Mid term progress

Axel Hirschel, 10656146

Table of contents (and what is in it)

Research questions

- Which classifiers perform best on classifying Dutch municipality documents after a domain adaptation from the Dutch parliament?
    - How can the classifiers be optimized to perform on the parliamentary data?
        - What thresholds are optimal in detecting topics of multi-labeled documents?
        - How can classifiers be optimized to deal with the large in-class variety of the documents?
    - How well do the classifiers adapt to a new domain?
        - How well do the algorithms perform on the new domain?
        - What characteristics of misclassified documents per classifiers?

Preliminary results*

Parliamentary data:

Algorithm	F1	Recall	Precision	Accuracy
Logistic Regression	0.77	0.72	0.84	0.54
Naive Bayes	0.09	0.05	0.98	0.06
Random Forest	0.45	0.30	0.92	0.27
SVM	0.79	0.72	0.84	0.55
CNN - normal	0.67	0.67	0.67	0.37
CNN - multiple filters	0.74	0.75	0.72	0.46
CNN - split up documents	0.74	0.80	0.69	0.44
Par2Vec with Logistic Regression	0.66	0.59	0.74	0.38
CNN - k-max pooling	NNB	NNB	NNB	NNB

Municipalities data:

Algorithm	F1	Recall	Precision	Accuracy
Naive Bayes	0.04	0.03	0.06	0.16
Logistic Regression	0.25	0.20	0.33	0.30
Random Forest	0.23	0.45	0.15	0.02
SVM	0.19	0.13	0.43	0.34
CNN - normal	0.23	0.25	0.21	0.35
CNN - multiple filters	0.19	0.22	0.17	0.11
CNN - split up documents	NNB	NNB	NNB	NNB
Par2Vec with Logistic Regression	0.17	0.27	0.13	0.25
CNN - k-max pooling	NNB	NNB	NNB	NNB

*Best F1 score is chosen per algorithm

Future weeks

14 - 20 May	Finish CNN - k-max pooling Test CNN with k-max pooling Re-write Introduction, Methodology and related work Create overview of all versions that need to be tested
21 - 27 May	Start testing Begin with results and conclusion Tweak algorithm to perform better on domain adaptation
28 May - 3 June	continue testing continue with results and conclusion Tweak algorithm to perform better on domain adaptation
4 - 10 June	Finish results Finish conclusion and future work **7 June:** Finish first draft thesis
11 - 17 June	Rewrite based on feedback Spell-check Add last observations **14 June:** Finish thesis
18 - 24 June	Write Blog for OSF (less technical) Practise Presentation **21 June:** Presentation
25 June - 1 July	Finish blog Finish up at OSF