Overview

1. Introduction
   1.1 What is Text Mining
   1.2 Documents and Document Collections
   1.3 Representation of Documents and Features
   1.4 Basic Text Mining Tasks
   1.5 Advanced Text Mining Tasks

2. Document Classification
3. Clustering of Documents
4. Information Extraction
5. Discovery of Patterns and Trends
1.1 What is Text Mining

Text mining (TM) seeks to extract useful information from a collection of documents. It is similar to data mining (DM), but the data sources are unstructured or semi-structured documents.

The TM methods involve:
- Basic pre-processing / TM operations, such as identification / extraction of representative features (this can be done in several phases)
- Advanced text mining operations, involving identification of complex patterns (e.g. relationships between previously identified concepts)

TM exploits techniques / methodologies from data mining, machine learning, information retrieval, corpus-based computational linguistics
1.2 Documents and Document Collections

**Document collection** is a grouping of text-based documents. It can be either *static* or *dynamic* (growing over time).

**Document** is a unit of discrete textual data within a collection, representing usually some real world document, such as, a business report, memorandum, email, research paper, news story etc. A document can be a member of different document collections (e.g. *legal affairs* and *computing equipment*, if it falls under both).

1.2 Document Collection PubMed

An example of a real-world document collection: PubMed is the National Library of Medicin’s on-line repository of citation-related information for biomedical research papers. This on-line service contains abstracts for > 12 million research papers. About 40,000 new abstracts are added every month.

Search for a particular document:
- Keyword search is not very useful, as protein or gene returns 2,800,000 documents.
- Even a more specific term — epidermal growth factor receptor — returned 10,000 documents.

Identification of relationships across documents across documents:
- Manual attempts are labour-intensive / impossible to achieve.
- Automatic methods enhance the speed / efficiency of research activities.
1.2 Document Structure

Text documents can be:
- **unstructured**, i.e. free-style text (but from a linguistic perspective they are really structured objects)
- **weakly structured** adhering to some pre-specified format, like most scientific papers, business reports, legal memoranda, news stories etc.
- **semistructured** exploiting heavy document templating or style sheets.

1.3 Document Representation and Features

Irregular and implicitly structured representation is transformed into an explicitly structured representation.

We can distinguish:
- **feature based representation**, 
- **relational representation**.

In feature based representation that documents are represented by a set of features.
1.3 Document Representation and Features

Examples of some commonly used features:

- **Characters**
  enabling to recognize e.g. morphological features.
  So called bigrams (trigrams) represent sequences of 2 (or 3) characters.

- **Words**
  Often the term word-level tokens is used instead.
  Tokens can be annotated (e.g. with labels representing noun, verb etc.).
  Bag-of-words representation exploits words, but the order is ignored.
  Word stem represents a group of related words stripped of a suffix

- **Terms**
  may represent single words or multiword units, such as “White House”

---

Examples of some commonly used features (continued):

- **Concepts**
  For example, the concept identifier “car” can represent different words in the text, such as automobile, car, sport-car etc.
  Concepts are useful to represent synonyms and help to resolve polysemy.
1.3 Problem of High Dimensionality

Structured representations of natural language documents leads usually to very large number of features. For instance, one small collection of Reuters of 15,000 documents contains 25,000 non-trivial features (word stems). Some algorithms do not deal very well with large numbers of features and hence it is necessary to employ feature reduction techniques. Another problem is feature sparcity: Each document contains only a small number of all potential features.

1.4 Basic Text Mining Tasks

- Document classification (categorization)
- Information Retrieval
- Clustering / organization of documents
- Information extraction

More information to follow
1.4 Information Retrieval

Retrieval of documents in response to a “query document”
(as a special case, the query document can consist of a few keywords)

![Diagram of Information Retrieval]

1.4 Document Classification

Classification of documents into predefined categories (classes)

![Diagram of Document Classification]
1.4 Clustering / Organizing Documents

Unsupervised process through which documents are classified into groups called clusters.

1.4 Information Extraction

IE involves identification of certain entities in the text, their extraction and representation in a pre-specified format (e.g. a table).

<table>
<thead>
<tr>
<th>Price</th>
<th>Type</th>
<th>Location</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>119 700</td>
<td>T5</td>
<td>Gaia</td>
<td>70</td>
</tr>
<tr>
<td>132.180</td>
<td>T4</td>
<td>Loulé</td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

T5 Duplex em Gaia
Data: 2002-05-10 15:01:24 PST
Excelente localização no centro da cidade. 2 WC, despensa, terasa com marquise com 70 m2; 119700 euros; Tel. 966969663


Output: Filled in Template / Table
1.5 Advanced Text Mining Tasks

- Concept co-occurrence
  - Quantification of co-occurrence
- Identification of trends in data
  - Identification of new topics
- Summarization

1.5 Concept Co-occurrence

Detection of concept **co-occurrence** in documents, e.g.:

- Disease – Medical Drug (based on BioWorld articles)
  - Rheumatoid Arthritis - Rituximab
  - Rheumatoid Arthritis - Infliximab
  - Prostate Carriroma - APC8015 Vaccine
  - etc.

**Quantification of frequency of co-occurrence**

can be expressed in numerical form, or using a graphical representation (e.g. a circle graph; the width of the line indicates the strength of the connection)
1.5 Identification of Trends in Data

Identification of trends in data
How does the news concerning a particular disease (e.g. Rheum. Arthritis) and a particular medical drug (e.g. Rituximab) change over time?
How does the news concerning a particular company and a particular product (e.g. medical drug Rituximab) change over time?

Identification of new topics in the data
Did any new articles appear concerning certain type of company (e.g. a pharmaceutical company) and a particular type of product (e.g. A medical drug useful for treating lung cancer)?

Identification of disappearing topics in the data
Identification of a period covered by a certain topic

1.5 Summarization

Summarization of a single document
Selection of some sentences, summarizing the document

Summarization of several documents
Selection of a single representative document

Selection of representative sentences from different documents