21/11/20

Team Ekoln
Moritz Reihs
Luca Entremont
Isha Gupta

Task 3

We have used a Latent Dirichlet Allocation (LDA) algorithm to define the most meaningful list of manuals. LDA is a topic modelling algorithm which is very well suited to our case, as it quickly and deterministically clusters emails with similar content. It is also specifically aligned to the notion of ‘topics’ within the language of an email, which is exactly the basis on which we want to partition and create the manuals.

An LDA works by grouping words within a word group into an overarching ‘topic’, and then analysing the occurrence of words in each email to create connections between topics and emails. It is given a list of all word that appear across all emails and performs its analysis based on

The LDA algorithm has to be told how many manuals to form. We tested various numbers of manuals by looking at the resulting groups and checking which was of the highest quality. This turned out to be … .

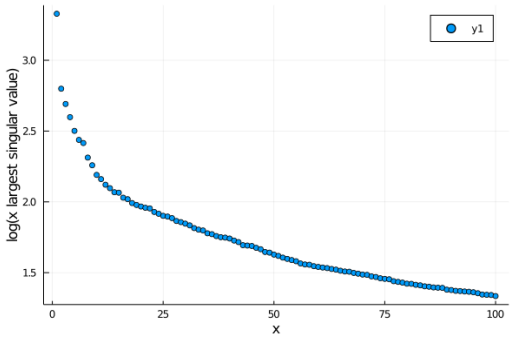
The results shows that the optimal list of manuals is as follows:

[…]

———————————————————————————————————————

An alternative approach to the clustering that we came up with is a singular value decomposition (SVD) analysis. This was based purely on the extraction of a single feature, namely which words appear both in the title and the the main body of the text.

We then plotted the eigenvalues that we obtained on a logarithmic scale and could extract the first 30-ish most significant to represent the most important manuals.



By then analysing the third factor of the decomposition (V) and comparing the first 30 columns, we could match each email to a manual.

Thereby we were also able to identify which manual receives the most emails.

21/11/20

Team Ekoln
Moritz Reihs
Luca Entremont
Isha Gupta