MSc in Data Science

Natural Language Analytics

Academic Year: 2017-2018

Exercise 3: Topic Modelling & NER

Delivery Date: **19/06/2018**

Question A: (80%)

This question concerns the modelling of data in topics. We will be using some papers from the NIPS (Neural Information Processing Systems) conference. NIPS is a machine learning conference.

You can download the data from Sam Roweis' website:
https://cs.nyu.edu/~roweis/data/nips12raw_str602.tgz

- **Natural Language Processing (NLP) (10%)** - Introducing basic text processing methods such as tokenisation, stop word removal, stemming and vectorising text via term frequencies (TF) as well as the inverse document frequencies (TF-IDF)
- **Topic Modelling with LDA (70%)** – Experimenting with the Latent Dirichlet Allocation (LDA). Experiment with the different parameters (number of topics, alpha parameter, etc.). Explore the dataset by finding for example the most representative document for each topic and the topic distribution across documents.

Question B: (20%)

This question concerns the Named Entity Recognition task. We will be using the Groningen Meaning Bank dataset. Download the 2.2.0 version of the corpus here: Groningen Meaning Bank Download.

GMB is composed of a lot of files, but we only care about the .tags files.

A file contains more sentences, which are separated by 2 newline characters. For every sentence, every word is separated by 1 newline character. For every word, each annotation is separated by a tab character.

The tags used are the following:

geo = Geographical Entity

org = Organization

per = Person

gpe = Geopolitical Entity

tim = Time indicator

art = Artifact

eve = Event

nat = Natural Phenomenon

There are also some subcategories, focus only on the main ones above, so you can remove them.

- **Named Entity Recognition (NER)** - Extract the Named Entities from the dataset and use the annotated data to learn a supervised classifier (using the IOB representation) for named-entity recognition and classification. Train your model using the Naïve Bayes classifier.

Helpful resources:

The following package provide some indicative, topic modeling packages:

1. NERC (Python):
    a. https://nlpforhackers.io/named-entity-extraction/
2. Topic modeling (Python):
    a. https://radimrehurek.com/gensim/index.html
    b. http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html
    c. https://nlpforhackers.io/topic-modeling/
3. Topic visualisation: http://pyldavis.readthedocs.io/en/latest/