

AnTeDe Lab 2

1 Notebook B

Using more Data to train the model increased the stability of the classifier. It knows more words connected to the cities, and can therefore classify a wider range of sentences. In the sentence 'there is no lake.' 'lake' is the only valid token. Due to the occurrences of 'lake' in the training set, the probability of lake occurring in a sentence is highest if the sentence is of class geneva.

The sentence 'It is the city of Zwingli' was wrongfully classified as being connected to bern. After a closer look it can be seen that only the word city is used by the classifier. Given that the word city occurs only once for each bern and zurich in training, the greater prior probability of bern is probably the reason for the misclassification. The additional training document labeled to zurich increased the prior probability of zurich. Now the posterior probability of zurich given the word city is the dominant factor.

2 Notebook C

When the stopwords aren't removed in the vectorizer, but TF IDF features are used the performance doesn't drop as much as without the features. This is caused by the TF IDF features which penalise frequent terms like stop words.

In the confusion matrix we can see which on which classes the classifier works good, and on which ones it has problems. Often the misclassified test documents are classified as a similar class than the true one. So we can conclude that the classifying by with a BoW approach can indeed do a decent job in classifying a topic of a document, but to differentiate between subtopics in a general topic it has its difficulties. This conclusion is to be expected since BoW classifiers don't include the context which is important to find the small differences between subtopics.

The proposed SGD classifier had a 5% better accuracy than the MNB classifier. The macro precision was slightly lower, however the recall and F1 score improved. We also tried some other classifiers from sklearn like the support vector machine classifier or the decision tree classifier. Using sklearn's standard settings none of these came up to the performance of the SGD classifier. The support vector machine classifier came closest with only a 1% accuracy drop.