SK[AI] IS THE LIMIT: SERVICE MERGING STRATEGY

Srikanth Gurram Anastasia Sycheva Sai Yaswanth

ETH Zurich, Switzerland

ABSTRACT

Find a solution for the evaluation of the best possible services to merge with each other into a single service. Try to come up with ready-to-use code that benchmarks different combinations of 2 or more services and show the resulting differences. Try to run automatic tests of one or more merges in parallel.

1. DATA PREPARATION

- Clean and lemmatize the subject - removing the prefixes and alphanumeric strings.
- Clean and lemmatize the body.
- Filter the mails based on language.

2. TEXT REPRESENTATION

First we concatenate all emails that correspond to certain topic together. Following that we use SPACY package to represent every email as a fixed dimensional vector (dim = 300). We cross check the validity of our representation with the help of PCA plots (umap and tsne could also serve the purpose)

An alternative approach would be to cluster the vector representation of all individual email separately and use the labels in the resulting cluster to find which services tend to overlap the most.

Doc2Vec and tf-idf are two other valid approaches for fixed size text representation that we did not time to experiment with.

3. CLUSTERING

We perform hierarchiclal clustering[1] on the resulting dataset with 55 entries. Our dendogram suggests that the emails could be represented by 15-17 differen.

Due to time constraints we did not have time to experiment with different linkage criteria and distance measures. Our inference suggests that services belonging to one cluster could be grouped together.

[1] we have also tried MeanShift and DBSCAN algorithms however the results were not very intuitive

4. BENCHMARKING

Our best performing classification model was a BERT based classification model initialized with pretrained weights. We use $dbmdz/bert-base-german-uncased$ *by Hugging Face as our pretrained weights.*

We use the clusters to merge the classes and train a Bert model to benchmark the new merged class labels.