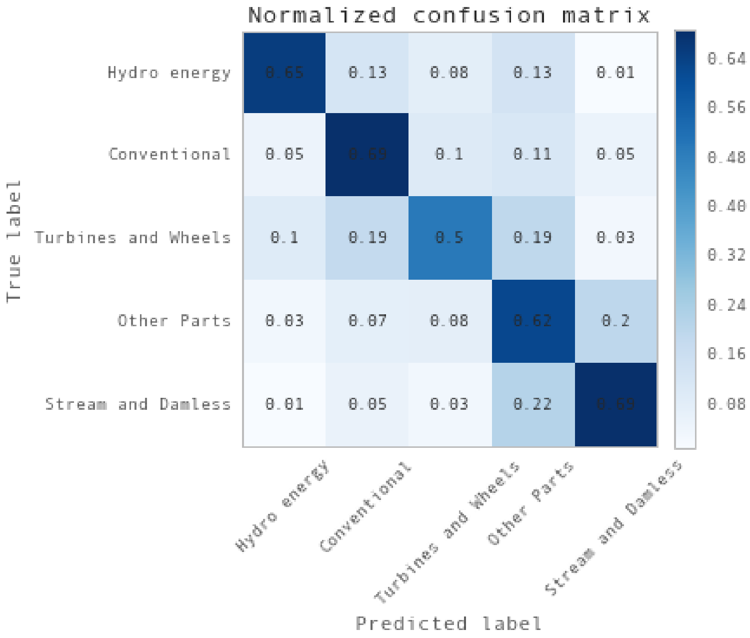


FIGURE 4.6: Confusion matrix of DTM classification results

4.4 Clustering Results

Here, we found that the vector spaces created by the DIM, and particularly the DTM, also excelled at effectively clustering documents. Scores for the various clustering metrics we calculated can be found in table 4.6. The DTM received the highest silhouette score at .55, more than double the score of .22 received by LDA, meaning it provided the densest and most distinct clustering of the documents irrespective of their class labels.

When we take into account the true class labels of each patent, using the normalized mutual information (NMI), we see that the DTM again produces the highest score. This suggests a higher agreement between the DTM's K-means predicted labels and the true labels compared to those of the other models. However, the DIM achieved the highest adjusted rand index (ARI) of the three models indicating a high similarity between its K-means predicted labels and the true labels as well. Irrespective of this, the traditional static LDA received both the lowest NMI and ARI scores.

Finally, we observed that the dynamic models both produced higher homogeneity and completeness scores than the static LDA model, with the DTM again leading slightly over the DIM. The DTM achieved a homogeneity of .21 indicating that its corresponding kmeans cluster assignments each contained more members of a single class than had the other models' cluster assignments. Additionally the completeness of the DTM, measuring how well all members of a given class are assigned to the same cluster, was the highest at .207 with the DIM *just* behind at .205 and LDA at .175.

To visualize the vector spaces generated by each model we made use of a handful of embedding algorithms, namely t-Distributed Stochastic Neighbor Embedding (**t-SNE**) , Principle Component Analysis (**PCA**), and Linear Discriminant Analysis (**LDA**), to project the spaces down to 2 dimensions. The results of this visualization can be seen in figure 4.7. It is interesting to observe how strikingly different the vector spaces produced by the DTM, DIM and LDA model are when plotted side by side. However some themes run constant throughout such as the proximity of "turbines and wheels" and "other parts" patents, represented by the two shades of green. This is substantiated by the classification confusion matrices illustrated in section 4.3 where we see that classifiers tended to have difficulty in assigning labels for "turbines and wheels" and "other parts" patents, more commonly mistaking them for each other than for other classes.

TABLE 4.6: Clustering results

metric	**DTM**	**DIM**	**LDA**
Silhouette	0.559518	0.370500	0.227973
NMI	0.210324	0.208029	0.177195
ARI	0.140005	0.158060	0.108182
Homogeneity	0.213513	0.210767	0.178718
Completeness	0.207182	0.205327	0.175685

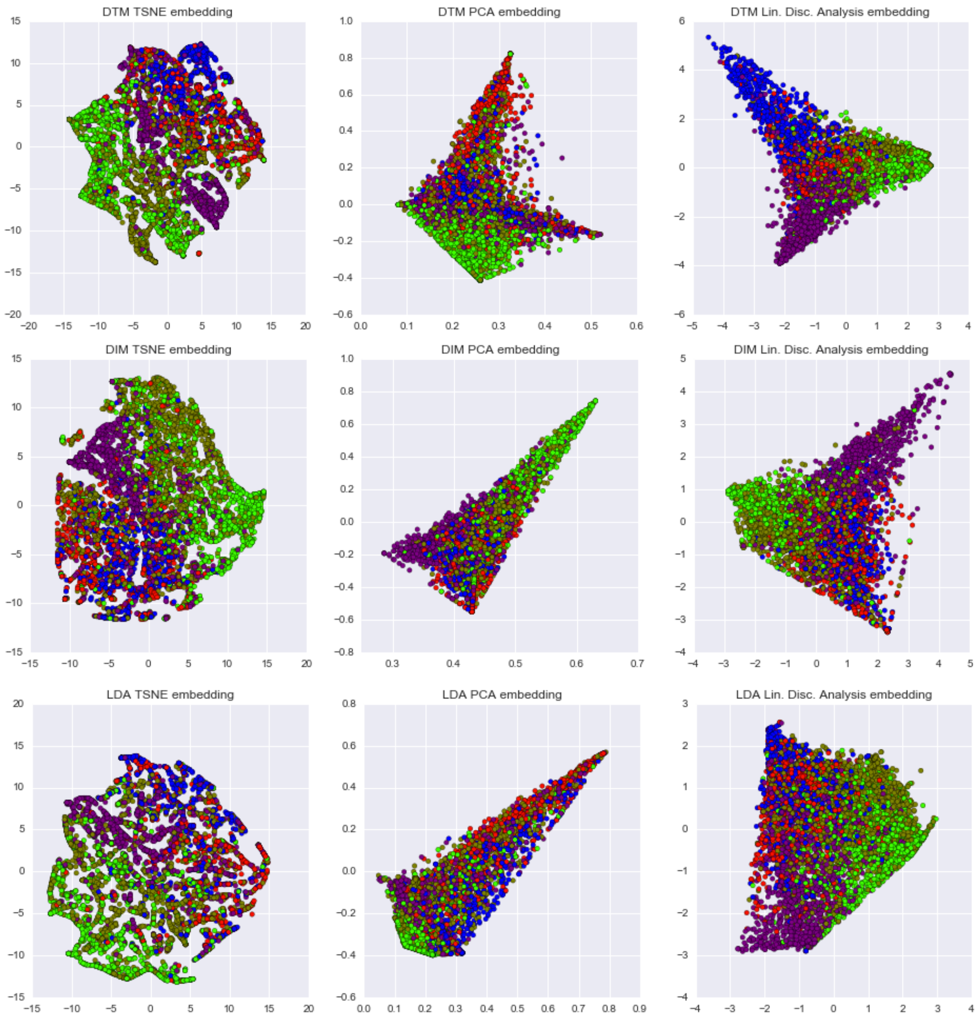


FIGURE 4.7: TSNE, PCA and Linear Discriminant Analysis embeddings for the vector spaces generated by DTM, DIM, and LDA model

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis we have implemented two extensions of latent dirichilet allocation for topic modeling of large corpuses, namely the dynamic topic model (**DTM**) and dynamic influence model (**DIM**). The aim of these extensions is to leverage and encode the chronological information contained in a set of documents to improve the quality of inferred topics. We trained both models on hydo electric patent abstracts from the August 2014 EPO Worldwide **PATSTAT** Database and demonstrated their effectiveness in a series of topic coherence, document classification and document clustering experiments. We found that both the DTM and DIM consistently outperformed traditional static LDA across all experiments.

As a preliminary experiment, we qualitatively inspected for a sample topic how the evolution of its word distribution complemented the known industry history of its corresponding patents. We found that not only did the dynamic models identify a major technological trend in stream and damless hydro electric technology, but successfully identified patents representative of this technical change.

More quantitatively, we tested the coherences of the topics each model produced, using a collection of measures established in the literature, that were designed to correlate with human judgements, namely **c_v**, **c_uci**, **c_npmi**, and **u_mass**. While the u_mass coherences proved inconclusive due to the metric's comparatively low correlation with human judgement, the results of the other metrics c_v, c_uci, and c_npmi were unanimous. The DTM produced verifiably more coherent topics than DIM, which also scored higher than static LDA.

When we tested the efficacies of the document vector spaces produced by each model at document classification we found a similar theme. Particularly, that the DTM based classifier yielded the highest precision, recall, and F1 score when classifying the patents, while the DIM based classifier experienced modestly diminished performance and the LDA based classifier experienced noticeably diminished performance. By the time we tested the clustering potential of each model's vector space the theme was clear.

The DTM again proved superior, tending to produce more dense and distinct clusters with labels more closely agreeing with the true labels compared to those of the DIM or LDA model.

We attribute the improved performance of the dynamic models to their richer posteriors. By creating a Markov Chain of variational parameters controlling the document topic proportions and topic word distributions, the dynamic models effectively encode the temporal information latent in the corpus. This in turn produces more detailed topics that prove useful when compared to the "one size fits all" topics generated by traditional LDA.

In conclusion, we have empirically demonstrated the enhanced performance of dynamic topic models at producing coherent topics, and vectorizing documents for document classification and clustering. We therefore recommend the use of these models in situations where the latent temporal information of documents is important and one wishes to produce powerful vector representations for subsequent tasks.

5.2 Future Work

While for the scope of a thesis we believe our evaluation of the dynamic topic models to be fairly comprehensive, several questions remain after this research. Moving forward, these questions posit possible options for future investigation, the first of which would be to quantitatively asses the validity of the document influence estimates produced my the DIM. One might correlate inferred document influence with forward citations, or go further by comparing them with pagerank scores for each document. Subsequent research might pursue the following additional research questions.

- We have observed that the DTM and DIM can outperform traditional LDA at a variety of tasks. How well do these models fare against other traditional document vectorization benchmarks such as the tf-idf and bag-of-words models?
- Additionally, one might test the dynamic models against other algorithms that also encode the temporal information latent in a corpus. How would the DTM fair against similar LDA extensions such as those that monitor the birth and death of topics?
- Finally, it is of practical interest how the performance and training times of these algorithms scale with certain variables. How do factors such as document length, corpus size, and number of classes affect performance?

Appendix A

Appendix Title Here

Bibliography

Blei, D. and J. Lafferty (2009). "Topic Models". In: *Text Mining: Theroy and Applications, Taylor and Francis*.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, pp. 113–120. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143859. URL: http://doi.acm.org/10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

Cai, Deng et al. (2008). "Modeling Hidden Topics on Document Manifold". In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA: ACM, pp. 911–920. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458202. URL: http://doi.acm.org/10.1145/1458082.1458202.

Chang, Jonathan et al. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Neural Information Processing Systems*. Vancouver, BC. URL: docs/nips2009-rtl.pdf.

Gerrish, Sean and David M. Blei (2010). "A Language-based Approach to Measuring Scholarly Impact". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Ed. by Johannes Fürnkranz and Thorsten Joachims. Omnipress, pp. 375–382. URL: http://www.icml2010.org/papers/384.pdf.

Griffiths, T. L. and M. Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.Suppl. 1, pp. 5228–5235.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of Classification* 2.1, pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075. URL: http://dx.doi.org/10.1007/BF01908075.

Khan, J. and G. Bhuyan (2009). "Ocean Energy: Global Technology Development Status". In: *ANNEX I - Review, Exchange and Dissemination of Information on Ocean Energy Systems*. IEA OES. URL: http://www.energybc.ca/cache/tidal/annex_1_doc_t0104-1.pdf.

Kumar, R. V. and K. Raghuveer (2013). "Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation". In: *IAES International Journal of Artificial Intelligence* 1.2, pp. 27–35. URL: http://dx.doi.org/10.11591/ij-ai.v2i1.1186.

Li, Lei and Yimeng Zhang (2010). "An empirical study of text classification using Latent Dirichlet Allocation". In:

Lu, Yue, Qiaozhu Mei, and Chengxiang Zhai (2011). "Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA". In: *Inf. Retr.* 14.2, pp. 178–203. ISSN: 1386-4564. DOI: 10.

1007/s10791-010-9141-9. URL: http://dx.doi.org/10.1007/s10791-010-9141-9.

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai (2007). "Automatic Labeling of Multinomial Topic Models". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: ACM, pp. 490–499. ISBN: 978-1-59593-609-7. DOI: 10.1145/1281192.1281246. URL: http://doi.acm.org/10.1145/1281192.1281246.

Mimno, David et al. (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 262–272. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145462.

Nagwani, N. K. (2015). "Summarizing large text collection using topic modeling and clustering based on MapReduce framework". In: *Journal of Big Data* 2.1, pp. 1–18. ISSN: 2196-1115. DOI: 10.1186/s40537-015-0020-5. URL: http://dx.doi.org/10.1186/s40537-015-0020-5.

Newman, David, Edwin V. Bonilla, and Wray L. Buntine (2011). "Improving Topic Coherence with Regularized Topic Models." In: *NIPS*. Ed. by John Shawe-Taylor et al., pp. 496–504. URL: http://dblp.uni-trier.de/db/conf/nips/nips2011.html#NewmanBB11.

Newman, David et al. (2006). "Analyzing Entities and Topics in News Articles Using Statistical Topic Models". In: *Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics*. ISI'06. San Diego, CA: Springer-Verlag, pp. 93–104. ISBN: 3-540-34478-0, 978-3-540-34478-0. DOI: 10.1007/11760146_9. URL: http://dx.doi.org/10.1007/11760146_9.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: ACM, pp. 399–408. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685324. URL: http://doi.acm.org/10.1145/2684822.2685324.

Rosner, Frank et al. (2014). "Evaluating topic coherence measures". In: *CoRR* abs/1403.6397. URL: http://arxiv.org/abs/1403.6397.

Steyvers, Mark et al. (2004). "Probabilistic Author-topic Models for Information Discovery". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. Seattle, WA, USA: ACM, pp. 306–315. ISBN: 1-58113-888-1. DOI: 10.1145/1014052.1014087. URL: http://doi.acm.org/10.1145/1014052.1014087.

Vinh, Nguyen Xuan, Julien Epps, and James Bailey (2009). "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: ACM, pp. 1073–1080. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553511. URL: http://doi.acm.org/10.1145/1553374.1553511.

Wallach, Hanna M. et al. (2009). "Evaluation Methods for Topic Models". In: *Proceedings of the 26th International Conference on Machine Learning*

(ICML). Ed. by Léon Bottou and Michael Littman. Montreal: Omnipress, pp. 1105–1112.

Wei, Xing and W. Bruce Croft (2006). "LDA-based Document Models for Ad-hoc Retrieval". In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '06. Seattle, Washington, USA: ACM, pp. 178–185. ISBN: 1-59593-369-7. DOI: 10.1145/1148170.1148204. URL: http://doi.acm.org/10.1145/1148170.1148204.

Xie, Pengtao and Eric P. Xing (2013). "Integrating Document Clustering and Topic Modeling". In: *CoRR* abs/1309.6874. URL: http://arxiv.org/abs/1309.6874.

Xu, Wei, Xin Liu, and Yihong Gong (2003). "Document Clustering Based on Non-negative Matrix Factorization". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. SIGIR '03. Toronto, Canada: ACM, pp. 267–273. ISBN: 1-58113-646-3. DOI: 10.1145/860435.860485. URL: http://doi.acm.org/10.1145/860435.860485.