THE EXAM

UNIVERSITY OF ESSEX
Postgraduate Examinations 2017

TEXT ANALYTICS

Time allowed: TWO hours

The paper consists of THREE questions.

Candidates must answer ALL questions.

The questions are NOT of equal weight.

The percentages shown in brackets provide an indication of the proportion of the total marks for the PAPER which will be allocated.

Please do not leave your seat unless you are given permission by an invigilator. Do not communicate in any way with any other candidate in the examination room.

Do not open the question paper until told to do so.

All answers must be written in the answer book(s) provided. All rough work must be written in the answer book(s) provided. A line should be drawn through any rough work to indicate to the examiner that it is not part of the work to be marked.

The use of pocket calculators is allowed. All the work of the examination, including any rough work, must be done in book(s) provided and you have been told you may have.
THE QUESTIONS

• Each question worth about 30%
• The typical question:
  – A general part worth 5-6 points asking you to provide a definition (e.g., ‘what is named entity recognition’) / examples of applications of a TA technique
  – A more theoretical question (‘how is PMI defined?’)
  – A question in which you have to demonstrate you understand the concepts by putting it into practice

TOPICS COVERED IN THE MODULE

• Preprocessing
• Clustering
• Text categorization
• Sentiment analysis
• Named entity recognition
• Wikipedia and Disambiguation to Wikipedia
• Coreference
• Relation extraction
• Social Media
• Summarization
CLUSTERING

• What is clustering
• Methods:
  – The vector space model
  – K-Means, Agglomerative clustering
  – Topic models

WHAT KIND OF QUESTIONS YOU MAY EXPECT

• Examples of general questions:
  – What are the uses of text clustering?
• Examples of theoretical questions:
  – Explain the K-means algorithm
  – What is TF/IDF
• Examples of practical questions:
  – Show how clustering could be used in a practical application
  – Apply some method (e.g., agglomerative clustering) in a toy example (e.g., the question gives you a few vectors)
TEXT CLASSIFICATION

• What is text classification
• Methods
  – Naïve Bayes vs Decision Trees (also: various other algorithms in the labs)
  – Feature selection (PMI, Information Gain ... )
• Applications
  – Text categorization
  – Spam detection
WHAT KIND OF QUESTIONS YOU MAY EXPECT

• General questions: what is text classification?
• Examples of theoretical question:
  – Explain how decision tree are learned
  – Explain what is information gain
• Practical question
  – Show how text classification could be used in a practical application
  – Apply some method in a toy example

---

A REAL EXAMPLE

---

Question 1:
Text classification techniques are used in a variety of applications.

(a) Explain what is meant by text classification and give three examples of applications of such methods. [5%]

(b) Bayesian methods are a popular approach to text classification.

(i) Explain the basis of the Bayesian approach to text classification and briefly outline the main advantages over alternative approaches. [5%]

(ii) Explain the main assumptions of a Naive Bayes classifier. Why is it called "Naive"? [5%]

(c) Imagine you are asked to develop an automated email sorter for a company. The company wants to make a distinction between three types of emails: technical (about companies), financial, and the rest ("irrelevant"). You decide to use a Naive Bayes approach.

(iii) What pre-processing steps (before running the Naive Bayes Classifier) do you consider applying to the input text? [4%]

(iv) Assume that on average 50% of the emails are technical, 40% are financial, and 10% are of no interest. Assume further you have collected the following statistics about the posterior probability of categories given the three classes, where 0.00 stands for some very small value.

<table>
<thead>
<tr>
<th>Categorization</th>
<th>Technical</th>
<th>Financial</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rent</td>
<td>0.01</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Doc</td>
<td>0.00</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>CRM</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>HR</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sales</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Business</td>
<td>0.02</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>Projects</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
SENTIMENT ANALYSIS

• What is sentiment analysis
• Approaches:
  – A form of supervised text classification / Lexically-based
  – Levels of granularity
• Resources
  – Sentiment lexica, etc

WHAT KIND OF QUESTIONS YOU MAY EXPECT

• Examples of general questions:
  – Give two examples of use of sentiment analysis
• Examples of theoretical questions:
  – Explain what is aspect-based sentiment analysis
  – Give examples of words with positive and negative polarity
• Examples of practical questions:
  – Apply text classification techniques to develop a sentiment analyzer for a particular company / product
NAMED ENTITY RECOGNITION

- The supervised approach to NER
  - NER as a classification problem
  - The IOB format
  - Sequence labelling
  - Features most used
- The semi-supervised approach
- Distant learning
WHAT KIND OF QUESTIONS YOU MAY EXPECT

• Examples of general questions:
  – What is NER and what issues you have to deal with

• Examples of theoretical questions:
  – Explain how the IOB notation works
  – What kind of machine learning approach works best with NER and why

• Examples of practical questions:
  – Explain how you could develop a NER system for a non-standard domain

COREFERENCE

• Issues in coreference
  – E.g., main problems with identifying mentions

• Theoretical ideas:
  – The Vilain et al evaluation algorithm
  – The Soon et al algorithm

• Applications
WHAT KIND OF QUESTIONS YOU MAY EXPECT

• Examples of general questions:
  – What kind of problems does a coreference resolver encounter?

• Examples of theoretical questions:
  – Explain (some aspect of) the Soon et al algorithm
  – How is precision computed in the Vilain et al evaluation algorithm

• Examples of practical questions:
  – Show how the Soon et al algorithm could be used in a practical application
  – Compute the MUC score in a toy example

A REAL EXAMPLE

Question 2

Resolving intra-document and inter-document coreference is a key aspect of text mining. The algorithm proposed in [2001] by Soon et al. has been very influential on subsequent work on coreference and it is still used as a baseline for modern systems.

(a) One of the most influential features of the algorithm is the way it characterizes coreference resolution as a classification problem. Briefly explain the notion of a learning instance used in the algorithm.

(b) Suppose you have the following annotated piece of text for training a coreference model using the Soon et al. algorithm, where square brackets are used to indicate entities and <span>subscript indices</span> are used to indicate the annotated antecedent:

(The algorithm proposed in [2001] by Soon et al. has been very influential in subsequent work on coreference and it is still used as a baseline for modern systems.)

Explain how the encoding algorithm proposed by Soon et al. works and what it will do in this example, i.e., which positive and negative training instances the algorithm would generate for precedent [1].

(c) Given the same annotated text as in part (b),

(1) explain the encoding scheme used by the algorithm, i.e., in which order it would tag potential antecedents [7%]

(2) what is your hypothesis about the training instances the algorithm would generate for precedent [2]? [7%]

(d) Briefly discuss limitations of the algorithm and the set of features proposed by Soon et al. Can you think of any improvements? [7%]
SOCIAL MEDIA

• Text analytics with social media
  – Types of social media
  – Issues that arise
• Content-based analysis
• User-based analysis

WHAT KIND OF QUESTIONS YOU MAY EXPECT

• Examples of general questions:
  – What kind of issues arise when applying text analytics methods to social media?
• Examples of theoretical questions:
  – How to do user geolocation using Twitter info
  – Event extraction from social media
• Examples of practical questions:
  – Suppose a company asks you to set up a framework for doing event extraction using Twitter. What types of issues you have to deal with?