Data Mining I

Text Mining
Outline

1. What is Text Mining?
2. Text Preprocessing
3. Feature Creation
4. Feature Selection
5. Pattern Discovery
Motivation for Text Mining

Approximately 90% of the world’s data is held in unstructured formats.

Examples:
- web pages
- emails
- customer complaint letters
- corporate documents
- scientific papers
- books in digital libraries

Source: Oracle Corporation
Text Mining

The extraction of implicit, previously unknown and potentially useful information from a large amount of textual resources.
Some Text Mining Applications

1. Classification of news stories or web pages
2. Email and news filtering / SPAM detection
3. Sentiment analysis
4. Clustering documents or web pages
5. Search term auto-completion
6. Information extraction
Mixture of Document Clustering and Classification
Sentiment Analysis

- The basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level.

- Polarity values
  - positive, neutral, negative
  - likert scale (1 to 10)

- Application examples
  - Document level
    - analysis of tweets about politicians
  - Feature/aspect level
    - analysis of product reviews
Search Log Mining

- Analysis of search queries issued by large user communities.

- Applications
  1. Search term auto-completion using association analysis
  2. Query topic detection using classification
Information Extraction

- Information extraction is the task of automatically extracting structured information from unstructured or semi-structured documents.

- Subtasks

  1. Named Entity Recognition and Disambiguation
     - “The parliament in Berlin has decided …“
     - Which parliament? Which Berlin?

  2. Relationship Extraction
     - PERSON works for ORGANIZATION
     - PERSON located in LOCATION

  3. Fact Extraction
     - CITY has population NUMBER
     - COMPANY has turnover NUMBER [Unit]
Search versus Discovery

- **Search/Query** (Goal-oriented)
  - Query Processing
  - Information Retrieval

- **Discovery** (Opportunistic)
  - Data Mining
  - Text Mining

Structured Data

Text
The Text Mining Process

1. Text Preprocessing
   - Syntactic and/or semantic analysis

2. Feature Generation
   - Bag of words

3. Feature Selection
   - Reduce large number

4. Data Mining
   - Clustering
   - Classification
   - Association analysis
2. Text Preprocessing

1. Tokenization
2. Stopword Removal
3. Stemming
4. POS Tagging
Syntactic and Linguistic Text Preprocessing

- **Simple Syntactic Processing**
  - Text Cleanup (remove punctuation and HTML tags)
  - Tokenization (break text into single words or N-grams)

- **Advanced Linguistic Processing**
  - Word Sense Disambiguation
    - Determine which sense a word is having.
    - Normalize synonyms (United States, USA, US)
    - Normalize pronouns (he, she, it)
  - Part Of Speech (POS) Tagging
    - Parse sentences according to grammar
    - Determine function of each term
    - e.g. John (noun) gave (verb) the (det) ball (noun).
Stopword Removal

- Many of the most frequently used words in English are likely to be **useless** for text mining.
- These words are called **Stopwords**.
  - examples: the, of, and, to, an, is, that, …
  - typically text contains about 400 to 500 such words
  - for an application, an additional domain specific stopwords list may be constructed
- Why should we remove stopwords?
  - Reduce data set size
    - stopwords account for 20-30% of total word count
  - Improve effectivity of text mining methods
    - stopwords may confuse the mining algorithm
More Examples of Stopwords

a about above across after again against all almost alone along already also although always am among an and another any anybody anyone anything anywhere are area areas aren't around as ask asked asking asks at away b back backed backing backs be became because become becomes been before began behind being beings below best better between big both but by c came can cannot can't case cases certain certainly clear clearly come could couldn't d did didn't differ different differently do does doesn't doing done don't down downed downing downs during e each early either end ended ending ends enough even evenly ever everybody everyone everything everywhere f face faces fact facts far felt few find finds first for four from full fully further furthered furthering furtherings furthers g gave general generally get gets give given gives go going good goods got greater greatest group grouped grouping groups h had hadn't has hasn't have haven't having he he'd he'll here here's hers herself he's high higher highest him himself his how however how's i i'd if i'll i'm important in interest interested interesting interests into is isn't it its it's itself i've j just k keep keeps kind knew know known knows l large largely last later latest least less let lets let's like likely long longer longest m made make making man many may me member members men might more most mostly mr mrs much must mustn't my myself n necessary need needed needing needs never new newer newest next no nobody non noone nor not nothing now nowhere number numbers o of off often old older oldest on once one only open opened opening opens or order ordered ordering orders other others ought our ours ourselves out over own p part parted parting parts per perhaps place places point pointed pointing points possible present presented presenting presents problem problems put puts q quite r rather really right room rooms s said same saw say says second seconds see seem seemed seeming seems sees several shall shan't she she'd she'll she's should shouldn't show showed showing shows side sides since small smaller smallest so some somebody someone something somewhere state states still such sure t take taken than that that's the their theirs them themselves then there therefore there's these they they'd they'll they're they've thing things think thinks this those though thought thoughts three through thus to today together too took toward turn turned turnings turns two u under until up upon us use used uses v very w want wanted wanting wants was wasn't way ways we we'd well we'll wells went were we're weren't we've what's when what's where whose which while who whom who's whose why why's will with within without won't work worked working works would wouldn't x y year years yes yet you you'll
Stemming

- Techniques to find the stem of a word.
  - Words: User, users, used, using ➔ Stem: use
  - Words: Engineering, engineered ➔ Stem: engineer

- Usefulness for Text Mining
  - improve effectivity of text mining methods
    - matching of similar words
  - reduce term vector size
    - combing words with same stem may reduce the term vector as much as 40-50%.
Some Basic Stemming Rules

– remove endings
  • if a word ends with a consonant other than s, followed by an s, then delete s.
  • if a word ends in es, drop the s.
  • if a word ends in ing, delete the ing unless the remaining word consists only of one letter or of th.
  • If a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a single letter.
  • …...

– transform words
  • if a word ends with “ies” but not “eies” or “aies” then “ies $\Rightarrow$ y.”
Preprocessing Operators in RapidMiner

- To use the operators, you need to install the Text Processing Extension first.
3. Feature Generation

- Text Preprocessing
- Text Transformation (Feature Generation)
- Feature Selection
- Data Mining / Pattern Discovery
- Interpretation / Evaluation
Term-Document Matrix

| Term      | A  | B  | C  | D  | E  | F  | G  | H  | I  | J  | K  | L  | M  | N  | O  | P  | Q  | R  | S  | T  | Σ  |
|-----------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| oil       | 5  | 12 | 2  | 1  | 1  | 7  | 3  | 3  | 5  | 9  | 5  | 4  | 5  | 4  | 3  | 4  | 5  | 3  | 3  | 1  | 85 |
| price     | 5  | 6  | 2  | 2  | 0  | 8  | 1  | 2  | 10 | 5  | 1  | 5  | 2  | 0  | 3  | 3  | 3  | 3  | 3  | 0  | 63 |
| opec      | 0  | 15 | 0  | 0  | 0  | 8  | 1  | 2  | 6  | 5  | 2  | 2  | 4  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 47 |
| mln       | 0  | 4  | 0  | 0  | 2  | 4  | 1  | 0  | 0  | 3  | 9  | 0  | 0  | 0  | 0  | 0  | 3  | 3  | 0  | 2  | 31 |
| market    | 2  | 5  | 0  | 0  | 0  | 3  | 0  | 2  | 0  | 10 | 1  | 2  | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 3  | 30 |
| barrel    | 2  | 0  | 1  | 1  | 0  | 4  | 0  | 0  | 1  | 3  | 3  | 0  | 1  | 1  | 0  | 3  | 3  | 1  | 0  | 2  | 26 |
| bpd       | 0  | 4  | 0  | 0  | 0  | 7  | 0  | 0  | 0  | 2  | 8  | 0  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 1  | 23 |
| dlrs      | 2  | 0  | 1  | 2  | 2  | 2  | 1  | 0  | 0  | 4  | 2  | 0  | 0  | 0  | 0  | 1  | 1  | 1  | 5  | 0  | 23 |
| crude     | 2  | 0  | 2  | 3  | 0  | 2  | 0  | 0  | 0  | 5  | 2  | 0  | 2  | 0  | 0  | 0  | 2  | 0  | 1  | 21 |
| saudi     | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 5  | 7  | 1  | 4  | 0  | 0  | 0  | 0  | 0  | 0  | 18 |
| kuwait    | 0  | 0  | 0  | 0  | 0  | 10 | 0  | 1  | 0  | 3  | 0  | 1  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 17 |
| offic     | 0  | 0  | 0  | 0  | 0  | 5  | 1  | 1  | 0  | 1  | 4  | 3  | 1  | 0  | 0  | 0  | 0  | 1  | 0  | 17 |
| meet      | 0  | 6  | 0  | 0  | 0  | 3  | 0  | 1  | 0  | 1  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 14 |
| pct       | 0  | 0  | 0  | 0  | 2  | 0  | 2  | 2  | 2  | 1  | 0  | 0  | 1  | 0  | 1  | 0  | 1  | 1  | 0  | 2  | 14 |
| product   | 1  | 6  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 4  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 13 |
| accord    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 5  | 1  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 12 |
| futur     | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 12 |
| minist    | 0  | 0  | 0  | 0  | 0  | 3  | 0  | 0  | 1  | 3  | 1  | 2  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 12 |
| govern    | 0  | 0  | 0  | 0  | 0  | 5  | 0  | 6  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 11 |
| month     | 0  | 1  | 0  | 0  | 0  | 2  | 2  | 0  | 1  | 0  | 5  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 11 |
| report    | 0  | 1  | 0  | 0  | 0  | 1  | 8  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 11 |
| sheikh    | 0  | 0  | 0  | 0  | 3  | 0  | 0  | 5  | 2  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 11 |
| industri  | 0  | 2  | 0  | 0  | 0  | 1  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 10 |
| produc    | 0  | 0  | 0  | 0  | 4  | 1  | 1  | 0  | 3  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 10 |
| quota     | 0  | 2  | 0  | 0  | 0  | 4  | 0  | 0  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 10 |
| reserv    | 0  | 0  | 0  | 3  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 3  | 3  | 0  | 0  | 0  | 0  | 10 |
| world     | 0  | 1  | 0  | 0  | 0  | 1  | 3  | 0  | 1  | 1  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 1  | 10 |

Σ        | 48 | 204| 34 | 39 | 46 | 219| 219| 73 | 161| 180| 208| 57 | 61 | 54 | 56 | 68 | 89 | 44 | 147| 32 | 2039 |
Feature Generation

- Document is treated as a **bag of words** (or terms)
  - each word or term becomes a feature.
  - order of words/terms is ignored.

- Each document is represented by a vector.

- Different techniques for vector creation:
  1. **Binary Term Occurrence**: Boolean attributes describe whether or not a term appears in the document.
  2. **Term Occurrence**: Number of occurrences of a term in the document (problematic if documents have different length).
  3. **Terms Frequency**: Attributes represent the frequency in which a term appears in the document (number of occurrences / number of words in document)
  4. **TF-IDF**: see next slide
The TF-IDF Term Weighting Scheme

- The TF-IDF weight (term frequency–inverse document frequency) is used to evaluate how important a word is to a corpus of documents.
  - TF: Term Frequency (see last slide)
  - IDF: Inverse Document Frequency.

\[ w_{ij} = tf_{ij} \times idf_i. \]
\[ idf_i = \log \frac{N}{df_i} \]

- Gives more weight to rare words.
- Give less weight to common words (domain-specific “stopwords”).
Feature Generation in RapidMiner

1. Specify files to process
2. Select feature generation method
4. Feature Selection

- Not all features help!
- Learners might have difficulty with high dimensional data.
Prune methods

- specify if and how **too frequent** or **too infrequent** words should be ignored.

Different options:

- **Percentual**
  - ignore words that appear in less / more than this percentage of all documents.

- **Absolute**
  - ignore words that appear in less / more than that many documents.

- **By Rank**
  - specifies how many percent of the most frequent / infrequent words are ignored.
- POS tagging may be helpful for feature selection.
- Sometimes you want to focus on certain classes of words:
  - Adjectives (JJ.) for sentiment analysis
    - good, bad, great
  - Nouns (N.) for text clustering
    - red and blue cars are similar
    - red and blue trousers are similar
- Rapidminer supports
  - PENN tag system for English
  - STTS tag system for German
  - Filtering conditions are expressed as regular expressions
5. Pattern Discovery

Methods:
1. Cluster Analysis
2. Classification
3. Association Analysis
5.1 Document Clustering

Goal
- Given a set of documents and a similarity measure among documents find clusters such that
  - documents in one cluster are more similar to one another
  - documents in separate clusters are less similar to one another
- using some clustering algorithm.

Applications
- Topical clustering of news stories
- Email message thread identification
- Grouping of document versions

Question
- Which similarity measures are a good choice for comparing document vectors?
The Jaccard coefficient is a popular similarity measure for vectors consisting of asymmetric binary attributes.

\[
dist(x_i, x_j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}
\]

Number of 11 matches / number of not-both-zero attributes values

- **Binary term occurrence vector**
  - 1 represents occurrence of specific word
  - 0 represents absence of specific word
Example: Jaccard Coefficient

- Example document set
  d1 = “Saturn is the gas planet with rings.”
  d2 = “Jupiter is the largest gas planet.”
  d3 = “Saturn is the Roman god of sowing.”

- Documents as binary term occurrence vectors

<table>
<thead>
<tr>
<th></th>
<th>Saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>Jupiter</th>
<th>largest</th>
<th>Roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Jaccard similarities between the documents
  - \( \text{sim}(d1,d2) = 0.44 \)
  - \( \text{sim}(d1,d3) = 0.27 \)
  - \( \text{sim}(d2,d3) = 0.18 \)
Cosine Similarity

- Popular similarity measure for comparing weighted document vectors such as term-frequency or TF-IDF vectors.

\[
\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{||d_1|| \cdot ||d_2||}
\]

where \( \cdot \) indicates vector dot product and \( ||d|| \) is the length of vector \( d \).

- Example

\[
d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0 \\
d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2
\]

\[
d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
\]

\[
||d_1|| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481
\]

\[
||d_2|| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.424
\]

\[
\cos(d_1, d_2) = 0.3150
\]
Example: Cosine Similarity and TF-IDF

- A **commonly used combination** for text clustering.
- Each document is represented by vectors of TF-IDF weights.

- Sample document set:
  - “Saturn is the gas planet with rings.”
  - “Jupiter is the largest gas planet.”
  - “Saturn is the Roman god of sowing.”

- First document as TF-IDF vector:
  \[
  \begin{align*}
  w_{ij} &= tf_{ij} \times idf_i \\
  idf_i &= \log \frac{N}{df_i}
  \end{align*}
  \]

  \((1/7 \times \log(3/2), 1/7\times\log(3/3), 1/7\times\log(3/3), ..., 0, 0, 0, ...)\)

  \[
  \text{Saturn is the Jupiter largest Roman}
  \]

  ![Diagram](image)
Example: Cosine Similarity and TF-IDF

- Sample document set
  d1 = “Saturn is the gas planet with rings.”
  d2 = “Jupiter is the largest gas planet.”
  d3 = “Saturn is the Roman god of sowing.”

- Documents as TF-IDF vectors

<table>
<thead>
<tr>
<th></th>
<th>Saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>Jupiter</th>
<th>largest</th>
<th>Roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d3</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
</tbody>
</table>

- Cosine similarities between the documents
  - $\cos(d1,d2) = 0.13$
  - $\cos(d1,d3) = 0.05$
  - $\cos(d2,d3) = 0.00$
5.2 Document Classification

- Given: A collection of labeled documents (training set).
- Find: A model for the class as a function of the values of the features.
- Goal: Previously unseen documents should be assigned a class as accurately as possible.
- Applications
  - Topical classification of news stories or web pages
  - SPAM detection
  - Sentiment analysis
- Classification methods commonly used for text
  1. Naive Bayes
  2. Support Vector Machines
  3. but KNN or decision trees may also work
Example Application: Sentiment Analysis

- Given: A text
- Goal: Assign a class of sentiment to the text
  - e.g., positive, neutral, negative
  - e.g., sad, happy, angry, surprised

- Can be implemented as supervised classification task
  - requires training data
  - i.e., pairs like <text; sentiment>
Labeling data for sentiment analysis
• is expensive, like every data labeling task

Reviews from the Web may be used as labeled data.

There exist various large corpora of reviews for public download
• Amazon Product Data by Julian McAuley: 142 million reviews from Amazon
• WebDataCommons: 70 million reviews from 50,000 websites that use RDFa or Microdata markup
Preprocessing for Sentiment Analysis

- Recap – we started our processing with:
  Simple Syntactic Analysis
  • Text Cleanup (remove punctuation, HTML tags, …)
  • Normalize case
  • …

- However, reasonable features for sentiment analysis might include
  • punctuation: use of “!”, “?” , “?!”
  • smileys (usually encoded using punctuation: ;-) )
  • use of visual markup, where available (red color, bold face, …)
  • amount of capitalization (“screaming”)
Text Classification Tricks

- Finding selective words
  - weight words according to their correlation with label
  - select top-k words with highest correlation

- Sentiment lexicons
  - use external dictionary of opinion words
  - Bing Liu’s List
    http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar
  - restrict Rapidminer word list to these words
Summary

- **Main challenge in Text Mining: Preprocessing of text**
  • in order to be able to apply well known Data Mining algorithms.

- **There are lots of alternative preprocessing techniques**
  • thus you need to experiment in order to find out which work well for your use case.

- **Text Mining can be tricky, but “ok”-ish results are easily achieved.**
References for this Slideset

- Survey Article
  - Horto et al: A Brief Survey of Text Mining

- Videos
  - Text Mining with RapidMiner by Vancouver Data
  - DWS Screencast: Text Mining with Rapidminer by Robert Meusel

- Additional Tools
  - Stanford NLP: Collection of open-source NLP tools
  - GATE: Feature rich open-source text processing toolkit

- Additional DWS lectures covering Text Mining
  - Text Analytics by Simone Ponzetto
  - Web Mining by Simone Ponzetto and Christian Bizer
    (more on sentiment analysis)
Next Week: Introduction to Student Projects

- **EVERY** student needs to attend the lecture next week (29.03.2017) or

- Inform me via email **BEFORE** the lecture
  - why she/he cannot attend and
  - If she/he already has a group or wants to get assigned to a group

- If you do not attend and do not send a mail, you will be **DELETED** from the participants list.