GÖTEBORGS UNIVERSITET

Språk-BANKEN

Topic modeling and named entity recognition

Leif-Jöran Olsson

Språkbanken, University of Gothenburg

2014-02-14

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Overview

- Topic modeling
- Named entity recognition
- EXPath app packages

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Topic modeling

- Basically putting labels on texts (or verifying them)
- Mallet a machine learning for language toolkit library for statistical natural language processing, document classification, clustering, topic modeling, and information extraction.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

A popular usecase for topic modeling

- Give me texts like the one(s) I have selected.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Named entity recognition (NER)

- Basically putting labels on words in texts (you see it is similar)
- Stanford NLP a group of statistical natural language processing libraries for tokenization, part-of-speech tagging, named entity recognition, classification, parsing, and coreference.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Stanford NER

- 3 classes: Person, organization, location
- A general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models (CRFClassifier).
- By training your own models, you can build sequence models for any task.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Mallet can do NER too

- Includes implementations of several classification algorithms, including Naïve Bayes, Maximum Entropy, and Decision Trees.
- Can also evalute classifiers.
- Can also be done with Finite state transducers (FST).

GÖTEBORGS UNIVERSITET

Språk-BANKEN

The EXPath apps

- Available on github and in public repo.
- Show examples

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Final Comments

- Many ways to do it
- Different level of languages support
- Licensing
-