

Wikification of news

Mart Busger Op Vollenbroek  
Olivier Louwaars S2814714  
Project Text Analysis  
M. Nissim  
mr. L. Kloppenburg

# Mart

# Olivier

The moment Mart and I started this project, we approached it as a team. Being in one class for ¾ year now, we both know what we can do ourselves and what the other is capable of too. Being on similar levels of programming, the dividing off the tasks was not really necessary. Each one of us would take a turn to write some code, while the other was watching and thinking about the problems ahead. Taking turns resulted in a program we both support and understand, and we would have (eventually) reached similar programs if we would have to do it on our own. Although the exercises for PTA prepared us a bit for the final assignment, it still was hard to combine all the small programs. Mostly due to the strict demands in formatting and order, using unordered data containers as sets and dictionaries was not ideal. From time to time we felt like we missed the knowledge, which probably has cost us a lot of time and workarounds.

During the final project, I focused mainly at the machine learning part. During the course information retrieval, we already learned how classifiers work which helped me here. As the classes are not commonly used in other classifiers, we had to start from scratch and build our own. Knowing the Stanford NER tagger is one of the best available, training it on our data had the best perspective. Using a online tutorial, it was fairly easy to do so, and it seems to classify with great performance. Of course this is only on the given train data, so it still has to proof itself on the real data. A lot hard was the wikifying itself, after the important entities were recognized. The method we eventually used was to look the entity up in WordNet for a definition, and then scanning the Wikipedia API suggestions for the same words. Although this seems a logical approach, the machine thinks different and comes up with very strange URL’s sometimes (still always related to the original entity). It will always be hard for a machine to achieve the same results as humans because of the lack of context that makes human decide to link to a certain page.