Data Mining MTAT.03.183

Text Mining

Jaak Vilo

2011 Fall
Thanks

• Viara Popova: Data Mining 2011
Outline

- Why text?
- What is Text Mining?
- Why is text difficult?

- Representations of text
- Tasks and applications

- Document classification
- Document retrieval
- Information extraction

introduction

foundations

tasks
Why Text Mining?

• Text understanding
  – If computers could answer our emails, talk to us, inform us about what is going on in the world, ...
  – Turing test
• But this is too difficult!
• So, we solve smaller subtasks that are still useful but less ambitious
  – Finding specific bits of information on the web, context sensitive spell-checking, authorship analysis, extracting facts or summarising documents
Why is text difficult?

*Time flies like an arrow.*

- Text is subtle, abstract, vague, ambiguous – w.r.t. concepts and relationships
- Many ways to represent the same concept
- Humans use a lot of domain/context/implicit knowledge to make sense of text
- Involves reasoning and probabilistic factors
- Involves social knowledge and interaction with humans
- High dimentionality – huge number of possible features describing the data
- High volume of data
However...

• Text is highly redundant
• Vast amount of data available
• Simpler methods solving less complex practical tasks can be very successful
What is Text Mining?

• Text Mining is the **automatic** discovery of potentially **useful**, previously **unknown** information, from **textual** resources.

• Related areas:
  – Natural language processing – computational linguistics
  – Machine learning – data analysis, knowledge discovery
  – Information retrieval – search, DB
  – Semantic Web – knowledge representation & reasoning
Text representations

- Characters
- Words phrases
- POS tagging
- Taxonomies/thesauri
- Vector-space model
- Language models
- Full parsing
- ...

- Representation depends on the task
Character level

• Sequences of characters with their frequencies – sequences of length 1, 2, 3, ... and how often they occur in the text

• Advantages – simple, avoids the complexities of text
  – Language detection, copy detection, ...

• For most tasks – too simple
Word level

• The characters are organised in words – tokens
• The most commonly used representation
• Tokenisation is not too difficult
  – Chinese?
Word properties

- **Synonymy**: different words, same meaning (goal, objective)
- **Homonymy**: same word, different unrelated meanings (can = able to, can = container)
- **Polysemy**: same word, multiple related meanings (paper = material, paper = article, paper = newspaper)
- **Hyponymy**: one word denotes a subclass of another (laptop → computer)
Word properties

• Word frequencies follow **Power-law distribution** (the frequency of an event varies as a power of some attribute of that event)

from theguardian

Photograph: Murdo MacLeod
Zipf law

- Frequency of item inversely proportionate to its (frequency) rank

\[ f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} (1/n^s)}. \]

- For word frequencies \( s \) is close to 1
- Linear in a log-log graph
Stop words

• Carry very little meaning, mainly functional role and are very frequent
  – Are often removed to improve the results

• Language-specific and application-specific
Stop words

For example:
a, an, and, the, of, on, ...

Or:
able, about, above, abroad, according, accordingly, across, actually, adj, after, afterwards, again, against, ago, ahead, ain't, all, allow, allows, almost, alone, along, alongside, already, also, although, always, am, amid, amidst, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, ...
Stemming

• What should we do with:
  see, seen, seeing, sees, saw, ...
  formalization, formalizations, formally, formalize, formalizing, formal,
  form, forms, forming, formation, formative, ...

• Stemming - transforming a word into its stem
  (normalized form)

• Martin Porter, 1980, de facto standard for English
  formalizing \rightarrow formalize \rightarrow formal \rightarrow form

http://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf

• For feature selection, query broadening, etc.
Phrases level

• n-gram - a contiguous sequence of n items from a given sequence of text

• Google n-gram corpus, 2006 - up to 5-grams
  
  Size: approx. 24 GB compressed text files
  Number of tokens: 1,024,908,267,229
  Number of unigrams: 13,588,391
  Number of bigrams: 314,843,401
  Number of trigrams: 977,069,902
  Number of fourgrams: 1,313,818,354
  Number of fivegrams: 1,176,470,663

• Google Books n-gram corpus, 2009 - 5.2 million books in 6 languages (also by year)

• [http://books.google.com/ngrams](http://books.google.com/ngrams)
Graph these case-sensitive comma-separated phrases: machine learning, data mining, support vector, decision tree.
between 1950 and 2008 from the corpus English with smoothing of 3.
Part-of-Speech level

- Tag words based on their function: verb, noun, adjective, adverb, pronoun, preposition, conjunction, interjection, ...

<table>
<thead>
<tr>
<th>We</th>
<th>like</th>
<th>this</th>
<th>lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>pronoun</td>
<td>verb</td>
<td>determiner</td>
<td>noun</td>
</tr>
</tbody>
</table>

(there is more to it – sub-categories, case, gender, tense, etc.)

- For named entity extraction, feature selection
- Can be manually coded rules, but usually learned with statistical techniques
Taxonomies / thesauri level

• Representation of words based on their meaning
• Linked by relationships:
  – hierarchical relationship
  – equivalence relationship
  – associative relationship
• General or domain-specific
• WordNet – the most well-known thesaurus (English and other languages)
WordNet

- Contains **nouns, verbs, adjectives and adverbs**
- Grouped into sets of cognitive synonyms (synsets) - 117,000 synsets
- Semantic relations link synsets:
  - super-subordinate relation (hyponymy)
    - cat $\rightarrow$ mammal
  - part-whole relation (meronymy)
    - room $\rightarrow$ house
  - antonyms
    - light $\rightarrow$ dark
  - etc.
Vector-space model

• The most commonly used representation
  Bag-of-words representation

• Documents are represented as sparse numeric vectors – processed with linear algebra operations

• Numbers represent word frequencies in the document or 0 if word doesn’t appear

• What about the structure of the text?

• Tasks – classification, clustering, visualization, etc.
Vector-space model

„The key idea is to construct a service network to represent all input and output dependencies between data attributes and operations captured in the service interfaces, and to apply centrality measures from network theory in order to quantify the degree to which an attribute belongs to a given subsystem.“

<table>
<thead>
<tr>
<th>add</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>append</td>
<td>0</td>
</tr>
<tr>
<td>apply</td>
<td>1</td>
</tr>
<tr>
<td>attribute</td>
<td>2</td>
</tr>
<tr>
<td>best</td>
<td>0</td>
</tr>
<tr>
<td>belong</td>
<td>1</td>
</tr>
<tr>
<td>capture</td>
<td>1</td>
</tr>
<tr>
<td>carry</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Too simple – why?
TF-IDF

\[
tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}
\]

t – term

\(d\) – document

tf – term frequency
TF-IDF

$t$ – term
$d$ – document

$$
tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}
$$

**tf** – term frequency

**D** – document corpus

**IDF** – Inverse Document Frequency

$$
idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}
$$

$$
tfidf(t, d, D) = tf(t, d) \times idf(t, D)
$$
Language models

• Probability of next word given a sequence of words
• Tasks: speech recognition, OCR, handwriting recognition, machine translation, spell checking, etc.
Parse tree

- Structural information about the sentence
- Not unique

Fruit flies like a banana

- Not used much in TM
Tasks

- Document summarization
- Text segmentation
- Document classification
- Document clustering
- **Information extraction**
- Machine translation
- Spell checking
- Sentiment analysis
- Discourse analysis

**Sub-tasks:**

- Named-entity recognition (NER)
- Word sense disambiguation
- Sentence splitting
- POS tagging
- Parsing

...
Document classification

- Classification of news articles by topic: business, science, entertainment, etc.
- Adding descriptive terms to documents to improve accessibility in a large library
- Spam filtering
- Language guessing
- Genre detection
- ...
Document classification

• Supervised learning – on textual data
• We can use bag-of-words representation
  – Stop words removal
  – Stemming
  – Etc.
• Huge number of features, sparseness – some methods work better than others
  – Support vector machines
  – Naïve Bayes classifier
  – K-nearest neighbour
  – Etc.
Bayesian spam filtering
(text classification)

• http://en.wikipedia.org/wiki/Bayesian_spam_filtering
Computing the probability that a message containing a given word is spam

Let's suppose the suspected message contains the word "replica". Most people who are used to receiving e-mail know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches. The spam detection software, however, does not "know" such facts, all it can do is compute probabilities.

The formula used by the software to determine that is derived from Bayes' theorem

\[
Pr(S|W) = \frac{Pr(W|S) \cdot Pr(S)}{Pr(W|S) \cdot Pr(S) + Pr(W|H) \cdot Pr(H)}
\]

where:

- \(Pr(S|W)\) is the probability that a message is a spam, knowing that the word "replica" is in it;
- \(Pr(S)\) is the overall probability that any given message is spam;
- \(Pr(W|S)\) is the probability that the word "replica" appears in spam messages;
- \(Pr(H)\) is the overall probability that any given message is not spam (is "ham");
- \(Pr(W|H)\) is the probability that the word "replica" appears in ham messages.
Document categorization

- Multiple categories relevant for a document
  - A separate classifier for each category

- A hierarchy of categories (taxonomy)
  - Separate classifiers for the leaves
  - Separate classifiers for all nodes
  - A hierarchy of classifiers

http://dir.yahoo.com/
Information retrieval

- A large collection of documents
- The user inputs a query
- Retrieve the documents most relevant for the query
- Vocabulary mismatch
- Bag-of-words representation
Document similarity

Documents are vectors (tf-idf weights)

Query – a vector

Find vectors with smallest angle between query and doc

Cosine similarity

$$\text{sim}(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$
Information extraction

• Extract specific type of information from text
  – Not documents but concepts and relationships
    – [http://www.youtube.com/watch?v=Ccum_Cu9sP8](http://www.youtube.com/watch?v=Ccum_Cu9sP8)

• TweetBeat
  – [http://www.youtube.com/watch?v=g3AqdI DYG0c](http://www.youtube.com/watch?v=g3AqdI DYG0c)
Videos

Hurricane Sandy: View the Superstorm Through the Eyes of Twitter

By filtering tweets that discuss the storm, get a first hand look at how the entire country was talking about the storm as it approached, made landfall, and took its course over the U.S. eastern seaboard.

US Presidential Election 2012 on Twitter

During the U.S. Presidential Election, Twitter reported 327,452 tweets per minute as news outlets began to predict an Obama victory. President Obama’s victory tweet quickly became the most-retweeted message in history, setting the record in just 22 minutes after the tweet was posted. This heat map has adjusted the color spectrum to show the intensity and location of tweets favorable to either President Barack Obama or Governor Mitt Romney over the course of Election Day, from the first polls opening to after President Obama’s victory speech.

The Global Twitter Heartbeat — Analyzing real-time data on the SGI UV

Heer Kalev Leetaru discuss how he went about constructing the data analysis pipeline, ingesting, analyzing and creating the visualizations on the SGI UV 2000.
Heat Maps of Sentiment on Twitter

Hurricane Sandy

RED represents more Negative sentiment.
BLUE represents more Positive sentiment.
Download full resolution image (ZIP)

US Presidential Election 2012

RED represents tweets about Romney.
BLUE represents tweets about Obama.
Download full resolution image (ZIP)

US Sentiment from Live Twitter Feed

Global Sentiment from Live Twitter Feed

Jaak Vilo and other authors
UT: Data Mining 2009
Vascular endothelial growth factor is induced by the inflammatory cytokines interleukin-6 [?] and oncostatin m [?] in human adipose tissue in vitro and in murine adipose tissue in vivo.


Department of Internal Medicine II, Medical University Vienna, Waehringer Guertel 18-20, A-1090 Vienna, Austria.

OBJECTIVES: It is believed that adipose tissue acts as an endocrine organ by producing inflammatory mediators and thereby contributes to the increased cardiovascular risk seen in obesity. A link between adipose tissue mass and angiogenesis has been suggested. Vascular endothelial growth factor (VEGF) seems to be implicated in this process. Members of the glycoprotein (gp)130 ligand family regulate VEGF expression in other cells. METHODS AND RESULTS: We used tissue explants as well as primary cultures of preadipocytes and adipocytes from human subcutaneous and visceral adipose tissue to investigate whether the gp130 ligands oncostatin M [?] (OSM [?]), interleukin-6 [?] (IL-6 [?]), leukemia inhibitory factor [?] (LIF [?]), and cardiotrophin-1 [?] (CT-1 [?]) regulate VEGF expression in human adipose tissue. Human subcutaneous and visceral adipose tissue responded to treatment with IL-6 [?] and OSM [?] with a significant increase in VEGF production. Human preadipocytes were isolated from subcutaneous and visceral adipose tissue. Adipocyte-differentiation was induced by hormone-supplementation. All cell types responded to IL-6 [?] and OSM [?] with a robust increase in VEGF protein production and a similar increase in VEGF-specific mRNA. Furthermore, IL-1beta synergistically enhanced the effect of OSM [?] on VEGF production. AG-490, a JAK/STAT inhibitor, abolished the OSM [?] -dependent VEGF induction almost completely. In mice, IL-6 [?] and OSM [?] increased serum levels of VEGF and VEGF mRNA and vessel density in adipose tissue. CONCLUSION: We speculate that the inflammatory cytokines IL-6 [?] and OSM [?] might support angiogenesis during adipose tissue growth by upregulating VEGF.
Information extraction tasks

• Identifying fields (named entity recognition)
• Relationships between fields (record association)
• Normalization and deduplication

• Usually the fields are given – we know what we are looking for
Approaches to IE

• Hand-crafted rules

• Learning from data
  – Annotated documents – training data
  – Design appropriate features
  – Train a classifier to produce annotations to a new document

• The documents are not the training instances!

• Local features:
  – Contents, text just before/after, separators on boundaries
Example features

• Begins with a number
• Contains only capitals
• Ends with punctuation
• Contains a city name (gazetteers)
• Is part of a noun phrase (POS tagging)
• Is in bold font
• Is in HTML title tag (HTML formatting)
• Ends in „ov“ or „ev“ or „ski“
• Is in stop word list
• ...
How to define instances

- Candidate classification – predefine the candidates and use them as instances
- Sliding window – every sequence (between m and n – e.g. from the training data)
- Boundary detection – discover start and end separately and associate them appropriately
- Finite state machines – Hidden Markov Models
Project 1.2
BioMedical
Midterm review
24.11.2012
Jaak Vilo
Medical IT

Terminology, Ontology

Descriptive Analysis

Information Extraction

ETL

OLAP

Query Reports

Text mining results

STACC Software Technology and Applications Competence Center
Text corpora

- East Tallinn Hospital (Oct 2010)
  - Cancer (mostly) /1420 patients, texts, 0.1M words, 0.8MB /
- E-Health (since spring 2012)
  - 0.95M rows, 9.5M words, 68MB of text (ambulatory)
  - 0.6M rows, 11.8M words, 96MB of text (stationary)
  - Working on new extraction with added data
- **GP texts** (since August 2012)
  - 1995-2010, 45 units, **146 doctors, 165K patients, 800K** diagnoses (6702 different), 2.5M analyses (287 codes), 820K prescriptions (613 ATC codes, 110 449 different texts), 113K X-ray studies.
- Drug labels
  - 3992 info sheets, 68MB of text
  - 4011 drug properties, 111MB of text
- SNOMED, ICD-10, ATC, ... ontologies.
Information extraction and mining

- Text issues. Ca - circa, Calcium, Carcinoma, ...

- NER

- Diagnoses (ICD)

- Drugs (ATC)

- Complaints (discover)
Automatic Extraction of Patient’s Complaints

- **Goal**
  - Build a system enabling automatic extraction of patients’ complaints from free text.
  - Create a list of common complaints.
- **Data**: E-Health Anamnesis Text Corpus (600K records)
- **Method**: Statistical named entity recognition
  - Part of the data has been manually labeled and fed to the system to learn rules which describe complaint-entities in text. The rules are formed based on word’s morphological, grammatical, semantical and contextual attributes. Resulting system applies these rules to identify complaints in future text.
- **Results**
  - Precision and recall at around 70% so far.
1. Primary hypertension is diagnosed
2. Type 2 diabetes mellitus is diagnosed
3. Receipt for antibiotics
4. Acute bronchitis is rediagnosed
5. Multiple drugs against bronchitis

- Bronchitis
- Diabetes
- Hypertension
Profile of patients with particular diagnosis

Example: F10 – Alcohol related disorders

Among patients who have diagnosis „F10 - Alcohol related disorders“ it is:

- **39 times higher** representation of „K70 - Alcoholic liver disease“ thin in reference group
- **2-4 times higher** representation of depression and sleep disorders (F3-F5)
- **Approximately 2-3x lower** representation of acute upper respiratory infections (J00-J06)
I* and M* diagnoses (heart (blue) and muscle-related)

Cold (blue) and heart (red)

Incidents related to falling
Text-mining for Diagnose-specific phrases (SPEXS2)

• The aim:
  • identifying characteristic clinical phrases for various diagnoses or diagnose groups using frequency profiling

• Examples:
  • J01 survetunne siinuste prk
  • J20 kopsudel bialter räginaid
  • J06 nohu köha t
  • J06 kurk veidi hüpereemiline
  • N30 sagenenud valulik urineerimine
  • M54 nimmest liikuvus piiratud

• Possible Uses:
  • Early detection of chronic disorders
  • Text-mining based diagnostics
  
  22000 sentences (ca 1MB), vs 250000 sent. (11mb); 3-7 word phrases, >5 occurrences: ca 25 sec