Introduction to Natural Language Processing and Data Mining

Workshop @ ISMW

Instructor: **Dmitry Ustalov**

IMM UB RAS & UrFU (Yekaterinburg, Russia)
ITMO University (Saint Petersburg, Russia)

Outline

- Introduction
- Data Preprocessing
- Text Clustering
- Topic Modeling
- Conclusion

Introduction

- You have probably attended the previous lecture.
- Group task presentations are planned for Friday.
- Here, we will consider a few practical aspects:
  - data preprocessing,
  - text clustering,
  - topic modeling.
- Software:
  - Weka: http://www.cs.waikato.ac.nz/~ml/weka/
  - MALLET: http://mallet.cs.umass.edu/

Introduction (cont.)

- I have created a small text corpus for this workshop.
  - Wikipedia is used (CC BY-SA).
- The corpus *contains* 3 topics, each topic has 10 texts:
  - food,
  - music,
  - Nokia.
- Please, download it.

http://ustalov.imm.uran.ru/pub/ismw-topics.tar.gz

Data Preprocessing

Data Preprocessing

- In practice, textual data are noisy.
- Hence, it is necessary to filter out some content:
  - spelling issues,
  - unnecessary parts-of-speech,
  - low frequency terms,
  - encoding issues.
- You may do it programmatically.
  - Tokenization, stemming, tf-idf, etc.
- But you do not have to.

Text Clustering

Text Clustering

- Weka is an acronym for Waikato Environments for Knowledge Analysis.
  - http://www.cs.waikato.ac.nz/~ml/weka/
- It is written in Java, contains various machine learning algorithms, and is often used in NLP tasks.
- We will apply Weka for text clustering, but let's explore the downloaded corpus.

Attribute-Relation File Format

- Weka uses the ARFF format, ARFF = CSV + header.

```
@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,Iris-
versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
```

…

Text Clustering with Weka (cont.)

- Run Weka and choose the Explorer option.
- Open the *ismw-topics.arff* file.
- Use the *unsupervised.attribute.StringToNominal* filter to convert the last attribute (*topic*) to the nominal variable.
- Use the *unsupervised.attribute.StringToWordVector* filter to create a vector space from the texts.
- Play with clustering, attribute selection and (optionally) classification.

Text Clustering with Weka (cont.)

- Remove the existent *topic* attribute.
  - It was included for demonstration purposes only.
- Use the *unsupervised.attribute.AddCluster* filter to provide texts with clusters.
  - Do not forget to configure this filter!
- Use the supervised.attribute.AttributeSelection to refine the attributes by some criterion.
  - NB: the "Undo" button.
- Save the resulted dataset in any file format that Weka supports.

Topic Modeling

Topic Modeling

- MALLET is a Java-based toolkit for topic modeling.
  - http://mallet.cs.umass.edu/
- The following processing steps:
  - provide it with the texts,
  - build an internal representation of them,
  - train a topic model.

Topic Modeling with MALLET

- Importing:

```
bin/mallet import-dir --input <path>
--output <file>.mallet --keep-sequence
--remove-stopwords
```

- Training:

```
bin/mallet train-topics --input
<file>.mallet --num-iterations 2000
--num-topics 20 --output-state
topic-state.gz --output-topic-keys
topic-keys.txt --output-doc-topics
doc-topics.txt
```

Conclusion

- Text preprocessing is often required to remove unnecessary words or data.
- Weka is good for data exploration and testing various algorithms on your data.
  - http://www.cs.waikato.ac.nz/~ml/weka/
- MALLET is a powerful tool for topic modeling.
  - http://mallet.cs.umass.edu/
- Plan before do.
  - Do not be afraid of trying the new.

Questions?

Dmitry Ustalov, https://ustalov.name/en/.

Feel free to contact me regarding this course.

• dau+ismw@imm.uran.ru

Thank you!

These slides will be uploaded to the ISMW website soon.