Finding Niche Topics using Semi-Supervised Topic Modeling via Word Embeddings

Gerald Conheady, Derek Greene

School of Computer Science, University College Dublin, Ireland
gerry.conheady@ucdconnect.ie ✉, derek.greene@ucd.ie

Abstract. Topic modeling techniques generally focus on the discovery of the predominant thematic structures in text corpora. In contrast, a *niche topic* is made up of a small number of documents related to a common theme. Such a topic may have so few documents relative to the overall corpus size that it fails to be identified when using standard techniques. This paper proposes a new process, called **Niche**+, for finding these kinds of niche topics. It assumes interactions with a user who can provide a strictly limited level of supervision, which is subsequently employed in semi-supervised matrix factorization. Furthermore, word embeddings are used to provide additional weakly-labeled data. Experimental results show that documents in niche topics can be successfully identified using **Niche**+. These results are further supported via a use case that explores a real-world company email database.

1 Introduction

In certain text corpus exploration tasks, users will be primarily interested in the predominant topics that naturally appear in the data. At other times, users will aim to discover documents related to a selection of topics of particular interest. We define a *niche topic* as a small set of documents from a corpus that the user considers to be linked together by a highly-coherent theme. It is expected that a user can provide example documents and typical words for a niche topic, i.e. a limited level of supervision. An example of this might be a 'data breach' topic, where a user is interested in the discovery of unacceptable leaks of patient data through a health organization's email database. There might be millions of emails in the database and the number of data breaches would be expected to be low, therefore we could naturally regard this as a niche topic within the overall dataset. Ideally a user investigating this data could potentially provide a small sample of emails and terms related to data breaches, so as to help to discover other related content. A second related example might be a 'Product Functionality Query' topic, where a user is interested in the discovery of content from customers who have been querying the functionality of a product.

Unsupervised algorithms, such as Non-negative Matrix Factorization (NMF) [7], have been used to uncover the underlying topical structure in unlabeled text

corpora [1]. NMF might potentially identify a topic associated with a small number of documents, when a very large number of topics is specified. However, this has computational implications and also leads to challenges in interpreting the resulting topic model. Specifically, it is often impractical to ask a user to scan through hundreds or thousands of topics in order to find one or two niche topics. Semi-supervised NMF (SS-NMF) algorithms have been proposed which use background information, in the form of word and document constraints, to guide the factorization process in order to produce more useful topic models [8]. The **Niche**+ process described later in this paper uses SS-NMF techniques.

Word embeddings have been applied in a range of natural language processing tasks, where words are represented by vectors in a vector space [9]. Words with related meanings will tend to be close together in this space. In Section 3 we apply the **Weak**+ approach supervision for topic modeling, [2], which uses word embeddings to generate additional "weakly-labeled" data. This supervision takes the form of a list of candidate words that are semantically related to a small number of "strong" words supplied by an expert to describe a topic.

Our experiments on annotated corpora in Section 4 show that, when this weak supervision is fed to **Niche**+, other example documents from the niche topic can be found. In Section 5 we describe a use case involving a real-world email corpus provided by an enterprise software manufacturer. We show that, by using highly-limited supervision, the **Niche**+ process can successfully identify specific topics of interest from among a larger set of more general topics.

2 Related Work

2.1 Topic Modeling

Topic modeling allows for the discovery of themes in a collection of documents in an unsupervised manner. It differs from keyword searches that try to match documents directly to a particular subset of words or phrases, whereas in topic modeling themes are based on grouping documents that have a similar usage of words. While probabilistic approaches have often been used for topic modeling, approaches based on NMF [7] have also been successful.

Semi-supervised learning often involves using limited labeled data to improve the performance of algorithms which are normally unsupervised. For instance, methods have been proposed for incorporating *constraints* into matrix factorization [8]. For text data, this typically involves providing supervision in the form of constraints imposed on documents and terms, suggested by a human expert who is often referred to as the "oracle". The *Utopian* system, which implements this approach, has demonstrated improved topic modeling results [6].

The best way to provide labeled data for semi-supervised learning is by continuous human interaction with the algorithm [5]. In topic modeling, a key practical challenge is to provide a user with an easy way to explore a large collection of text. The user needs to be free to select from the output of the topic model to

highlight areas for improvement or further analysis. The ability to manipulate both documents and terms in a topic is needed. This approach is used in our experiments where the oracle is asked to provide topic documents and topic words for supervision, in order to guide the topic modeling process. The oracle is also asked to provide feedback on the documents found in the first run to determine whether they belong to the topic or not. This information is used to provide negative supervision during the second run. In other words, the feedback is used to exclude the documents and words from the niche topic.

2.2 Word Embeddings

NMF does not directly take into account semantic associations between words. Related meanings of words, such as between 'car' and 'bus', do not explicitly influence the factorization process. Techniques based on word embeddings attempt to take into account the semantic relatedness between pairs of words, as derived from a large corpus of text. Many applications of word embeddings are based on the use of a neural network as in the original *word2vec* model [9]. The input and output layers have one entry for each word in the vocabulary n. The hidden layer is considered the dimension layer and has d entries. It allows the output from the hidden layer to be represented by a n x d matrix. This representation measures the semantic associations between words in a corpus.

3 Methods

3.1 Characterizing Niche Topics

The characteristics of a niche topic might typically include both its distinctiveness compared to the overall corpus and the *heterogeneity* of the niche. The *distinctiveness* influences how easy it is to find documents in the niche topic and can be measured using the *cosine ratio*. Given a corpus of documents assigned to k topics, where each document is assigned to one topic, we quantify the *cosine ratio* as follows. Firstly, we compute a topic-topic similarity matrix $\mathbf{S}$, where an off-diagonal entry S_{ij} indicates the mean cosine similarity between all pairs of documents in topic i and topic j, and a diagonal entry S_{ii} indicates the mean cosine similarity between all pairs of documents in the same topic i. We refer to S_{ii} as the *within-topic similarity* for topic i. The *between-topic similarity* for topic i is the average of the values S_{ij} where $j \neq i$. The *cosine ratio* for topic i is its within-topic similarity divided by its between-topic similarity. A higher value for this ratio indicates a niche topic which is more coherent and well-separated relative to the rest of the topics present in the corpus.

The heterogeneity of a niche topic can be established by a manual review of the sub-themes of documents within the niche. For instance, sub-themes of a topic such as "sport" could relate to soccer, rugby and tennis. Although clearly part of the "sports" topic, the vocabulary of the documents would be specific to each

sub-theme. In a small group of niche documents, it is expected that the higher the number of sub-themes, the more difficult the documents are to find, as it is more heterogeneous.

3.2 Weak+

The **Weak**+ approach has been proposed to provide a form of limited supervision for topic modeling, where word embeddings are used to generate additional "weakly-labeled" data . The Wikipedia *word2vec* [10] model provides an excellent source of generic semantic relationships of words. However, it cannot fully reflect the idiosyncratic semantic relationships between words within individual subject domains. In order to overcome this limitation, supervision words are first chosen from a *word2vec* model generated from the corpus. These words are only selected if they also appear in the top 500 similar words coming from the Wikipedia *word2vec* model.

3.3 Niche+

We now discuss the **Niche**+ approach for identifying niche topics in a corpus. It uses a semi-supervised strategy based on a simplified version of the Utopian algorithm [6]. The oracle-provided documents and words and the **Weak**+ "weakly-labeled" words are input to **Niche**+ to provide supervision. The relevant notation used for this discussion is summarized in the table below.

Notation	**Description**
m	Number of documents in the corpus
n	Number of words in the corpus
k	User-specified number of topics
$A \in \mathbb{R}^{mxn}$	Document-term matrix
$W \in \mathbb{R}^{mxk}$	Document by topic matrix
$H \in \mathbb{R}^{kxn}$	Topic by word matrix
$W_r \in \mathbb{R}^{mxk}$	Supervision matrix with topic weights for documents in $\mathbf{W}$
$H_r \in \mathbb{R}^{kxn}$	Supervision matrix with topic weights for words in $\mathbf{H}$
$M_W \in \mathbb{R}^{kxk}$	Masking matrix for $\mathbf{W}$ with cells set to 1 for topics supervised
$M_H \in \mathbb{R}^{nxn}$	Masking matrix for $\mathbf{H}$ with cells set to 1 for topics supervised
$D_H \in \mathbb{R}^{nxn}$	Diagonal matrix used for automatic scaling

The Utopian matrix factorization algorithm minimizes the objective in (1):

$$||A - WH||_F^2 + ||(W - W_r)M_W||_F^2 + ||(H - H_r D_H)M_H||_F^2 \quad (1)$$

This requires the $\mathbf{H}$ matrix to be recalculated column by column for every iteration until the stopping criteria is reached. It demands large resources for a corpus with a large vocabulary. As a result a simpler form of the objective was adopted as in (2) which only requires the $\mathbf{H}$ matrix to be updated once per iteration. The diagonal matrix $\mathbf{D_H}$ is no longer required and is eliminated.

$$||A - WH||_F^2 + ||(W - W_r)M_W||_F^2 + ||(H - H_r)M_H||_F^2 \quad (2)$$

The non-negativity constrained least squares with active-set method and column grouping, [4], *nnlsm_ activeset*, is used. The **W** matrix continues to be updated as in (3) per the Utopian algorithm.

$$W \leftarrow \underset{W \geq 0}{argmin} \left\| \begin{bmatrix} H^T \\ M_W \end{bmatrix} W^T - \begin{bmatrix} A^T \\ M_W W_r^T \end{bmatrix} \right\|_F^2 \quad (3)$$

The update process for the **H** matrix is changed as in (4):

$$H \leftarrow \underset{H \geq 0}{argmin} \left\| \begin{bmatrix} W \\ M_W \end{bmatrix} H - \begin{bmatrix} A \\ H_r.M_H \end{bmatrix} \right\|_F^2 \quad (4)$$

The *nnlsm_ activeset* algorithm solves **X** for $min \left\| AX - B \right\|_F^2$ X >=0 element-wise and is used to solve **W** and **H**. Positive supervision is implemented by setting the supervision weights for documents and words to positive values and negative supervision by setting the weights to 0. **Niche+,** building on the original SS-NMF process, is carried out using the following steps:

1. Normalize the document-term matrix using TF-IDF.
2. Reconstitute the documents from the document-term matrix.
3. Generate an extended list of supervision words using the **Weak+** process.
4. Apply the SS-NMF algorithm on the document-term matrix **A**
   (a) Initialise the **W** and **H** matrices with random values. Initialize the $\mathbf{W_r}$, $\mathbf{H_r}$, $\mathbf{M_W}$ and $\mathbf{M_H}$ matrices with zeros.
   (b) Set the $\mathbf{W_r}$ and **W** matrices weights for each document for the topic to be supervised.
   (c) Set the $\mathbf{H_r}$ and **H** matrices weights for each word to be supervised.
   (d) Set the $\mathbf{M_W}$ and $\mathbf{M_H}$ weights for the topic to be supervised.
   (e) Select another topic and set its weights to be the reverse of the supervised topic.
   (f) Repeat until the objective converges or the maximum iteration is reached:
      i. Using *nnlsm_ activeset* solve for **H** and then for **W**.
      ii. Recalculate the objective.

The **Niche+** process guides the discovery of niche topic documents using the words and documents provided by the user, along with the extended list of semantically linked words provided by **Weak+**.

4 Evaluation

4.1 Experimental Setup

The aim of our experiments is to investigate whether the **Niche**+ process can be used successfully to find niche topics. The corpora listed in Table 1 are used for the evaluation. They were chosen as they come with a ground truth and provide niche topics that meet our definition of being a 'small set of documents from the corpus that the users considers to be linked together'.

Table 1. Summary of datasets used in the evaluation. For each topic, we report the within-topic similarity, between-topic similarity, and the ratio for the two.

Dataset	m	k	**Topic**	**Entries**	**Within**	**Between**	**Cosine Ratio**
EU-PR	9,677	12	Antitrust	30	0.14	0.03	4.48
20-NG	18,828	20	Electronics	30	0.05	0.01	4.84
			Politics	30	0.07	0.01	5.75
			Religion	30	0.07	0.01	5.94
			Med	30	0.07	0.01	6.75
Complaints	66,804	90	Bankruptcy	50	0.12	0.03	4.05
			Data Privacy	54	0.09	0.04	2.18
			Adding Money	65	0.12	0.04	3.38

The *EU-PR* dataset consists of press releases describing activities relating to the European Parliament across 12 different policy areas [3], where some policy areas are naturally covered more frequently than others. Our second corpus is the widely-used *20 Newsgroups* (*20-NG*) collection of approximately 18K posts from 20 Usenet newsgroups. The *Complaints* corpus is a collection of over 66K records from the US Consumer Financial Complaints dataset provided by Kaggle, categorized into 90 different types, such as 'Data Privacy', 'Bankruptcy' and 'Foreign Currency Exchange'.

For each corpus, we use the topic(s) with the lowest *cosine ratio* as these are the most difficult to find. Niche topics are created for the *EU-PR* and *20-NG* topics by deleting all the documents for the topics, except the first 30. The 90 *Complaints* dataset is a much larger dataset and its topics have sizes ranging from a single document to over 6000 documents. The 'Adding Money', 'Data Privacy' and 'Bankruptcy' topics are selected as naturally occurring niche topics.

It is important to minimize the user burden of providing labeled documents. To simulate this, only the first five unique documents in each topic are used as oracle documents. We calculate centroid vectors for each annotated ground truth topic, and then rank the corresponding words based on their centroid weights. In this way we can select the top five words as the oracle-given words for the niche. These words are used to construct *word2vec* embeddings on each corpus using

a skip-gram model, with vectors of 100 dimensions and the document frequency threshold set to a minimum of 5.

We next run the **Niche**+ process with step 4 repeated 50 times. We fix the number of topics k to be the number of ground truth topics in each corpus. Weights are set to 10 for the oracle given documents and words. Weights are set to 1 for **Weak**+ generated words, as these are not to be as influential as the oracle given words. The mask weights are set to 1 for the topics supervised. All other weights are set to 0.

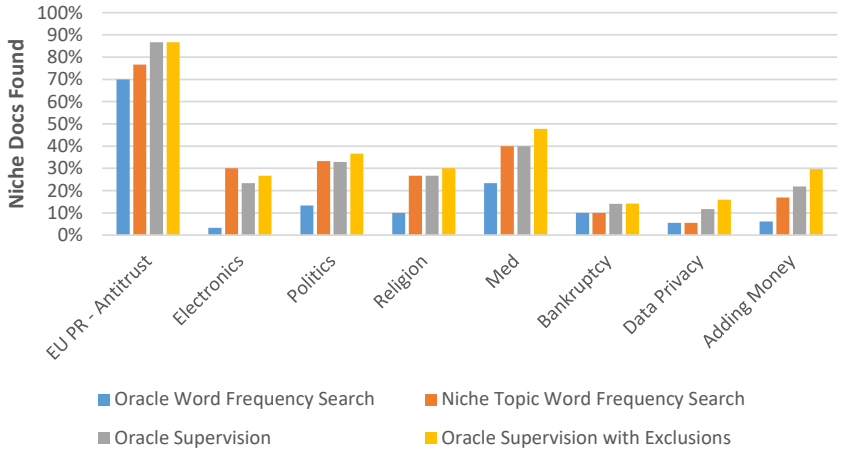
In order to simulate further feedback from the oracle, the initial runs are followed by an 'exclusion' run. Documents found in the original run that are not part of the niche topic are subject to negative supervision by setting their weights to 0. The weights of 5 prominent words from these documents are set to 0 to provide further negative supervision. The percentage of documents found using **Niche**+ is compared to a word frequency search based on the oracle words.

4.2 Results

Firstly, we use Normalized Mutual Information (NMI) [11] to measure the accuracy of document assignments arising from the topic models, relative to the ground truth document assignments. This is done by counting the number of correct documents found for each topic using the ground truth labels. The results show little difference in NMI scores between the runs with and without niche topic supervision. This is explained by the fact that the **Niche**+ process concentrates on improving accuracy for a single topic only. The NMI scores for the *EU-PR* and *20-NG* corpora range from 0.65 to 0.82 and for the *Complaints* corpus from 0.35 to 0.36.

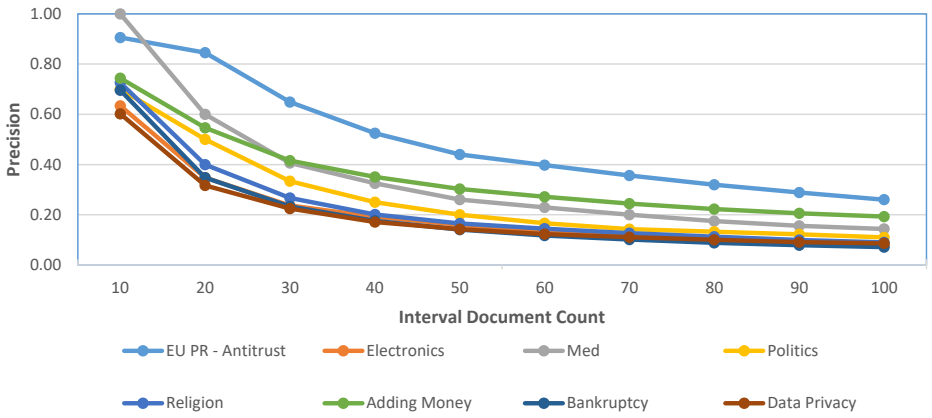
Next we use the percentage of documents found for each niche topic as a measure of success. Weightings by topic for each document are output from **Niche**+. They are ranked into the top 100 documents for each topic. Figure 1 shows the percentage of documents that are labeled as being part of the niche topic. An oracle-based word frequency search is run using the words generated by centroid calculation for the oracle documents. This is all that can be done when there is no ground truth, as in our later use case. The results show 70% of documents are found in the *EU PR* 'Antitrust' topic, from 3% to 23% documents for the *20 Newsgroups* topics and from 6% to 10% for the *Complaints* topics. A 'best case' niche topic word frequency search is run using words generated by a separate centroid calculation for all the documents in the niche topic. This can be done as we have a ground truth. Improved results show that 77% of documents are found in the *EU PR* 'Antitrust' topic, from 27% to 40% of documents for the *20 Newsgroups* topics and from 6% to 17% for the *Complaints* topics. **Niche**+ oracle supervision results are as high as 87% for the 'Antitrust' topic, 40% for the *20 Newsgroups* topics and 22% for the *Complaints* topics. **Niche**+ oracle supervision with exclusions showed further improvement reaching 87% for the 'Antitrust' topic, 48% for the *20 Newsgroups* topics and 30% for the *Complaints* topics.

Fig. 1. Percentage of documents found for niche topics, using different models.



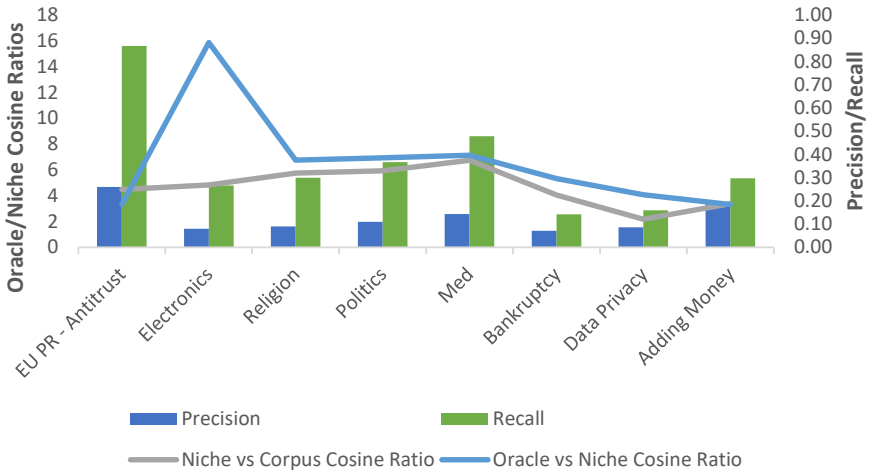
The precision interval analysis in Figure 2 presents the number of niche documents for different levels of precision, considering the top 10 to 100 documents found. We see that typically over 80% of documents are found within the first 30 documents.

Fig. 2. Precision interval analysis for different niche topics.



It is expected that where the oracle documents are similar to the niche topic and the niche topic is distinct from the corpus, that the results will be most successful. i.e. a low oracle to niche *cosine ratio* and a high niche to corpus *cosine ratio* will find more niche documents. These ratios are shown in Figure 3.

Fig. 3. Comparison of cosine ratios vs precision and recall for niche topics.



The "antitrust" topic in the *EU-PR* dataset has the lowest oracle to niche documents *cosine ratio* of 3.3. This is reflected in its high precision and recall scores. The "electronics" topic has the lowest precision and recall scores for the 20ng dataset. Its oracle to niche documents *cosine ratio* is high at 15.8 showing that the oracle documents do not represent the niche well. The highest *20-NG* topic results are for the "med" topic with an oracle to niche documents *cosine ratio* of 7.1. This is slightly higher than the scores of 6.9 and 6.8 for the "religion" and "politics" topics. However, the *cosine ratio* for the niche documents to the corpus documents is 6.8 compared to 5.9 and 5.8 for the "religion" and "politics" topics, implying that the niche topic for "med" is more distinct than the others. A similar pattern is seen in the *Complaints* dataset where the "adding money" topic has the lowest oracle to niche documents ratio and the "bankruptcy" topic the highest. The combination of how well the oracle documents reflect the niche and how distinct the niche is in the corpus determines the success level.

A further manual analysis of the "med" topic reveals a high level of heterogeneity as defined in Section 3.1. It can be divided into distinct sub-themes such as 'back pain', 'lactose intolerance', 'smoking' and others. All the documents are clearly linked to the "med" topic. The oracle documents include one related to the 'lactose intolerance' sub-theme and the results include similar documents. However, the oracle documents do not include any relating to the 'smoking' sub-theme and none of the 'smoking' documents in the niche topic are found. Although the relationship between the sub-themes is easily detected manually, the **Niche**+ process does not make the connection. The "electronics" topic shows a few clear sub-themes such as 'searches for circuits', 'data transmission' and 'car radar'. The oracle documents do not contain any documents relating to these sub-themes and this may explain the poorer performance.

5 Case Study

5.1 Experimental Setup

Enterprise email archives can contain hundreds of millions of emails. The ability to discover niche topics in archives can assist enterprises to audit and manage business processes. A software manufacturer has extracted 279K emails from their email archive for our real-world use case. The emails are unlabeled and cover twelve years of activity from their customer support department. The niche topics provided for analysis relate to a 'visa application', an 'accounting package upgrade' and the 'moving of email archive volumes'.

An initial clean-up of the emails is required. Only the subject and the main body of the email text are used. Details removed include forwarded messages, original messages, confidentiality notices, signatures, URLs ,and email addresses. Only emails with at least 50 characters are selected for the creation of a TF-IDF normalized document-term matrix. Words are filtered based on a minimum document frequency of 30.

The **Weak**+ process generates 95 extra words for supervision from the original user provided words. Table 2 shows the first 10 words generated by user word for the 'accounting package' topic. The words generated are semantically close to the user words in the context of an accounting package upgrade. The words selected for 'quote' relate to seeking and providing of information relating to the cost of the accounting package upgrade. This reflects the user's domain whereas a more generic approach may have interpreted 'quote' as relating to citations from literature.

Niche+ is then run to find 20 topics. This choice is based on an inspection of the data. The supervision weights of the five documents and words supplied by the user are set to 10. The supervision weights of the **Weak**+ generated words are set to 1.

Table 2. Sample of words generated by Weak+

User Word	**Generated Words**
accounting	structuring, budgeting, balances, evaluations, profitability, forecasting, advisory, accountability, competencies, methodology
system	configured, determine, improve, testing, reduce, process, utilizing, application, environment, development
quote	clarifies, mention, explanation, anyway, linked, basically, viewpoint, clearly, clarifying, incorrect
requirement	indicator, workable, disclaimers, consistent, scrutiny, definitions, unacceptable, furthermore, mandates, maintain
upgrade	updating, interface, cost, latest, test, invoicing, platform, automate, storage, service

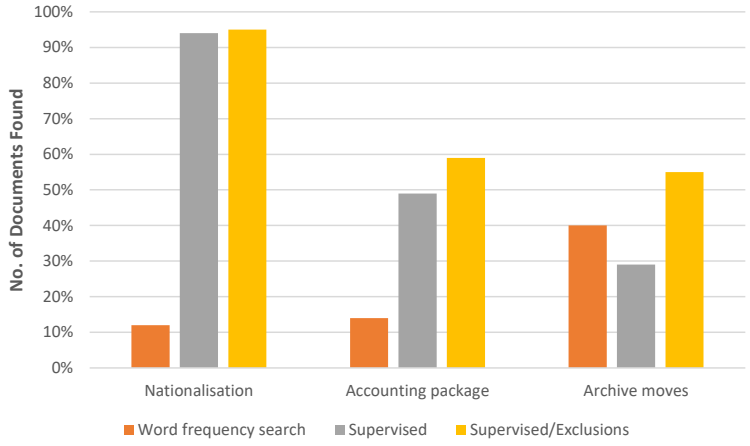
5.2 Discussion of Results

Based on the topic model produced by our approach, the first 100 documents for each topic are ranked based on their weightings as in Section 4.2. The judgment of a user expert, who is familiar with the data, is that 94% of the documents for the 'Visa' topic, 49% for the 'Accounting Package' and 29% for the 'Archive' relate to the topic, as seen in Figure 4. A word frequency search of the email corpus, as used in Section 4.2, with the user given words, results in finding 12% of the documents for the 'Visa' topic and 14% for the 'Accounting Package' and 40% for the 'Archive' topic.

In the case of the 'Accounting Package' topic, many of the off-topic documents relate to other package upgrades, such as Microsoft Windows upgrades. The documents the user excludes from the 'Archive' topic include many relating to a similar product, from a competitor, that is not of user interest. All documents and 5 words identified as not belonging to the topic are used for exclusion runs for each topic, as described in Section 4.1. The number of documents found increases in all supervision/exclusion runs to 95% for the 'Visa' topic, 59% for the 'Accounting Package' and 55% for the 'Archive' topic.

Overall, this use case shows that the **Niche**+ process can successfully find niche topics in real-world datasets, such as a large email corpus.

Fig. 4. Number of documents found for three niche topics in the email corpus, relative to human judgments.



6 Conclusions and Future Work

This paper has shown that input from an oracle (e.g. a "human-in-the-loop") during topic modeling can improve results. In particular, when trying to identify

small niche topics in a large unstructured text corpus, a user's domain expertise can be essential. An initial set of inputs from a user helps the discovery of such niche topic documents. A second round of input, either in the form of inclusions or exclusions, can further improve the results.

It has also been shown that *cosine ratio* is a good predictor of the number of niche documents that are found. This opens up the opportunity to guide users in their selection of suitable documents for niche topic supervision by looking at the *cosine ratios* for the oracle documents.

However, the **Niche+** process is not always successful in finding documents relating to sub-themes in the niche that do not have oracle examples, such as in the case of the 'smoking' sub-theme, as seen in Section 4.2. The process cannot currently reach out to semantically linked sub-themes. This will be an area of further investigation.

References

1. Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models - Going beyond SVD. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 1–10, 2012.
2. Gerald Conheady and Derek Greene. Weak Supervision for Semi-supervised Topic Modeling via Word Embeddings. *Language, Data, and Knowledge. LDK 2017.*, pages 150–155, 2017.
3. J.P. Cross and D. Greene. Capturing and explaining the policy agenda of the european commission between 1986-2016: A quantitative text analysis approach. Under review, 2017.
4. Jingu Kim. nonnegfac-python @ github.com.
5. Patrik Ehrencrona Kjellin. A Survey On Interactivity in Topic Models. 7(4):456–461, 2016.
6. D Kuang, J Choo, and H Park. Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. *Partitional Clustering Algorithms*, pages 1–28, 2015.
7. D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.
8. Tao Li, Chris Ding, and Michael I Jordan. Solving Consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, volume 2, pages 577–582, 2007.
9. Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, 2013.
10. Radim Rehurek. gensim 1.0.0rc1 : Python Package Index.
11. Alexander Strehl and Joydeep Ghosh. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.