UNIVERSITY OF BIRMINGHAM

Ontology Construction

Csongor Barabasi – 1636980

1. Outlining the Task

The task was to automatically classify the seminar announcement emails according to an ontology.

2. The Ontology Tree

By scanning through the training emails I constructed an ontology tree by hand, which looks as following:

- → science
    - chemistry
    - engineering
        - computer
            - vision
            - robotics
            - software
            - interaction
        - electronics
    - mathematics
    - biology
        - environment
        - medicine
    - physics
    - psychology
- → business
    - marketing
    - career
        - graduate
        - employment

3. The Algorithm

First step, I deleted all the stopwords (and, the, in etc.) from all the training emails. Then, using the TF-IDF (Term Frequency, Inverse Document Frequency) algorithm *(En.wikipedia.org,2017,Tf-idf)* I extracted the 8000 most frequent words which appear in the training data and saved them to a text file, each word in a separate line.

Next, I looped through all the emails which have to be classified. First I extract the most important words of the current email, by first looping through my word database and getting those which also appear in the email. I add to these words every word from the *'Topic'* line from the header and the 10 most used words in the email body. Then I go through the ontology tree, while I reach a leaf. During the process I check whether the current ontology node word appears in the text or not. If it does, I proceed further on that branch. If not, then I consider all the ontology words on the current level and choose the one which has the highest similarity with the most important words from the email. Similarity is computed using **word2vec** trained on the *GoogleNews* dataset.

4. Improvements

Further improvements could be made to this algorithm. First would be to train my **word2vec** model on a dataset which is closer to the vocabulary used in the emails. Next major improvement would be to work on the ontology tree, to make it more detailed. Lastly, since it is an unsupervised algorithm, I would take another approach to the classification if I would have more time. I would try using k-Means clustering to cluster the emails based on their most important words chosen by the **TF-IDF** algorithm, and then try to assign to each cluster the right ontology word.

Reference:

1. En.wikipedia.org. (2017). Tf-idf. Available at: https://en.wikipedia.org/wiki/Tf%E2%80%93idf