SK[AI] IS THE LIMIT: MISSING MANUALS

Srikanth Gurram Anastasia Sycheva Sai Yaswanth

ETH Zurich, Switzerland

ABSTRACT

Which manuals are missing on the list? Try to perform a clustering of the E-Mails, so that new manual categories can be identified and the Helpdesk team can create such how-to manuals.

Introduction

In our training data set, for more than 1800 emails there was no associated manual. This indicates a simultaneously a problem and a huge room for improvement of the customer experience. Our notebook aims to guide the new manual creation process.

Unfortunately due to time constraints we were not able to try all the approaches that we had in mind. We have explored the two most promising: hierarchical clustering and lda topic modelling

Data preparation

To save some time, we have used the text preparation pipeline provided in the baseline. To better serve out task we have extended the number of stop words that are filtered out. Experimenting with different sets of **POS** tags could be an interesting improvement however we could not invest too much time into it.

Text representation

Our goal at this stage is to represent an email text by a fixed dimensional vector. To achieve this we use **tf-idf** vectorizer. Other approaches we have considered include: doc2vec and various ways of aggregating word embeddings of words in sentences (i.e. sum or average).

Dimensionality reduction techniques such as PCA, t-sne and umap are useful tools to examine the validity of our representation.

Hierarchical clustering

We use the text representations from the previous section to perform hierarchical clustering of more than 1800 emails. We us a dendogram to select the optimal number of clusters. We get intuition about the resulting clusters with the help of word clouds. Histograms of other categorical variables might contain other valuable queues about the potential issue, its impact and severity

LDA topic modelling

A potential disadvantage of our previous approach is the fact that it implicitly assumes that 1 email is dedicated to 1 topic only. To overcome it, we perform the LDA topic modelling on all emails without manuals. The number of topics is selected using the

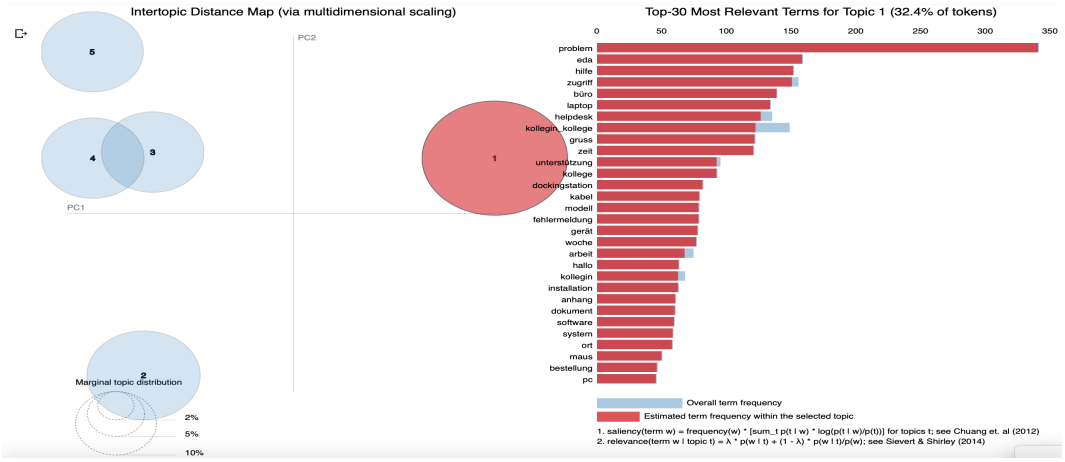

Fig. 1. Topic visualization for LDA