Università des Saarlandes
WiSe 2016/17
Computational Linguistics

This folder contains a small library.

Modules

clean.py

The script stores the `clean(doc)` function: given a string, removes punctuation and stopwords, lemmatizes the words, returns the string.

clean_with_tags.py

The script stores the `clean_with_tags(doc)` function: given a string, removes punctuation and stopwords, lemmatizes the words, concatenates to each lemma the tag that was most often associated with it in the InScript CAKE annotations, returns the string. The best tag for each lemma is retrieved from the word_best_tag dictionary created from the script 'tags-word_best_tag.py'.

lda_model.py

The script stores two main functions:

- `save_lda_with_tags(doc_complete,n_topics)`: given a list of raw documents, it cleans them with the clean(doc) function, and runs the lda model with iterations 100????? . It saves the dictionary, corpus and lda_model to file.
- `save_lda_with_tags(doc_complete,n_topics)`: given a list of raw documents, it cleans them with the clean_with_tags(doc) function, and runs the lda model with iterations ????????? . It saves the dictionary, corpus and lda_model with tags to file.

tags_eval.py

The script stores the function `save_tags_eval(lda_filename,num_topics,results_filename)`: given an lda_model filename, the number of topics it was trained with, and a filename to store the results, it saves the 3 tags with the highest counts per topic (ordered from the most to the least frequent).

tags-word_best_tag.py

The script was used to create 2 files:

- **tags** is a list of all tags used for the CAKE scenario in the InScript corpus
- **word_best_tag** is a dictionary that stores the best tag from the CAKE scenario associated with each word (if any, because not all words were tagged)

Francesca Capuano