

AnTeDe Lab 2

1 Notebook A

2 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. The sentence 'There is no lake' is still misclassified due to not understanding the negations.

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.

3 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.