Topic Modeling on the Crowd RE Dataset using Word Embeddings

Nicholas Alan Andrew Patrick Ford, Kim Julian Gülle

Technische Universität Berlin
nicholas.ford@campus.tu-berlin.de, kim.j.guelle@campus.tu-berlin.de

Abstract. We aimed to automatically derive topics from a dataset of 2966 requirement sentences. The requirements were previously collected, tagged and categorized by crowd workers as part of the Crowd RE project. We preprocessed the data in an NLP pipeline using state of the art NLP methods and applied topic modeling techniques to the results in order to cluster the data. This includes Latent Dirichclet Allocation, word embeddings using word2vec and the recently published Word Mover's Distance. We visiualized our findings in 2D scatter plots with the help of Principal Component Analysis and Stochastic Neighbor Embedding (t-SNE). All our programming work was done in Python and is strongly dependend on the nltk and gensim library. The requirement sentences had 5 different domains already assigned to them (including a domain called *Other*, which was neither of the other 4). Thus we expected to find 4 different clusters, with some noise between them, caused by the requirements of the *Other* domain. Having followed several approaches for topic modeling we found out that the Word Mover's Distance is probably the most promising, still we were only able to find 3 reasonably distinct clusters. The final verification, whether this means the pre-assigned categories are wrong or the dataset is too small for an automated topic modeling (or a combination of both) is a manual process and may be part of future work.

1 Introduction

In this paper, we aim to automatically analyze the Smart Home requirements collected Murukannaiah et al. for the Crowd RE project[20] in 2016. We will put ourselves in the perspective of a fictitious product owner, who wants to answer the following question:

> *Given a set of requirement sentences, what kind of features are my potential customers interested in the most?*

We consider our product owner to be working in a company which builds smart home appliances and deem the Crowd RE requirements to be the result of a survey that company has performed. The collected requirements therefore are the foundation of our analysis.

Considering the number of requirements (2966), we want to automate our analysis using the Python programming language and a word2vec model to derive a set of categories where the collected requirements can be assigned to. Finally, we want to answer our initial question based on the categories we found and the number of requirements assigned to each of the categories.

2 The Crowd RE Dataset

In an attempt to "facilitate large scale user participation in RE" [20] 609 Amazon Mechanical Turk users[1] were asked to submit requirements for smart home applicances in the Crowd RE project. The result was a dataset containing 2966 requirements, related to the domains *Energy, Entertainment, Health, Safety* and *Other*. The requirements were collected in two phases.
In the first phase the crowd workers were asked for their requirements of a smart home. The phase comprised three stages in which the workers were given a number of requirements and they were asked to add 10 requirements which are distinct to what they have seen. The requirements had to be submit through a form to ensure the requirement sentences follow the user story format[2]. Furhtermore, one of the aforementioned domains had to be selected as the *application domain* of the requirement. Finally, a comma separated list of tags could be added to the requirement. The resulting requirement would then look as follows:

> *"**As a** pet owner, **I want** my smart home to let me know when the dog uses the doggy door, **so that** I can keep track of the pets whereabouts."*[3]

In the second phase, the crowd workers were presented with the requirements produced in the first phase and they were asked to rate the requirements with regard to their clarity, usefulness and novelty. Note that for our analysis though, we only rely on the results of phase one and we mentioned the second phase solely for the sake of completeness.

3 Background

3.1 Natural Language (Pre-)Processing

In order to successfully perform an analysis of the dataset, we first needed to better understand the composition of the data. In a first step we therefore created and analyzed a corpus of requirements. These requirements were formulated as user stories with the defined pattern as described above. Next we compared the results to the Brown Corpus [9], a much larger generic corpus with words taken from books and news articles.

[1] `https://www.mturk.com/`, last visited 2020-01-15

[2] As a [role] I want [feature] so that [benefit].

[3] The keywords marekd in bold text represent the placeholders which were already provided by the form to preserve the user story format.

Indicator	Crowd RE	Brown
Number of Tokens (unique)	90,844 (5,024)	1,034,378
Number of Lexical Words	52,266	542,924
Vocabulary Size (Lexical Words)	4,906	4,6018
Vocabulary Size (Stems)	3,398	29,846
Average Sentence Length (Tokens)	31	18
Average Sentence Length (Lexical Words)	18	10
Lexical Diversity	0.011	0.054

Table 1. Data from the analysis of the Crowd RE dataset

In Table 1 we can see the number of tokens and lexical words is much larger in the Brown dataset which is a result of a wider variety of words in this kind of texts and is also because the brown dataset contains approximately 10 times more lexical words than the Crowd RE dataset. Even though the requirement sentences tend to be much longer, which may have also been caused by the prescribed user story format, the lexical diversity is lower. Requirements use domain-specific expressions, so the same or similar words appear more often in the written requirements[8]. Additionally, the usage of synonyms shall be avoided because it may add ambiguity which is not intended in requirements. In general these results show typical features which are part of a representative dataset that contains only requirements.

In order to derive meaningful data from a dataset, we had to perform some Natural Language Processing (NLP) first, before further analyzing the data. A range of NLP techniques exist, which can be used to prepare the data for our kind of analysis [24][8]. The following list briefly describes the techniques we used in our research:

- **Tokenization** means separating the text into a sequence of tokens. The tokens are simply the single words that are part of the text. With tokenization, whitespaces and all punctuation are removed from the data. As result a list of tokens is generated. The easiest tokenization is just splitting all alphanumeric characters. [24]
- **Stopword-Removal** is removing common words from the data. They are often only required because of grammar or syntax. These words are not necessary to get the meaning of the text. [16]
- **Stemming** is a technique that reduces a word to just the root of the word. It eliminates duplicates that have the same meaning. This is important in NLP as the conjugation of a word is not important. We are just interested in the semantic information that is contained in the words. [16]

– **Bag-of-Words** is a technique that is used to simplify a sentence or document. The idea is to have a list of all containing words with the corresponding word-count in the text. It therefore only holds the word itself and the multiplicity (or in other words the frequency). It is used to have a numerical representation for the words which can be easier processed by a computer. [17]
– **TF-IDF** can be separated into two different indices. TF: term frequency is the rating how often a specific term occurs in the text. IDF: inverse document frequency is a measure how much information a single term provides in relation to a document. [13] The TF-IDF therefore is a rating how valuable a term for a document is which is represented by the formula: $TF\text{-}IDF = TF_{i,j} * log(\frac{N}{n_i})$.

3.2 Latent Dirichlet allocation

The Latent Dirichlet allocation is a technique that can be used to observe groups of similar data within a dataset. The LDA is a probabilistic model that works for discrete data where hidden topics are assumed. The LDA was supposed by Blei et. al in [6]. Within the LDA there are several terms that describe the data. The word is the basic unit of the discrete data. The collection of words is named a document and the set of documents is called corpus. The approach aims to find a limited number of topics that were latent inside of the documents of the corpus. To do so, the documents get *"represented as probability distributions over latent topics where each topic is characterized by a distribution over words"* [21]: The LDA uses all words that are inside of the collection of documents and generates a polynomial distribution over all terms inside of the documents. Afterwards, for each document a Dirichlet distribution is performed which assumes that each document only contains a limited number of topics which is the basic assumption of this approach.

3.3 Word Vectors and Word Embeddings

Being a probabilistic model, an LDA model describes the *"statistical relationship of occurrences rather than real semantic information embedded in words"* [21]. Without considering the semantic relationship between words, the similarity between words cannot be discovered, though [17]. This can result in too broad topics when performing topic modeling using LDA [21]. To overcome this shortcoming, continuous space neural network language models can be trained to capture both the syntactic and the semantic regularities of language. A common defining feature of such models is that each word is converted into high-dimensional real valued vectors (*word vectors*) via learned lookup-tables [19]. A property of these models is that *"similar words are likely to have similar vectors"* [19].

Word2Vec Several architectures for the calculation of word vectors exist (see [17,19] for a more detailed elaboration of these architectures). But according to Mikolov

et al., none of these *"architectures has been successfully trained on more than a few hundred of millions of words"* [17, p1], as they become computationally very expensive with larger data sets (this also applies to the previously mentioned LDA). Therefore, in 2013, Mikolov et al. proposed two optimized neural network architectures for calculating word vectors at a significantly reduced learning time, which allows to train a language model on data sets with billions of words instead: the *continuous bag-of-words model* (CBOW) and the *continuous skip-gram model* [17]. *"The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word"* [17][4]. Both are shallow neural network architectures consisting of an input layer, a projection layer and an output layer [17,22]. Once the language model is trained on any of these architectures, the projection layer holds a dense representation of any word vector, also called *word embedding* [5]. Following this method, Mikolov et al. could not only improve the speed of the learning, but they also found the word embeddings calculated using this method to preserve the syntactic and semantic regularities of the input words given to the neural network [19]. When representing the words in vector space it is then possible, to express these syntactic and semantic similarities by vector offsets, where all pairs of words sharing a particular relation are related by the same constant offset [19]. Figure 1 visualizes these offset-relations in the three-dimensional space. E.g. the Country-Capital plot shows how the word vectors for countries share similar offsets to the word vectors of the belonging capital. This allows to discover relations between words through algebraic operations with their vector representations. E.g. the analogy *Spain is to Madrid as Germany is to Berlin*, could be mathematically represented by the equation $X = vector("Spain") - vector("Madrid") + vector("Germany")$. The word whose vector is closest to X, measured by cosine distance, will be that of *Berlin* [17]. In addition to their initial research, Mikolov et al. created the word2vec open-source project, which incorporates the tool they used to create word embeddings from text corpora based on their promoted neural network architectures CBOW and skip-gram[5].

[4] Consider e.g. the requirements sentence *"As a smart home owner, I want my smart home to..."*. Given the the words ['As', 'a', 'smart', 'owner', 'I'] a CBOW based model would be trained to predict the word *home* as missing. On the other hand, given the word *home*, a skip-gram based model would be trained to predict the words which are most likely to surround the word *home*, so *smart* and *owner*.

[5] `https://code.google.com/archive/p/word2vec/`

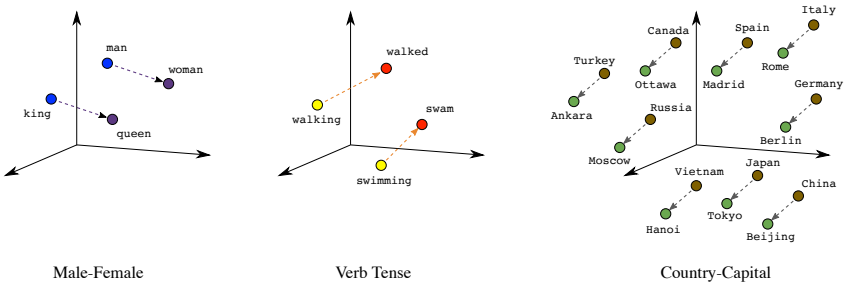


Fig. 1. Word analogies visualized using word embeddings in three-dimensional space [3].

Word Mover's Distance While word2vec is very sophisticated when it comes to generating quaility word embeddings, the method still has its weaknesses. Consider the two documents: *"My smart home should turn on my favorite music when I come to my home."* and *"My smart home shall play my most favored songs when I arrive at my place."* The sentences basically convey the same information. Plotting these sentences with word embeddings, some of their vectors will even be close. Especially if word-wise similarity is given, as e.g. with the pairs $< music, songs >$, $< come, arrive >$. The closeness of the sentences as a whole, though, can not be represented in the word2vec model alone. To overcome this shortage, Kusner et al. introduced the Word Mover's Distance (WMD) in 2015 [12]. The WMD is a distance function which can be used to calculate the distance between these kind of text documents. Based on previously created word embeddings (as for example those from word2vec), the *"distance be- tween[sic!] two text documents A and B is the minimum cumu- lative[sic!] distance that words from document A need to travel to match exactly the point cloud of document B"*[12, p2]. Using this method, the WMD reaches a high retrieval accuracy, while being completely free of hyper-parameters and therefore straight-forward to use.

4 Related Work

Multiple recent works on topic modeling apply the Latent Dirichlet Allocation in order to get an appropriate result for the hidden topics. Zhou et. al in [27] used this technique to automate a part of text mining. They used two different kind of dataset for their research. At first, they focus on articles from Wikipedia where they evaluated over 200,000 articles. They found out that from 50 topics they discovered, there are three topics with high probabilities compared to the others. As second analysis they used a set of twitter messages from 10,000 users. They again found 30 topics containing five topics with the highest probabilities of the

set of topics. As result they mentioned that the processing time of the suggested approach took quite long and might be improved in future works.

Building up a pre-processing pipeline for the topic modeling approach was also performed at several related works. In [10] Gemko et. al proposed a data pre-processing pipeline they used for an automatic glossary term extraction. Their pipeline contains the steps of Tokenization, POS-Tagging, Chunking and Lemmatization. Additionally, they apply some relevance filtering and specificity filtering afterwards. They also used the CrowdRE dataset and got well prepared data from their pre-processing pipeline to work with for their glossary term extraction.

A generally important python library was created in 2010 by Řehůřek et al. They wanted to automatically create a short list of articles similar to a given article [28]. To achieve this, they used Latent Semantic Analysis, as well as LDA in their approach and created a Python library called *gensim*, which aimed at implementing these techniques in a clear, efficient and scalable way [2]. Besides offering implementations of Latent Semantic Analysis and LDA, the *gensim* library also implements the word2vec tool created by Mikolov et al. and was recommended for the generation of word embeddings in a review on NLP toolkits performed in 2018 [24].

Using such word embeddings and the WMD, Qiang et al. accomplished topic modeling over short texts [22]. They also based their work on the findings of [17] and [12], as we do.

5 Proposed Approach

The Crowd RE dataset is available in form of a MySQL database dump, but the tables can also be downloaded separated into several *.csv* files [1]. For our research, we were only interested in the pure requirement sentences (without any ratings, or user characterization added to the data). We therefore reconstructed the sentences from the *requirements.csv* file, which is included in the downloaded data.

To have some measure to evaluate the proposed approach we need a labeling at the dataset that we can use to rate how good the topic modeling worked. At first we checked if we can use the user defined tags as soft labeling for the requirements. Unfortunately most of them are only matched once (Total tags: 2116, tags that only occur once: 1562). Additionally we evaluated the most common used tags and the coverage[6] of the requirements.

[6] specific number of tags covering a significant amount of requirements

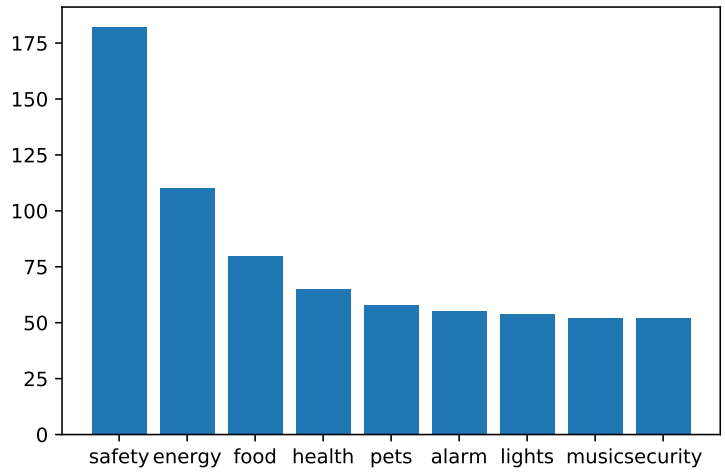


Fig. 2. Tag occurrence and coverage of the requirements

In Figure 2 the coverage of requirements by the given tags is shown. The fact that about $\frac{1562}{2116} \approx 73.8\,\%$ of the tags only occur once leads to a small coverage of requirements by the given tags. The variety of tags that may be assigned to the same topic is very high and the low coverage of requirements with the top 9 tags makes the tags not suitable for the soft labeling.

Another approach to get a labeling for the evaluation was to check the domains that were assigned to the requirements. The domains are separated into five groups: Health, Energy, Entertainment, Safety and Other. For the "Other" there are again user defined specific domains, but we focus on the five top level domains for our labeling.

5.1 NLP Preprocessing Pipeline

As initially described in subsection 3.1 we preprocessed our requirement documents using an NLP pipeline as shown in Figure 3. Implementing our solution in Python and following the common practice as suggested in [8], we made use of the NLTK library [4] to perform the NLP techniques needed for our analysis. As some of the requirements sentences contains special characters, some initial data cleansing was necessary, to remove these special characters (i.e. spaces, dots, apostrophes, slashes) as they would have otherwise been ranked in the later used bag of words. We used regular expressions as provided by the Python standard library in order to do so. For the tokenization, the stop-word-removal and the stemming we used the functions provided by the NLTK API.

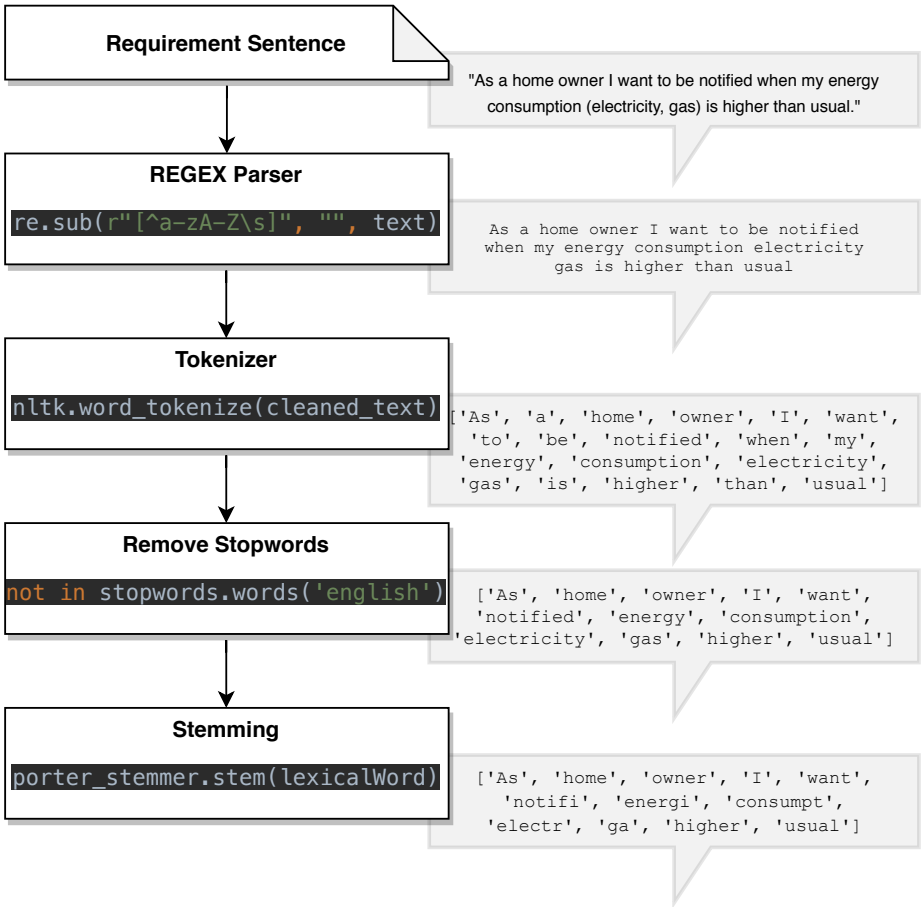


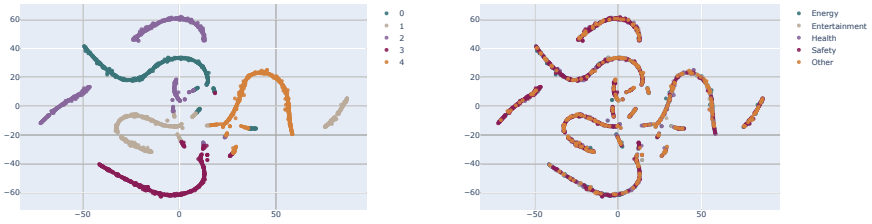
Fig. 3. Processing an exemplary requirement sentence through our NLP Preprocessing Pipeline.

5.2 LDA Approach

After we developed our pre-processing pipeline for the dataset and some basic analysis on the data we have we decided to use the LDA for a first topic modelling. The LDA approach serves as reference for the result we wanted to obtain by the neural network to have a result for evaluating.

For our LDA approach we used our pre-processed requirements. To apply the LDA we transformed the data in the following way. We created a matrix where each row represents one of the 2966 requirements. the columns are the single words of the requirements. But as the LDA needs a numerical representation of the words we first applied a bag-of-words to the single requirements. As the results were not sufficient we decided to calculate the TF-IDF to get weights for

the single words. After these steps we had a prepared matrix that holds the data that now can be used for the LDA.



(a) LDA Plot with colors to topics (b) LDA Plot with colors for domains

Fig. 4. LDA Result with TF-IDF (plotted with t-SNE)

In Figure 4 we can see the two-dimensional representation of the results of the LDA. The reduction of the dimensions is performed by t-SNE which tries to preserve the most differing dimensions. In Subfigure 4(a) the colors are mapped to the found topics which leads to separable clusters. But if we look for the expected mapping to the domains in Subfigure 4(b) we can see that the found cluster doesn't represent the expected clusters that where defined by the domains.

5.3 Word Embeddings

In order for our approach to also respect the semantic regularities of the requirements sentences, it is not enough to rely on the clusters generated by the LDA, as mentioned in subsection 3.3. We therefore create word embeddings for the corpus of requirements sentences we derived from the Crowd RE dataset, using the techniques described earlier.

word2vec As shown in Figure 5, we use the word2vec implementation of the *gensim* library to train language models on both word2vec architectures, with and without pre-processing the requirements through our NLP pipeline before. The outcome are 50-dimensional word vectors which form the basis for our clustering. With our dataset being relatively small, the created vectors may not capture all the semantic regularities, though [18]. We therefore also use the word vectors Mikolov et al. created in order to measure the performance of their word2vec architectures, in 2013 [17]. Instead of the approximately 50.000 words in the Crowd RE dataset, they trained their language model on the Google News dataset with 100 billion words. We thus expect the quality of these word vectors to be much higher and as such to positively affect our later results.

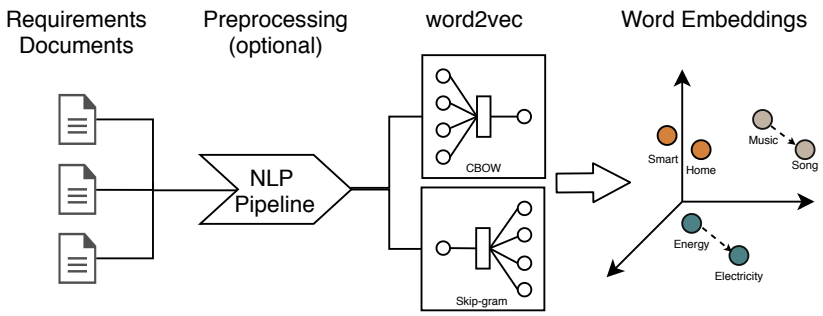


Fig. 5. Training a language model on our requirements corpus to create word embeddings using word2vec.

With the word embeddings, every word in the Crowd RE dataset can be represented as a multidimensional vector. We could now create a matrix, representing the whole vocabulary in word vectors, as shown in Figure 6 (1). Applying K-Means to the matrix, we would then cluster all the words in the dataset into topics. To cluster the dataset on a sentence level instead, we perform the following four steps:

1. We create a matrix for every sentence in the corpus, by replacing each word with its vector representation (see Figure 6 (2)). The x-dimension of the matrix depends on the length of the sentence, the y-dimension is determined by the length of the word vectors. Using our own 50-dimensional word vectors and e.g. a sentence with 10 words, the shape of the matrix will be 10×50.

2. As the sentences in the dataset are of different lengths, the x-dimensions of the matrices are different, too. We use Principal Component Analysis (PCA) [25] to reduce the different x-dimensions to length of the shortest sentence (or the lowest x-dimension respectively).

3. We combine all these sentence matrices in a single matrix T, which is a 3-dimensional matrix of the form $T \in \mathbb{R}^{n \times d \times s}$, where n is the total number of sentences, d is the dimension of the word vectors and s is the length of the shortest sentence in the dataset.

4. To cluster our results with K-Means, we need to further transform this matrix into two-dimensional space. We do this by reshaping the matrix T to a matrix T' with $T' \in \mathbb{R}^{n \times d * s}$[7].

Finally, we cluster the sentences by applying K-Means to the resulting matrix T'.

[7]This is done using the reshape-function of the numpy array implementation: `https://www.w3resource.com/numpy/manipulation/reshape.php`

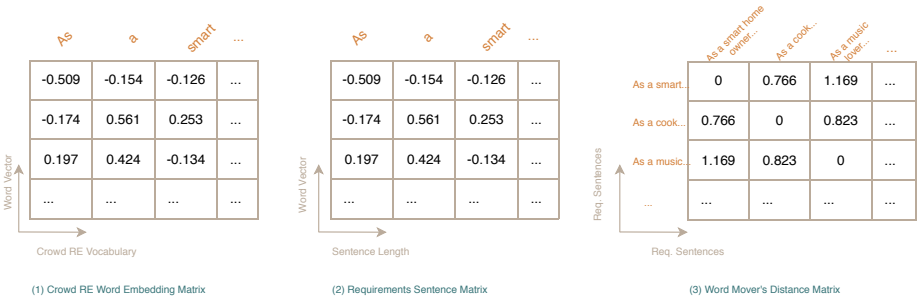


Fig. 6. Representing the Crowd RE dataset using word embeddings.

Word Mover's Distance As mentioned in section 3.3, the document or sentence wise similarity can not be captured fully using word vectors only. In our last approach, we therefore use WMD to again cluster the dataset. To calculate the distance between sentences with WMD, the sentences have to be represented by their word embeddings. Hence, we reuse the word vectors we created with the aforementioned word2vec models. Instead of a word based distance, we now create a distance matrix $D \in \mathbb{R}^{n \times n}$, with n being the total number of sentences, as follows:

1. For every sentence in the Crowd RE corpus, calculate the distances to every other sentence using WMD.
2. Save the distances in the matrix D, as shown in Figure 6 (3).

We then cluster the sentences by applying K-Means to the resulting matrix D. As D is 2-dimensional already, there is no need to reduce its dimensions before.

6 Findings

With each approach we generate topics for the Crowd RE dataset. In accordance with the predefined application domain labels, we expect to find four different topics: *Energy, Entertainment, Health* and *Safety*. Sentences labeled as *Other* are assumed to be visible as noise in the result. In our plots, we always plot the different requirement sentences. We use t-SNE to transform the embeddings into 2-dimensional space for our plots [15]. The coloring is based on the application domain they were associated with and is as follows: *Energy*, *Entertainment*, *Health*, *Safety* and *Other*.

6.1 LDA

Unfortunately, the result of the LDA doesn't match the expected topics. The approach itself creates some separable clusters, they can not be mapped to the

expected ones considering the soft labeled domains. Still, some similarity between the requirements that are plotted next to each other can be found.

The processing time of the LDA approach is very fast which means the performance of is very good compared to approaches with high computational effort. The overall result for the LDA approach is that we couldn't gather the expected topics from the given dataset.

6.2 word2vec & PCA

As shown in Figure 7 we could identify two clusters using the word2vec. Again, this did not match our expected 4 clusters. Any interpretation of the plotted clusters may be only speculative and highly subjective, which is why we did not make any assumptions with regard to the data quality, yet. Similar to the LDA, the performance was relatively good, though and we did not wait for our results for much longer than a couple of minutes.

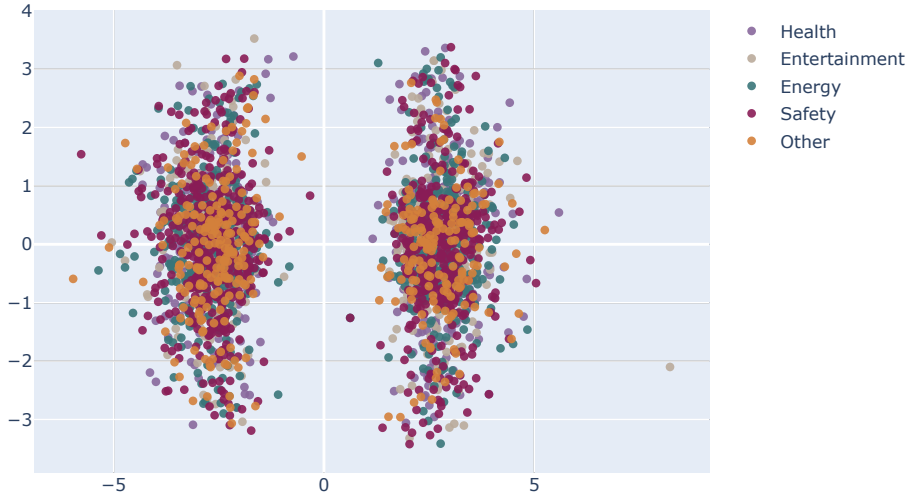


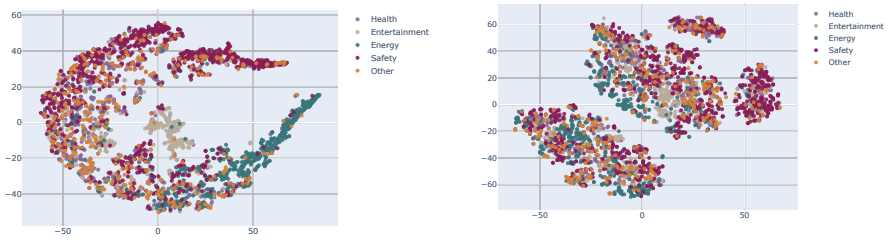
Fig. 7. word2vec result of the clustering using Google News word vectors

6.3 word2vec & Word Mover's Distance

Using Word Mover's Distance, we achieve the best results, in the sense that we can not only distinguish clusters visually (see Figure 8), but also by manual inspection of the sentence plotted next to each other. When looking at the results in Subfigure 8(a), we can see that the domains *Entertainment* (around the center) and *Energy* (stretching from (0,-45) to (10,57)) can be well distinguished.

Also, a cluster predominantly consisting of *Health* sentences becomes apparent in the region from (0,20) to (70,45). Judged by the pre-labeling only, it seems the clustering with our self-trained word vectors worked better. But manual inspection shows that the clustering based on the Google News vectors also brings new insights in the dataset. In Subfigure 8(b) we can see, how the demarcation between the clusters is a lot clearer. Also, even though the sentences seem unrelated at first, the rightmost cluster (the area between (45,-15) and (75,30)) mostly contains sentences related parenting and children. Furthermore, the topmost cluster between (20,45) and (40,65) contains requirement sentences about animals. It becomes apparent, that the dataset may be clustered into different clusters than the 4 domain-based clusters we initially anticipated.

The higher quality of our results comes with a drawback of performance, though. On a current Intel i5-9600K 6-core processor with 3.7 GHz and 32 GB of memory attached, the calculation of the Word Mover's Distance matrix took approx. 45 minutes (even after splitting up the calculation to be done in 12 parallel threads and making use of the symmetry of the WMD).



(a) Self-trained word vectors (b) Google News word vectors

Fig. 8. Results of the Word Mover's Distance

7 Conclusion

To sum up our research we can say that the used approaches work well on large data sets and our dataset is compared to common used ones small. Also the soft labeling of the data that was done by the users themselves doesn't have a good quality. To increase the quality of the labeling manual checking with a lot of effort needs to be performed. The bad quality of the labeling might also result from a missing common understanding of the domains within the crowd workers.

In the end still a lot of manual work needs to be done to derive valuable results from the dataset regarding what features or feature categories are wanted the most in smart home application.

Finally we can summarize that the idea of CrowdRE works to some extend. One can gather a lot of requirements that are assigned to the smart home topic, but it is very difficult to analyze them automatically and requires a lot of effort to prepare them for further work.

8 Discussion

As expected, our dataset was probably too small to achieve any better results. In this context, it is important to know how the accuracy of the word2vec phrase detection dropped to 66% when Mikolov et al. trained their model on a "smaller" dataset of 6 billion words[18, p7]. *Smaller* at least in comparison to their final training set, but this is still a lot larger than our dataset by a factor of almost 120.000.

LDA is a typical topic modeling technique, which is proven to work well on large documents. Short texts instead, contain very limited word co-occurrence information. This hinders the LDA to work well on short texts, as we have seen [23].

With our word2vec approach we have to reduce the dimensions of the sentence matrices using PCA and also to reshape the final matrix from 3 to 2 dimensions. As the PCA provides an approximation of the original data, it cannot be avoided that some data is lost in the process [25]. Our word2vec results may therefore be impacted by the dimensionality reduction.

"The underlying reason is that the document similarity can not be accurately measured under BoW representations due to the extreme sparseness of short texts." [14].

More time would have been needed for the evaluation of our results. Both the tags, as well as the application domains were set by the crowd-workers themselves. The quality of these assignemnts has not been proven yet and we used this data only for lack of proper testing data. For example, one of the requirements with the content "I want my smart home to sync with my biorhythm app and turn on some music that might suit my mood when I arrive home from work so that I can be relaxed" was related to the emphEntertainment domain. In our last approach this sentence was found to be in the *Health* domain using the Word Mover's Distance. We could not say that this assignement was definitely wrong, though. So it could be we find our approach a lot more successful after a thorough analysis of all the clusters.

Finally, we lacked prior knowledge of the field of topic modeling and machine learning in general. Though we performed our research with technical and professional care in all conscience and under consideration of commonly accepted

principles, there may be a lot of potential for further optimization (which goes beyond changing the hyper-parameters of our models). Future works could be done on an improved set of data, by only analyzing those requirements which have clearly defined domains. Such dataset could be achieved, by cleaning the current Crowd RE dataset by the means of manually labeling the requirement sentences. This includes verifying the currently assigned application domains, reassigning some domains and also creating new domains, which may not have been in the domain-selection when the requirements were created.

When manually reviewing the dataset, our results could also be improved through cleaning the dataset. In accordance with [7] and [11] cleaned datasets have a much higher impact on the training results of ML models than the optimization of hyper parameters.

- Li et al. also created a classifier using WMD [14]. Future works could use their approach for a comparison of the generated clusters on the Crowd RE dataset

- *"in general more data (as opposed to simply relevant data) creates better embeddings."* [12]

- The calculation of the WMD matrix was relatively time consuming. Wu et al. propose a different distance measure for document clustering, which compared to the WMD *"can achieve much lower time complexity with the same accuracy."* [26].

Acronyms

CBOW	Continuous bag-of-words
CSV	Comma Separated Value
LDA	Latent Dirichlet allocation
ML	Machine Learning
PCA	Principal Component Analysis
RE	Requirements Engineering
TF-IDF	Term Frequency, Inverse Document Frequency
WMD	Word Mover's Distance

References

1. Crowd RE Dataset, https://crowdre.github.io/murukannaiah-smarthome-requirements-dataset/, last visited: 2020-01-15
2. Gensim Python library, https://radimrehurek.com/gensim/index.html, last visited: 2020-01-19
3. Machine Learning Crash Course | Embeddings: Translating to a Lower-Dimensional Space, https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space, last visited: 2020-02-24
4. Nltk library, https://www.nltk.org/, last visited: 2020-01-18
5. Word embeddings | TensorFlow Core, https://www.tensorflow.org/tutorials/text/word_embeddings, last visited: 2020-02-24
6. Blei, D.M.: Latent Dirichlet Allocation p. 30
7. Chu, X., Ilyas, I.F., Krishnan, S., Wang, J.: Data cleaning: Overview and emerging challenges. In: Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16. pp. 2201–2206. ACM Press, http://dl.acm.org/citation.cfm?doid=2882903.2912574
8. Ferrari, A.: Natural language requirements processing: from research to practice. In: Proceedings of the 40th International Conference on Software Engineering Companion Proceeedings. pp. 536–537. ACM Press (2018), http://dl.acm.org/citation.cfm?doid=3183440.3183467
9. Francis, W.N.: A standard corpus of edited present-day american english 26(4), 267–273 (1965), https://www.jstor.org/stable/373638
10. Gemkow, T., Conzelmann, M., Hartig, K., Vogelsang, A.: Automatic Glossary Term Extraction from Large-Scale Requirements Specifications. In: 2018 IEEE 26th International Requirements Engineering Conference (RE). pp. 412–417. IEEE, Banff, AB (Aug 2018), https://ieeexplore.ieee.org/document/8491159/
11. Krishnan, S., Wang, J.: Data cleaning: A statistical perspective - overview and challenges part 2, https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxkYXRhY2xlYW5pbmd0dXRvcmlhbHNpZ21vZDE2fGd4OjJhMzc4ZWExM2U3MzA3MGE, ACM SIGMOD/PODS Conference
12. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. pp. 957–966. ICML'15, JMLR.org
13. Leskovec, J., Rajaraman, A., Ullman, J.D.: Data Mining. In: Mining of Massive Datasets (2014)
14. Li, C., Ouyang, J., Li, X.: Classifying extremely short texts by exploiting semantic centroids in word mover's distance space. In: The World Wide Web Conference. pp. 939–949. WWW '19, Association for Computing Machinery, https://doi.org/10.1145/3308558.3313397
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE 9, 2579–2605, http://www.jmlr.org/papers/v9/vandermaaten08a.html
16. Mhatre, M., Phondekar, D., Kadam, P., Chawathe, A., Ghag, K.: Dimensionality Reduction for Sentiment Analysis using Pre-processing Techniques. In: Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC). pp. 16–21 (2017)
17. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient estimation of word representations in vector space. pp. 1–12

18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality 26
19. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. Association for Computational Linguistics, https://www.aclweb.org/anthology/N13-1090
20. Murukannaiah, P.K., Ajmeri, N., Singh, M.P.: Toward automating crowd RE. In: 2017 IEEE 25th International Requirements Engineering Conference (RE). pp. 512–515. IEEE, http://ieeexplore.ieee.org/document/8049175/
21. Niu, L.Q., Dai, X.Y.: Topic2vec: Learning distributed representations of topics http://arxiv.org/abs/1506.08422
22. Qiang, J., Chen, P., Wang, T., Wu, X.: Topic modeling over short texts by incorporating word embeddings http://arxiv.org/abs/1609.08496
23. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation pp. 2270–2276
24. Solangi, Y.A., Solangi, Z.A., Aarain, S., Abro, A., Mallah, G.A., Shah, A.: Review on natural language processing (NLP) and its toolkits for opinion mining and sentiment analysis. In: 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS). pp. 1–4. ISSN: null
25. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis 2, 37–52 (1987)
26. Wu, X., Li, H.: Topic mover's distance based document classification. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT). pp. 1998–2002 (2017), ISSN: 2576-7828
27. Zhou Tong, H.Z.: A TEXT MINING RESEARCH BASED ON lda TOPIC MODELLING. Computer Science & Information Technology (CS & IT) pp. 201–210 (2016)
28. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: New Challenges for NLP Frameworks. ELRA, http://is.muni.cz/publication/884893/en