<table>
<thead>
<tr>
<th>Course 9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text Mining</strong></td>
</tr>
</tbody>
</table>

**Program**

1. **Introduction**
   - a) Examples of Natural Language Processing (NLP) applications
   - b) Current topics in text mining
   - c) An overview of NLP and text mining frameworks
   - d) Basic terminology and concepts
   - e) High-level topic overviews
   - f) Statistical issues and the main problems
   - g) A quick refresher: Bayesian statistics
   - h) The chain rule of probability
   - i) The Markov property
   - j) Practical

2. **Language Modeling**
   - a) Power laws: Zipf’s law
   - b) Stochastic processes
   - c) Probabilistic language models
   - d) N-gram probabilities
   - e) Sparsity and zero probabilities
   - f) Basic smoothing techniques
   - g) Advanced smoothing techniques
   - h) Model evaluation: entropy and perplexity
   - i) Practical

3. **String Processing**
   - a) Documentation extraction
   - b) Tokenization: lexer and scanner
   - c) Sentence segmentation
   - d) Part-of-speech tagging
   - e) Stemming vs. lemmatization
   - f) String metrics & matching
   - g) Locality sensitive hashing
   - h) Document similarity measures
   - i) TF-IDF details
   - j) Practical

4. **Text Classification**
   - a) Latent semantic analysis
   - b) Text classification approaches
   - c) Multinomial naive Bayes classifier
   - d) Maximum entropy (logistic regression) classifier
   - e) Sentiment analysis
   - f) Point-wise mutual information
   - g) Evaluation metrics: set scores
   - h) Practical

5. **Information Extraction**
   - a) Probabilistic graphical models
b) Markov random field  
c) Latent Dirichlet allocation  
d) Dynamic Markov models  
e) Sequence tagging: HMM, MEMM, and CRF  
f) Entity recognition and annotation  
g) Interaction extraction  
h) Practical

Bibliography

  [http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers](http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers)
  [http://gnosis.cx/TPiP/](http://gnosis.cx/TPiP/)

Prerequisites

Knowledge of math and statistics will be helpful, but only an understanding of Bayes’ rule is required.

Because many participants lack a strong computational background, GATE ([https://gate.ac.uk/](https://gate.ac.uk/)), a Java text mining and information extraction workbench, will be used as the working environment during the practicals. It has a graphical interface, requires no prior programming experience, and it is easy to use the Standford CoreNLP library from it ([http://nlp.stanford.edu/software/corenlp.shtml](http://nlp.stanford.edu/software/corenlp.shtml)). Prior Java programming exposure therefore is helpful, but not required.

Python experts can instead apply the practical examples using a mix of tools, including the NLTK, which has wrappers for Stanford CoreNLP. IPython Notebooks of the examples can be provided, but there is not enough time to discuss them or provide assistance for the Python implementations, too, so this path is only suggested for true “Pythonistas”.

Preparations

The practical sessions will use GATE, which should be installed from [https://gate.ac.uk/download/](https://gate.ac.uk/download/) before the course. The latest Stanford CoreNLP ZIP package should also be at least downloaded from [http://nlp.stanford.edu/software/corenlp.shtml](http://nlp.stanford.edu/software/corenlp.shtml). An excellent tutorial to quickly pick up some of the basics of using GATE is available at [https://gate.ac.uk/wiki/quick-start/](https://gate.ac.uk/wiki/quick-start/) and the GATE website has an extensive User and Developer guide. It is strongly suggested to work through the mentioned GATE tutorial before the course.