Topic Modeling on the Crowd RE Dataset using Unsupervised Machine Learning

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. Hier kommt eine kurze Zusammenfassung der Arbeit.

1 Introduction

ToDo: Find & add references
The result of a good software is depends on a lot more than the pure engineering. Software is designed to perform and automate a task, which otherwise would have to be executed manually. In order to fulfill the given task properly, it is necessary to completely understand the task's underlying problem. One strategic approach to develop this understanding is to conduct a thorough analysis of the requirements (of both the business and the software) and to document the results according to given standards. This process is called requirements engineering (RE). The work products of the RE process can be manifold. In the scope of this research study we will therefore focus on the actual requirements.

One part of the RE process is to collect, order and prioritize the requirements. Depending on the number of requirements, it may help or even be necessary to use some data analysis techniques to automatically derive some useful insights, which can then be used in the further decision-making process. In this paper, we will have a look at the requirements which were previously collected by ... in the Crowd RE project. From the perspective of a fictitious product owner we want to answer the following question:

> *Given a set of requirements, what kind of features are our 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 this company has performed. The collected requirements therefore are the foundation of our analysis.
Considering the sheer size of the requirements, we don't want to go through each of the requirements manually, but want to setup and use an unsupervised neural network to perform the analysis. We will use the Python programming language to do so and want to derive a set of categories where the collected requirements fall into.

Finally, we want to answer our initial question based on the categories we found and the number of requirements assigned to each of them.

2 The CrowdRE Dataset

- An approach towards scaling the RE process through the engagement of the general public.

- Necessary to use automated techniques to gain useful insights

- The dataset contains RE for smart home appliances

- Reqs were submit by 609 Amazon Mechanical Turk users (https://www.mturk.com/)

- Furthermore, the personal characteristics of the crowd workers who supplied the RE are recorded "including their demographics, personality, traits, and creative potential", as gathered in a presurvey (questions, Mini-IPIP scale, CPS)

- It was an attempt to "facilitate large scale user participation in RE" [2]

2.1 Challenges / Motivation / Benefits

Crowd RE made it possible to gather a large amount of data

Raw data is of little use, but to derive information from the data manually may be difficult and is error prone, e.g. when looking at the sheer amount of information gatherted

Also, human effort is a cost factor and the time is better spent on tasks which can not be automated, yet

We, as the authors, can be very happy to base our research on the Crowd RE dataset, as it is quite cumbersome to curate data which can be used to train and test automated techniques

Authors of the Crowd RE reqs already tagged their reqs into the domains Energy, Entertainment, Helath, Safety, Other -¿ Could be used for verification

3 Natural Language Processing

ToDo: What is NLP and what are the essential steps?

3.1 Word2Vect

ToDo: How does Word2Vect work and what is the advantage?

ToDo: How is categorization usually performed?

3.2 Latent Dirichlet Allocation

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 idea was to have another approach in the first step that we can use as intermediate result for the data and also to compare it to the result of the neural network to have some kind of benchmark or basis for a performance comparison. The LDA is a probabilistic model that can be used for discrete data. It is a statistical approach that can be used to generate a topic model for text corpora [1]. The technique starts with selection a number of expected topics. The LDA then use all terms 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 amount of topics. Target of the approach is to get the latent topcis that are are core of the document collection.

3.3 Unsupervised Neural Networks

ToDo: What kind of NN used to perform categorization?

ToDo: How does it work?

ToDo: What kind of preprocessing is necessary?

4 Related Works

ToDo: Add references to papers that have a similar approach...

5 Analysis

The Crowd RE dataset is available in form of a MySQL database dump. As an alternative, the tables can also be downloaded separated into several *.csv* files. As we aim to analyze the datasat using unsupervised learning techniques, we were only interested in the pure requirements (without any ratings, classifications or other user characterization added to the data whatsover). Using the *.csv* files was therefore sufficient for us and the following analysis is based on the reconstructed requirements that were generated from with the template. The template was given as follows: As a [role] I want [feature] so that [benefit]. The missing parts are taken from the file *requirements.csv* as they were entered into the template.

5.1 NLP

To compare the data we choose the Brown Dataset which is a huge dataset that was taken from books and articles.

Indicator	Crowd RE	Brown
Number of Tokens (unique)	90844 (5024)	1034378
Number of Lexical Words	52266	542924
Vocabulary Size (Lexical Words)	4906	46018
Vocabulary Size (Stems)	3398	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 bigger in the Brown dataset as the diversity of words is higher in this kind of text. But it's also because the dataset is much bigger than the Crowd RE dataset. Also obvious is that the sentence length at the requirements is lower as they tend to be formulated short and simple. Additionally the lexical diversity is also less as often the same words are used to write down requirements. And it is also necessary to use unique words for the description of the same feature to avoid ambiguity. To sum up we can say that the results are as expected from a dataset that contains only requirements.

5.2 Preprocessing

- Bag of words - Remove speical characters (spaces, dots, apostrophes, slashes), because otherwise they would have been ranked in the bag of words

6 Our approach

ToDo: Give the approach a name as title!

6.1 NLP Pipeline

ToDo: How does our NLP Pipeline look like and why

6.2 NLP Pipeline

ToDo: How does our approach with the Neural Network looks like?

7 Findings

ToDo: Our results

7.1 LDA

ToDo: How good does the LDA perform?

7.2 Neural Network

ToDo: How good does our approach with the Neural Network perform?

7.3 Comparison

ToDo: Compare the results of the both methods

ToDo: Finally, our initial questions can be answered as follows:

ToDo: Discuss the results (maybe a new chapter for that?

Acronyms

CSV	Comma Separated Value
LDA	Latent Dirichlet allocation
RE	Requirements Engineering

References

1. Blei, D.M.: Latent Dirichlet Allocation p. 30
2. 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/