Text Mining Overview

NCSU MSA Text Mining

Tao Xie

Major slides borrowed and adapted from KDD 2007 Tutorial Slides by Marko Grobelnik, Dunja Mladenic
Most Data are Unstructured (Text) or Semi-Structured…

- Email
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- …

- Customer complaint letters
- Contracts
- Transcripts of phone calls with customers
- Technical documents
- …

Text data mining has become more and more important…

(Adapted from J. Dorre et al. “Text Mining: Finding Nuggets in Mountains of Textual Data”)
What is Text Mining?

“…finding **interesting** regularities in large **textual** datasets…” (adapted from Usama Fayad)

- …where **interesting** means: non-trivial, hidden, previously unknown and potentially useful

“…finding **semantic** and **abstract** information from the surface form of textual data…”
Applications of Text Mining

- **Direct applications**
  - Discovery-driven (Bioinformatics, Business Intelligence, etc): We have specific questions; how can we exploit data mining to answer the questions?
  - Data-driven (WWW, literature, email, customer reviews, etc): We have a lot of data; what can we do with it?

- **Indirect applications**
  - Assist information access (e.g., discover latent topics to better summarize search results)
  - Assist information organization (e.g., discover hidden structures)
Why dealing with Text is Tough? (M.Hearst 97)

- Abstract concepts are **difficult to represent**
- “**Countless**” combinations of subtle, abstract relationships among concepts
- **Many ways** to represent similar concepts
  - E.g. space ship, flying saucer, UFO
- **High dimensionality**
- **Tens or hundreds of thousands of features**
Why dealing with Text is Easy? (M.Hearst 97)

- Highly redundant data
  - …most of the techniques count on this property

- Just about any simple algorithm can get “good” results for simple tasks:
  - Pull out “important” phrases
  - Find “meaningfully” related words
  - Create some sort of summary from documents
Who is in the text analysis arena?

Knowledge Rep. & Reasoning / Tagging

Semantic Web

Web 2.0

Computational Linguistics

Text Analytics

Search & DB

Information Retrieval

Data Analysis

Natural Language Processing

Machine Learning

Text Mining

http://en.wikipedia.org/wiki/Web_2.0
What dimensions are in text analytics?

- Three major dimensions of text analytics:
  - Representations
    - ...from character-level to first-order theories
  - Techniques
    - ...from manual work, over learning to reasoning
  - Tasks
    - ...from search, over (un-)supervised learning, to visualization, summarization, translation ...
Text-Mining

How do we represent text?
Levels of text representations

- Character (character n-grams and sequences)
- Words (stop-words, stemming, lemmatization)
- Phrases (word n-grams, proximity features)
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Levels of text representations

- **Character**
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

---

- Vector-space model
- Language models
- Full-parsing
- Cross-modality

---

- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Character level

- Character level representation of a text consists from **sequences** of characters...
  - ...a document is represented by a frequency distribution of sequences
  - Usually we deal with contiguous strings...
  - ...each character sequence of length 1, 2, 3, ... represent a feature with its frequency
Good and bad sides

- Representation has several important strengths:
  - ...it is very robust since avoids language morphology (patterns of word formation)
    - (useful for e.g. language identification)
  - ...it captures simple patterns on character level
    - (useful for e.g. spam detection, copy detection)
  - ...because of redundancy in text data it could be used for many analytic tasks
    - (learning, clustering, search)

- ...for deeper semantic tasks, the representation is too weak
Levels of text representations

- Character
- **Words**
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Word level

- The most common representation of text used for many techniques
  - ...there are many tokenization software packages which split text into the words

- Important to know:
  - Word is well defined unit in western languages – e.g., Chinese has different notion of semantic unit
Words Properties

- Relations among word surface forms and their senses:
  - **Homonomy**: same form, but different meaning (e.g. bank: river bank, financial institution)
  - **Polysemy**: same form, related meaning (e.g. bank: blood bank, financial institution)
  - **Synonymy**: different form, same meaning (e.g. singer, vocalist)
  - **Hyponymy**: one word denotes a subclass of another (e.g. breakfast, meal)

- Word frequencies in texts have **power distribution**:
  - …small number of very frequent words
  - …big number of low frequency words
Stop-words

- Stop-words are words that from non-linguistic view do not carry information
  - …they have mainly functional role
  - …usually we remove them to help the methods to perform better

- Stop words are language dependent – examples:
  - **English**: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ...
  - **Dutch**: de, en, van, ik, te, dat, die, in, een, hij, het, niet, zijn, is, was, op, aan, met, als, voor, had, er, maar, om, hem, dan, zou, of, wat, mijn, men, dit, zo, ...
  - **Slovenian**: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ...
Word character level normalization

- Hassle which we usually avoid:
  - Since we have plenty of character encodings in use, it is often nontrivial to identify a word and write it in unique form
  - …e.g. in Unicode the same word could be written in many ways – canonization of words:

<table>
<thead>
<tr>
<th>Source</th>
<th>NFD</th>
<th>NFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>A ́</td>
<td>Å</td>
</tr>
<tr>
<td>00C5</td>
<td>0041 030A</td>
<td>00C5</td>
</tr>
<tr>
<td>Ô</td>
<td>O ́</td>
<td>Ô</td>
</tr>
<tr>
<td>00F4</td>
<td>006F 0302</td>
<td>00F4</td>
</tr>
</tbody>
</table>

http://unicode.org/reports/tr15/

NFD: Normalization Form D
NFC: Normalization Form C
Stemming (1/2)

- Different forms of the same word are usually problematic for text data analysis, because they have different spelling and similar meaning (e.g. learns, learned, learning, ...)

- **Stemming** is a process of transforming a word into its stem (normalized form)
  - ...stemming provides an inexpensive mechanism to merge
For English is mostly used **Porter stemmer** at [http://www.tartarus.org/~martin/PorterStemmer/](http://www.tartarus.org/~martin/PorterStemmer/)

Example cascade rules used in English Porter stemmer

- ATIONAL -> ATE       relational -> relate
- TIONAL   -> TION      conditional -> condition
- ENCI     -> ENCE      valenci -> valence
- ANCI     -> ANCE      hesitanci -> hesitance
- IZER     -> IZE       digitizer -> digitize
- ABLI     -> ABLE      conformabli -> conformable
- ALLI     -> AL        radicalli -> radical
- ENTLI    -> ENT       differentli -> different
- ELI      -> E         vileli -> vile
- OUSLI    -> OUS       analogousli -> analogous
Levels of text representations

- Character
- Words
- **Phrases**
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

©Marko Grobelenik and Dunja Mladenic
Phrase level

- Instead of having just single words we can deal with phrases
- We use two types of phrases:
  - Phrases as frequent contiguous word sequences
  - Phrases as frequent non-contiguous word sequences (e.g., “Mary switches her table lamp off.”)
- The main effect of using phrases is to more precisely identify sense (one of the meanings of a word)
In September 2006 Google announced availability of n-gram corpus:

- [http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html](http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html)

Some statistics of the corpus:
- File sizes: approx. 24 GB compressed (gzip'ed) text files
- Number of tokens: 1,024,908,267,229
- Number of sentences: 95,119,665,584
- 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
Example: Google n-grams

- ceramics collectables collectibles 55
- ceramics collectables fine 130
- ceramics collected by 52
- ceramics collectible pottery 50
- ceramics collectible cookware 45
- ceramics collection, 144
- ceramics collection . 247
- ceramics collection </S> 120
- ceramics collection and 43
- ceramics collection at 52
- ceramics collection is 68
- ceramics collection of 76
- ceramics collection | 59
- ceramics collections , 66
- ceramics collections . 60
- ceramics combined with 46
- ceramics come from 69
- ceramics comes from 660
- ceramics community, 109
- ceramics community . 212
- ceramics community for 61
- ceramics companies , 53
- ceramics companies consultants 173
- ceramics company ! 4432
- ceramics company , 133
- ceramics company . 92
- ceramics company </S> 41
- ceramics company facing 145
- ceramics company in 181
- ceramics company started 137
- ceramics company that 87
- ceramics component ( 76
- ceramics composed of 85

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensable 40
- serve as the individual 234
- serve as the industrial 52
- serve as the industry 607
- serve as the info 42
- serve as the informal 102
- serve as the information 838
- serve as the informational 41
- serve as the infrastructure 500
- serve as the initial 5331
- serve as the initiating 125
- serve as the initiation 63
- serve as the initiator 81
- serve as the injector 56
- serve as the inlet 41
- serve as the inner 87
- serve as the input 1323
- serve as the inputs 189
- serve as the insertion 49
- serve as the insourced 67
- serve as the inspection 43
- serve as the inspector 66
- serve as the inspiration 1390
- serve as the installation 136
- serve as the institute 187
Levels of text representations

- Character
- Words
- Phrases
- **Part-of-speech tags**
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Part-of-Speech level

- By introducing part-of-speech tags we introduce word-types enabling to differentiate words functions
  - For text analysis, part-of-speech information is used mainly for “information extraction” where we are interested in e.g. named entities which are “noun phrases”
  - Another possible use is reduction of the vocabulary (features)
    - …it is known that nouns carry most of the information in text documents
- Part-of-Speech taggers are usually learned on manually tagged data
## Part-of-Speech Table

<table>
<thead>
<tr>
<th>part of speech</th>
<th>function or &quot;job&quot;</th>
<th>example words</th>
<th>example sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>action or state</td>
<td>(to) be, have, do, like,</td>
<td>EnglishClub.com is a web site. I like EnglishClub.com.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>work, sing, can, must</td>
<td></td>
</tr>
<tr>
<td>Noun</td>
<td>thing or person</td>
<td>pen, dog, work, music,</td>
<td>This is my dog. He lives in my house. We live in London.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>town, London, teacher, John</td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td>describes a noun</td>
<td>a/an, the, 69, some,</td>
<td>My dog is big. I like big dogs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>good, big, red, well,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>interesting</td>
<td></td>
</tr>
<tr>
<td>Adverb</td>
<td>describes a verb,</td>
<td>quickly, silently, well,</td>
<td>My dog eats quickly. When he is very hungry, he eats really quickly.</td>
</tr>
<tr>
<td></td>
<td>adjective or adverb</td>
<td>badly, very, really</td>
<td></td>
</tr>
<tr>
<td>Pronoun</td>
<td>replaces a noun</td>
<td>I, you, he, she, some</td>
<td>Tara is Indian. She is beautiful.</td>
</tr>
<tr>
<td>Preposition</td>
<td>links a noun to another word</td>
<td>to, at, after, on, but</td>
<td>We went to school on Monday.</td>
</tr>
<tr>
<td>Conjunction</td>
<td>joins clauses or sentences or words</td>
<td>and, but, when</td>
<td>I like dogs and I like cats. I like cats and dogs. I like dogs but I don't like</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cats.</td>
</tr>
<tr>
<td>Interjection</td>
<td>short exclamation,</td>
<td>oh!, ouch!, hi!, well</td>
<td>Ouch! That hurts! Hi! How are you? Well, I don't know.</td>
</tr>
<tr>
<td></td>
<td>sometimes inserted into a sentence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[http://www/englishclub.com/grammar/parts-of-speech_1.htm](http://www/englishclub.com/grammar/parts-of-speech_1.htm)
Part-of-Speech examples

<table>
<thead>
<tr>
<th>verb</th>
<th>noun</th>
<th>verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop!</td>
<td>John</td>
<td>works.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>noun</th>
<th>verb</th>
<th>verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>is</td>
<td>working.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pronoun</th>
<th>verb</th>
<th>noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>She</td>
<td>loves</td>
<td>animals.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>noun</th>
<th>verb</th>
<th>adjective</th>
<th>noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animals</td>
<td>like</td>
<td>kind</td>
<td>people.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>noun</th>
<th>verb</th>
<th>noun</th>
<th>adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tara</td>
<td>speaks</td>
<td>English</td>
<td>well.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>noun</th>
<th>verb</th>
<th>adjective</th>
<th>noun</th>
<th>adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tara</td>
<td>speaks</td>
<td>good</td>
<td>English.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pronoun</th>
<th>verb</th>
<th>preposition</th>
<th>adjective</th>
<th>noun</th>
<th>adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>She</td>
<td>ran</td>
<td>to</td>
<td>the</td>
<td>station</td>
<td>quickly.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pron.</th>
<th>verb</th>
<th>adj.</th>
<th>noun</th>
<th>conjunction</th>
<th>pron.</th>
<th>verb</th>
<th>pron.</th>
</tr>
</thead>
<tbody>
<tr>
<td>She</td>
<td>likes</td>
<td>big</td>
<td>snakes</td>
<td>but</td>
<td>I</td>
<td>hate</td>
<td>them.</td>
</tr>
</tbody>
</table>

Here is a sentence that contains every part of speech:

<table>
<thead>
<tr>
<th>interjection</th>
<th>pron.</th>
<th>conj.</th>
<th>adj.</th>
<th>noun</th>
<th>verb</th>
<th>prep.</th>
<th>noun</th>
<th>adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well,</td>
<td>she</td>
<td>and</td>
<td>young</td>
<td>John</td>
<td>walk</td>
<td>to</td>
<td>school</td>
<td>slowly.</td>
</tr>
</tbody>
</table>
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- **Taxonomies / thesauri**
  - Vector-space model
  - Language models
  - Full-parsing
  - Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Taxonomies/thesaurus level

- Thesaurus has a main function to connect different surface word forms with the same meaning into one sense (synonyms)
  - ... additionally we often use hypernym relation to relate general-to-specific word senses
  - ... by using synonyms and hypernym relation we compact the feature vectors

- The most commonly used general thesaurus is WordNet which exists in many other languages (e.g. EuroWordNet)
  - [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/)
WordNet – database of lexical relations

- WordNet is the most well developed and widely used lexical database for English
  - ...it consist from 4 databases (nouns, verbs, adjectives, and adverbs)
- Each database consists from sense entries – each sense consists from a set of synonyms, e.g.:
  - musician, instrumentalist, player
  - person, individual, someone
  - life form, organism, being

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Forms</th>
<th>Number of Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>94474</td>
<td>116317</td>
</tr>
<tr>
<td>Verb</td>
<td>10319</td>
<td>22066</td>
</tr>
<tr>
<td>Adjective</td>
<td>20170</td>
<td>29881</td>
</tr>
<tr>
<td>Adverb</td>
<td>4546</td>
<td>5677</td>
</tr>
</tbody>
</table>

A word sense is one of the meanings of a word.
WordNet sample entry of “bass”

Noun

- **S:** (n) **bass** (the lowest part of the musical range)
  - **direct hypernym** / **inherited hypernym** / **sister term**
  - **S:** (n) **pitch** (the property of sound that varies with variation in the frequency of vibration)
    - **direct hyponym** / **full hyponym**
    - attribute
    - **direct hypernym** / **inherited hypernym** / **sister term**
    - **derivationally related form**
- **S:** (n) **bass**, **bass part** (the lowest part in polyphonic music)
- **S:** (n) **bass**, **basso** (an adult male singer with the lowest voice)
- **S:** (n) **sea bass**, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- **S:** (n) **freshwater bass**, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S:** (n) **bass**, **bass voice**, **basso** (the lowest adult male singing voice)
- **S:** (n) **bass** (the member with the lowest range of a family of musical instruments)
- **S:** (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- **S:** (adj) **bass**, **deep** (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone"
WordNet – excerpt from the graph

26 relations (ex next slide)
WordNet Nouns: 116k senses
WordNet relations

- Each WordNet entry is connected with other entries in the graph through relations
- Relations in the database of nouns:

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>From lower to higher concepts</td>
<td>breakfast -&gt; meal</td>
</tr>
<tr>
<td>Hyponym</td>
<td>From concepts to subordinates</td>
<td>meal -&gt; lunch</td>
</tr>
<tr>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty -&gt; professor</td>
</tr>
<tr>
<td>Member-Of</td>
<td>From members to their groups</td>
<td>copilot -&gt; crew</td>
</tr>
<tr>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td>table -&gt; leg</td>
</tr>
<tr>
<td>Part-Of</td>
<td>From parts to wholes</td>
<td>course -&gt; meal</td>
</tr>
<tr>
<td>Antonym</td>
<td>Opposites</td>
<td>leader -&gt; follower</td>
</tr>
</tbody>
</table>
WordNet hierarchies (hyponym)

Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist
  => musician, instrumentalist, player
  => performer, performing artist
  => entertainer
  => person, individual, someone...
    => life, form, organism, being...
      => entity, something
  => causal agent, cause, causal agency
  => entity, something

Sense 7
bass --
(the member with the lowest range of a family of musical instruments)
=> musical instrument
  => instrument
  => device
    => instrumentality, instrumentation
      => artifact, artefact
    => object, physical object
      => entity, something
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

---

**Vector-space model**

- Language models
- Full-parsing
- Cross-modality

---

- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
The most common way to deal with documents is first to transform them into **sparse numeric vectors** and then deal with them with **linear algebra operations**

- …by this, we forget everything about the linguistic structure within the text
- …this is sometimes called “structural curse” because this way of forgetting about the structure doesn’t harm efficiency of solving many relevant problems
- This representation is referred to also as “Bag-Of-Words” or “Vector-Space-Model”
- Typical tasks on vector-space-model are classification, clustering, visualization etc.
Bag-of-words document representation

Journal of Artificial Intelligence Research

JAIR is a refereed journal, covering all areas of Artificial Intelligence, which is distributed free of charge over the internet. Each volume of the journal is also published by Morgan Kaufman....
Word weighting

- In the bag-of-words representation each word is represented as a separate variable having numeric weight (importance).

- The most popular weighting schema is normalized word frequency TFIDF (Term Frequency Inverse Document Frequency):

\[ tfidf(w) = tf \cdot \log \left( \frac{N}{df(w)} \right) \]

- \( tf(w) \) – term frequency (number of word occurrences in a document)
- \( df(w) \) – document frequency (number of documents containing the word)
- \( N \) – number of all documents
- \( tfidf(w) \) – relative importance of the word in the document

The word is more important if it appears several times in a target document. The word is more important if it appears in less documents.
Example document and its vector representation

- TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.

- [RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171] [ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119] [DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102] [DONALD:0.097] [COMMON:0.093] [GIVING:0.081] [OWNS:0.080] [MAKES:0.078] [TIMES:0.075] [SHARE:0.072] [JAMES:0.070] [REAL:0.068] [CONTROL:0.065] [ACQUIRE:0.064] [OFFERED:0.063] [BID:0.063] [LATE:0.062] [OUTSTANDING:0.056] [SPOKESMAN:0.049] [CHAIRMAN:0.049] [INTERNATIONAL:0.041] [STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]
Similarity between document vectors

- Each document is represented as a vector of weights $D = \langle x \rangle$

- Magnitude of vector difference between two doc vectors
  - Drawback: two similar-content docs can have significant vector difference if one is much longer than the other (same relative distribution of terms but absolute term frequencies of one larger)

- Cosine similarity (dot product) is the most widely used similarity measure between two document vectors
  - ...calculates cosine of the angle between document vectors
  - ...compensates for the effect of document length
  - ...similarity value between 0 (different) and 1 (the same)

$$Sim(D_1, D_2) = \frac{\sum_{i} x_{1i}x_{2i}}{\sqrt{\sum_{j} x_{j}^2} \sqrt{\sum_{k} x_{k}^2}}$$
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- **Language models**
  - Full-parsing
  - Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

©Marko Grobelnik and Dunja Mladenic
Language model level

- Language modeling is about determining probability of a sequence of words
  - The task typically gets reduced to the estimating probabilities of a next word given two previous words (trigram model):
    \[ P(w_i|w_{i-2}w_{i-1}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} \]
  - It has many applications including speech recognition, OCR, handwriting recognition, machine translation and spelling correction

©Marko Grobelnik and Dunja Mladenic
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- **Full-parsing**
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Full-parsing level

- Parsing provides maximum structural information per sentence
- On the input we get a sentence, on the output we generate a parse tree
- For most of the methods dealing with the text data the information in parse trees is too complex
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Cross-modality level

- It is very often the case that objects are represented with different data types:
  - Text documents
  - Multilingual texts documents
  - Images
  - Video
  - Social networks
  - Sensor networks

- …the question is how to create mappings between different representation so that we can benefit using more information about the same objects
Example: Aligning text with audio, images and video

- The word “tie” has several representations ([http://www.answers.com/tie&r=67](http://www.answers.com/tie&r=67))
  - Textual
  - Multilingual text
    - (tie, kravata, krawatte, …)
  - Audio
  - Image:
    - [http://images.google.com/images?hl=en&q=necktie](http://images.google.com/images?hl=en&q=necktie)
  - Video (movie on the right)

- Out of each representation we can get set of features and the idea is to correlate them
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- **Collaborative tagging / Web2.0**
- Templates / Frames
- Ontologies / First order theories
Collaborative tagging

- Collaborative tagging is a process of adding metadata to annotate content (e.g., documents, web sites, photos)
  - …metadata is typically in the form of keywords
  - …this is done in a collaborative way by many users from larger community collectively having good coverage of many topics
  - …as a result we get annotated data where tags enable comparability of annotated data entries
Example: flickr.com tagging

Tags entered by users annotating photos
Example: del.icio.us tagging

Tags entered by users annotating Web sites
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- **Templates / Frames**
- Ontologies / First order theories
Template / frames level

- Templates are the mechanism for extracting the information from text
  - ...templates always focused on specific domain which includes consistent patterns on where specific information is positioned
  - Templates are one of the basic methods for information extraction
Examples of templates of KnowItAll system

- Generic approach of extracting is described in
  - *Unsupervised named-entity extraction from the Web: An experimental study (Oren Etzioni et al)*

- KnowItAll system uses the following generic templates:
  - NP “and other” <class1>
  - NP “or other” <class1>
  - <class1> “especially” NPList
  - <class1> “including” NPList
  - <class1> “such as” NPList
  - “such” <class1> “as” NPList
  - NP “is a” <class1>
  - NP “is the” <class1>

- ...each template represents specific relationship between the words appearing in the variable slots
Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Ontologies level

- Ontologies are the most general formalism for describing data objects
  - ...got popular through Semantic Web and OWL (Web Ontology Language) standard
  - Ontologies can be from relatively simple ones (light weight described with simple) to heavy weight (described with first order theories).

An ontology is both a controlled vocabulary of things in the real world and captures the relations between them.
General Knowledge about Terrorism:
Terrorist groups are capable of directing assassinations:
(implies
  (isa ?GROUP TerroristGroup)
  (behaviorCapable ?GROUP AssassinatingSomeone directingAgent))

... If a terrorist group considers an agent an enemy, that agent is vulnerable to an attack by that group:
(implies
  (and
    (isa ?GROUP TerroristGroup)
    (considersAsEnemy ?GROUP ?TARGET))
  (vulnerableTo ?GROUP ?TARGET TerroristAttack))
Text-Mining

Typical tasks on text
Document Summarization
**Document Summarization**

- **Task**: the task is to produce shorter, summary version of an original document

Two main approaches to the problem:

- **Selection based** – summary is selection of sentences from an original document
- **Knowledge rich** – performing semantic analysis, representing the meaning and generating the text satisfying length restriction
Selection based summarization

- Three main phases:
  - Analyzing the source text
  - Determining its important points (units)
  - Synthesizing an appropriate output

- Most techniques adopt linear weighting model – each text unit (sentence) is assessed by the following formula:
  - \( \text{Weight}(U) = \text{LocationInText}(U) + \text{CuePhrase}(U) + \ldots \) (other statistics)
  - Cue phrases: words and phrases such as now and well which serve primarily to indicate document structure or flow, rather than to impart semantic information about the current topic.

- …output consists from topmost text units (sentences)
Example of selection based approach from MS Word

**Tutorial title**
Text Mining and Link Analysis for Web Data

Presenter contact information including the e-mail address

Dunja Mladenic
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Dunja.Mladenic@ijs.si
Phone: +386 1 4773 377

Marko Grobelnik
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Marko.Grobelnik@ijs.si
Phone: +386 1 4773 778

Aims/Learning objectives:

The aim of this tutorial is to present topics from the areas of text mining and link analysis in the relationship to the web data. The goal is to show the whole list of nontrivial problems appearing in everyday life and occasionally in professional work with the web and to show how they can be approached using text mining and link analysis techniques and tools. The goal is to make an overview of the available approaches, which are potentially useful for solving interesting problems connected to the documents and their linkage coming from the web structure.

Duration (half or full day)
Half day, but it could be scaled to full day

Scope (general topic area) and why it is relevant for WWW2004:

The tutorial’s relevance for the WWW2004 is in the presentation of analytic approaches used on the web data (text+links). In particular, the tutorial will focus on the possibilities offered by two very active and relevant subfields of data mining: text mining and link analysis. The relevance of these topics to the WWW2004 public is in extending possible activities, which could be used in shaping, understanding and potentially predicting the static and dynamic nature of the web. Analysis of such data offers typically new insights in the nature of the complex web data. Suitability of the tutorial for the WWW2004
Knowledge rich summarization

- To generate ‘true’ summary of a document we need to (at least partially) ‘understand’ the document text
  - …the document is too small to count on statistics, we need to identify and use its linguistic and semantic structure

- On the next slides we show an approach from (Leskovec, Grobelnik, Milic-Frayling 2004) using 10 step procedure for extracting semantics from a document:
  - …the approach extracts semantic network from a document and tries to extract relevant part of the semantic network to represent summary
Knowledge Rich Summarization Example

1. Input document is split into sentences
2. Each sentence is deep-parsed
3. Name-entities are disambiguated:
   - Determining that 'George Bush' == 'Bush' == 'U.S. president'
4. Performing Anaphora resolution:
   - Pronouns are connected with named-entities
5. Extracting of **Subject-Predicate-Object** triples
6. Constructing a **graph** from triples
7. Each triple in the graph is described with features for learning
8. Using machine learning train a model for classification of triples into the summary
9. Generate a summary graph from selected triples
10. From the summary graph generate textual summary document

Tom went to town. In a bookstore he bought a large book.

**NLPWin**

Tom went to town. In a bookstore he [Tom] bought a large book.

**WordNet**
Training of summarization model

- A model was trained deciding which **Subject-Predicate-Object** triple belongs into the target summary

---

**Document Semantic network**

**Summary semantic network**
Cracks appeared in the U.N. trade embargo against Iraq. The State Department reports that Cuba and Romania have struck oil deals with Iraq as others attempt to trade with Baghdad in defiance of the sanctions. Iran has agreed to exchange food and medicine for Iraqi oil. Saddam has offered developing nations free oil if they send their tankers to pick it up. Thus far, none has accepted.

Japan, accused of responding too slowly to the Gulf crisis, has promised $2 billion in aid to countries hit hardest by the Iraqi trade embargo. President Bush has promised that Saddam's aggression will not succeed.
Text Segmentation
Text Segmentation

- **Problem**: divide text that has no given structure into segments with similar content
- Find story boundaries, determine what stories go with one another, and discover when something new (unforeseen) has happened
- **Example applications**:
  - topic tracking in news (spoken news, e.g., CNN)
  - identification of topics in large, unstructured text databases
Text Segmentation (cont.)

Application in Information Retrieval:

- Performing query similarity measures against sections of document as supposed to the whole document
- When displaying the search result, you can display the most relevant portion of the document to the query. Similar to the way Google displays its search results
Hearst Algorithm for Text Segmentation

- **Algorithm**
  - **Initial segmentation**
    - Divide a text into equal blocks of k words
  - **Similarity Computation**
    - compute similarity between $m$ blocks on the right and the left of the candidate boundary
  - **Boundary Detection**
    - place a boundary where similarity score reaches local minimum

- …the approach can be defined either as optimization problem or as sliding window

©Marko Grobelnik and Dunja Mladenic
Supervised Learning
Document Categorization Task

- **Given**: set of documents labeled with content categories
- **The goal**: to build a model which would automatically assign right content categories to new unlabeled documents.

Content categories can be:
- unstructured (e.g., Reuters) or
- structured (e.g., Yahoo, DMOz, Medline)
Example learning algorithm: Perceptron

Input:
- set of documents $D$ in the form of (e.g., TFIDF) numeric vectors
- each document has label +1 (positive class) or -1 (negative class)

Output:
- linear model $w_i$ (one weight per word from the vocabulary)

Algorithm:
- Initialize the model $w_i$ by setting word weights to 0
- Iterate through documents N times
  - For document $d$ from $D$
    - // Using current model $w_i$ classify the document $d$
    - if $\sum (d_i * w_i) >= 0$ then classify document as positive
    - else classify document as negative
    - if document classification is wrong then
      - // adjust weights of all words occurring in the document
      - $w_{t+1} = w_t + \text{sign(true-class)} * \text{Beta} \ (\text{input parameter Beta}>0)$
      - // where $\text{sign(positive)} = 1$ and $\text{sign(negative)} = -1$
Text Categorization into hierarchy of categories

- There are several hierarchies (taxonomies) of textual documents:
  - Yahoo, DMOz, Medline, …

- Different people use different approaches:
  - …series of hierarchically organized classifiers
  - …set of independent classifiers just for leaves
  - …set of independent classifiers for all nodes
Yahoo! hierarchy (taxonomy)

- human constructed hierarchy of Web-documents
- exists in several languages
- easy to access and regularly updated
- captures most of the Web topics
- English version includes over 2M pages categorized into 50,000 categories
- contains about 250Mb of HTML files

http://www.dmoz.org/
http://dir.search.yahoo.com/

http://www.googleguide.com/directory.html
CALL FOR PAPERS

Fourth Computational Natural Language Learning Workshop

CoNLL-2000

Lisbon, September 14, 2000


CoNLL is the yearly workshop organized by SIGNLL, the Association for Computational Linguistics Special Interest Group on Natural Language Learning.

The meeting will be held in conjunction with ICGI-2000, the International Conference on Grammar Inference (http://vinci.inesc.pt/icgi-2000) and the Learning Language in Logic workshop (http://www.lti.fccn/LLL-2000) in Lisbon on Thursday, September 14, 2000, and will feature a shared task competition about learning of chunking. There will be joint sessions with ICGI-2000 and the LLL workshop on topics of common interest. Previous CoNLL meetings were held in Madrid, Sydney, and Bergen.

We invite submissions of abstracts on all aspects of computational natural language learning, including

- Computational models of human language acquisition
- Computational models of the origins and evolution of language
- Machine learning methods applied to natural language processing tasks (speech processing, phonology, morphology, syntax, semantics, discourse processing, language engineering applications)
  - Symbolic learning methods (Rule Induction and Decision Tree Learning, Lazy Learning, Inductive Logic Programming, Analytical Learning, Transformation-based Error-driven Learning)
  - Biologically-inspired methods (Neural Networks, Evolutionary Computing)
  - Statistical methods (Bayesian Learning, HMM, maximum entropy, SNOW, Support Vector Machines)
  - Reinforcement Learning
  - Active learning, ensemble methods, meta-learning
- Computational Learning Theory analyses of language learning
- Empirical and theoretical comparisons of language learning methods
- Models of induction and analogy in Linguistics

A special session of the workshop will be devoted to a shared task: the identification of phrases (syntactic constituents) with machine learning methods, a task called chunking.
Some predicted categories

<table>
<thead>
<tr>
<th>Rank</th>
<th>Probability</th>
<th>Word [Weight]</th>
<th>Category Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>LANGUAGE [0.0714]</td>
<td>/Computers_and_Internet/Software/Natural_Language_Processing/</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>LANGUAGE [0.0714]</td>
<td>/Computers_and_Internet/Software/Natural_Language_Processing/</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>NATURAL [0.0714]</td>
<td>/Computers_and_Internet/Internet/World_Wide_Web/Information_and_Documentation/</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>LANGUAGE [0.0429]</td>
<td>/Computers_and_Internet/Internet/World_Wide_Web/Information_and_Documentation/</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>PROCESSING [0.0036]</td>
<td>/Computers_and_Internet/Supercomputing_and_Parallel_Computing/</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>NATURAL [0.001]</td>
<td>/Computers_and_Internet/Supercomputing_and_Parallel_Computing/</td>
</tr>
<tr>
<td>4</td>
<td>0.99</td>
<td>GROUP [0.0087]</td>
<td>/Computers_and_Internet/Mobile_Computing/</td>
</tr>
<tr>
<td>6</td>
<td>0.99</td>
<td>GROUP [0.0041]</td>
<td>/Computers_and_Internet/Information_and_Documentation/Product_Reviews/Buyers_Guides/Software/</td>
</tr>
<tr>
<td>6</td>
<td>0.99</td>
<td>GROUP [0.0056]</td>
<td>/Computers_and_Internet/Software/</td>
</tr>
<tr>
<td>7</td>
<td>0.98</td>
<td>SEPTEMBER [0.0059]</td>
<td>/Computers_and_Internet/Software/Programming_Tools/Object_Oriented_Programming/Conferences/</td>
</tr>
<tr>
<td>8</td>
<td>0.98</td>
<td>SEPTEMBER [0.0087]</td>
<td>/Computers_and_Internet/Conventions_and_Conferences/</td>
</tr>
<tr>
<td>9</td>
<td>0.97</td>
<td>GROUP [0.0055]</td>
<td>/Computers_and_Internet/Software/</td>
</tr>
<tr>
<td>10</td>
<td>0.97</td>
<td>LEARNING [0.0022]</td>
<td>/Computers_and_Internet/Internet/Information_and_Documentation/</td>
</tr>
<tr>
<td>11</td>
<td>0.95</td>
<td>SEPTEMBER [0.0004]</td>
<td>/Computers_and_Internet/Communications_and_Networking/Conferences/</td>
</tr>
<tr>
<td>12</td>
<td>0.95</td>
<td>SPECIAL [0.0121]</td>
<td>/Computers_and_Internet/Internet/World_Wide_Web/Conferences/Past_Events/</td>
</tr>
<tr>
<td>13</td>
<td>0.93</td>
<td>PROCESSING [0.0356]</td>
<td>/Computers_and_Internet/Supercomputing_and_Parallel_Computing/Conferences/</td>
</tr>
<tr>
<td>14</td>
<td>0.92</td>
<td>MAXIMUM [0.0019]</td>
<td>/Computers_and_Internet/Software/Peripherals/Modems/</td>
</tr>
<tr>
<td>15</td>
<td>0.92</td>
<td>SUSSION [0.0857]</td>
<td>/Computers_and_Internet/Internet/World_Wide_Web/Announcement_Services/Rbots/</td>
</tr>
</tbody>
</table>
Content categories

- For each content category generate a separate classifier that predicts probability for a new document to belong to its category
Unsupervised Learning
Document Clustering

- Clustering is a process of finding natural groups in the data in an unsupervised way (no class labels are pre-assigned to documents)
- Key element is similarity measure
  - In document clustering, cosine similarity is most widely used
- Most popular clustering methods are:
  - EM/K-Means clustering
  - Hierarchical clustering
  - ...
K-Means clustering algorithm

- **Given:**
  - set of documents (e.g., TFIDF vectors),
  - distance measure (e.g., cosine)
  - $K$ (number of groups)

- **Determining k**
  - Domain knowledge can be used to help determine $k$
  - Different values of $k$ can be experimented with to determine the best value
  - Other learning methods can be used to learn $k$
K-Means clustering algorithm (cont.)

- **For each** of $K$ groups, initialize its centroid with a random document

- **While** not converging
  - Each document is assigned to the **nearest** group (represented by its **centroid**: center of gravity or average of its points)
  - For each group calculate new centroid

- **Termination conditions** (at convergence)
  - Reach a predefined number of iterations
  - No changes on doc assignments to clusters btw iterations
  - No changes on centroids btw iterations

©Marko Grobelnik and Dunja Mladenic
Expectation-Maximization (EM) Algorithm

- The EM algorithm is a general template for a family of algorithms
  - Currently very popular and widely used in NLP and machine learning

- EM can be seen as a version of k-means clustering

From Patrick Cash
EM Algorithm

- Model
  - Parameters: k points representing cluster centers
  - Hidden structure: for each data point, which center generated it?

- Two steps
  - Expectation step: Use current parameters to reconstruct hidden structure
  - Maximization step: Use that hidden structure to re-estimate parameters

From Patrick Cash
EM Example Application – K-means

- **k-means**
  - K-means can be seen as a special case of EM where the mean of the distribution is the only variable
  - E step: estimate cluster membership using distance metric
  - M step: move seeds to new cluster centers

From Patrick Cash
Problems with EM

- EM can be very sensitive to initialization
  - Clustering can get stuck in local minima
  - Other clustering algorithms can be used for initialization
- EM convergence can be slow

From Patrick Cash
Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of unlabeled examples.

![Dendrogram representation of animal classification]

- One option to produce a hierarchical clustering is recursive application of a partitional clustering algorithm to produce a hierarchical clustering.

From Manning and Raghavan
Hierarchical Agglomerative Clustering (HAC)

- Assumes a similarity function for determining the similarity of two instances.
- Starts with all instances in a separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

From Manning and Raghavan
A Dendogram: Hierarchical Clustering

- Dendrogram: Decomposes data objects into a several levels of nested partitioning (tree of clusters).

- Clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.
HAC Algorithm

Start with all instances in their own cluster. Until there is only one cluster:

Among the current clusters, determine the two clusters, $c_i$ and $c_j$, that are most similar.

Replace $c_i$ and $c_j$ with a single cluster $c_i \cup c_j$

From Manning and Raghavan
As clusters *agglomerate*, docs likely to fall into a hierarchy of “topics” or concepts.

From Manning and Raghavan
Properties of hierarchical and EM/K-Means clustering

Hierarchical Clustering:
- Preferable for detailed data analysis
- Provides more information than flat clustering
- Less efficient than flat (for n objects, n X n similarity matrix required)

EM/K-Means Clustering:
- Preferable if efficiency is a consideration or data sets are very large

From Patrick Cash
Latent Semantic Indexing

- Term-document matrices are very large
- But the number of topics that people talk about is small (in some sense)
  - Clothes, movies, politics, …
- Can we represent the term-document space by a lower dimensional latent space?
What it is

- From term-doc matrix $A$, we compute the approximation $A_k$.
- There is a row for each term and a column for each doc in $A_k$.
- Thus docs live in a space of $k << r$ dimensions.
  - These dimensions are not the original axes.
- But why?

From Manning and Raghavan
Problems with Lexical Semantics

- Ambiguity and association in natural language
  - **Polysemy**: Words often have a multitude of meanings and different types of usage (*more severe in very heterogeneous collections*).
  - The vector space model is unable to discriminate between different meanings of the same word.

\[
\text{sim}_{\text{true}}(d, q) < \cos(\angle(\vec{d}, \vec{q}))
\]
Problems with Lexical Semantics

- **Synonymy**: Different terms may have an **identical or a similar meaning** (weaker: words indicating the same topic).

- No associations between words are made in the vector space representation.

\[
\text{sim}_{\text{true}}(d, q) > \cos(\angle(d, q))
\]
Polysemy and Context

- Document similarity on single word level: polysemy and context

- contribution to similarity, if used in 1st meaning, but not if in 2nd

From Manning and Raghavan
Latent Semantic Indexing (LSI)

- Perform a **low-rank approximation** of **document-term matrix** (typical rank **100-300**)

**General idea**

- Map documents (and terms) to a **low-dimensional** representation.
- Design a mapping such that the low-dimensional space reflects **semantic associations** (latent semantic space).
- Compute document similarity based on the **inner product** (i.e., angle) in this **latent semantic space**

From Manning and Raghavan

©Marko Grobelnik and Dunja Mladenic
Goals of LSI

- Similar terms map to similar location in low dimensional space
- Noise reduction by dimension reduction
Latent Semantic Analysis

- **Latent semantic space**: illustrating example

![Diagram of latent semantic space](image)

*courtesy of Susan Dumais*

From Manning and Raghavan
Impact to clustering

- What does this have to do with clustering?
- Intuition: Dimension reduction through LSI brings together “related” axes in the vector space.

From Manning and Raghavan
Summary: Latent Semantic Indexing

- LSI is a statistical technique that attempts to estimate the hidden content structure within documents:
  - ...it uses linear algebra technique Singular-Value-Decomposition (SVD)
  - ...it discovers statistically most significant co-occurrences of terms
## LSI Example

### Original document-term mantrix

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
</tr>
</thead>
<tbody>
<tr>
<td>cosmonaut</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>astronaut</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>moon</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>car</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>truck</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Rescaled document matrix, Reduced into two dimensions

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim 1</td>
<td>-</td>
<td>1.62</td>
<td>-</td>
<td>0.60</td>
<td>-</td>
<td>0.97</td>
</tr>
<tr>
<td>Dim 2</td>
<td>-</td>
<td>0.46</td>
<td>-</td>
<td>0.84</td>
<td>-</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### High correlation although d2 and d3 don’t share any word

### Correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d2</td>
<td>0.8</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d3</td>
<td>0.4</td>
<td>0.9</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td>0.5</td>
<td>-0.2</td>
<td>-0.6</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d5</td>
<td>0.7</td>
<td>0.2</td>
<td>-0.3</td>
<td>0.9</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>d6</td>
<td>0.1</td>
<td>-0.5</td>
<td>-0.9</td>
<td>0.9</td>
<td>0.7</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Visualization
Why visualizing text?

- ...to have a top level view of the topics in the corpora
- ...to see relationships between the topics and objects in the corpora
- ...to understand better what’s going on in the corpora
- ...to show highly structured nature of textual contents in a simplified way
- ...to show main dimensions of highly dimensional space of textual documents
- ...because it’s fun!
Example: Visualization of PASCAL project research topics (based on published papers abstracts)

- theory
- natural language processing
- kernel methods
- multimedia processing
...typical way of doing text visualization

- By having text in the sparse vector Bag-of-Words representation, we usually perform a kind of clustering algorithm to identify structure, which is then mapped into 2D or 3D space.

- ...other typical way of visualization of text is to find frequent co-occurrences of words and phrases which are visualized, e.g., as graphs.

- Typical visualization scenarios:
  - Visualization of document collections
  - Visualization of search results
  - Visualization of document timeline
Graph based visualization

The sketch of the algorithm:

1. Documents are transformed into the bag-of-words sparse-vectors representation
   - Words in the vectors are weighted using TFIDF
2. K-Means clustering algorithm splits the documents into K groups
   - Each group consists from similar documents
   - Documents are compared using cosine similarity
3. K groups form a graph:
   - Groups are nodes in graph; similar groups are linked
   - Each group is represented by characteristic keywords
4. Using simulated annealing draw a graph
Graph based visualization of 1700 IST project descriptions into 2 groups
Graph based visualization of 1700 IST project descriptions into 3 groups
Graph based visualization of 1700 IST project descriptions into 10 groups
Graph based visualization of 1700 IST project descriptions into 20 groups
Tiling based visualization

The sketch of the algorithm:
1. Documents are transformed into the bag-of-words sparse-vectors representation
   - Words in the vectors are weighted using TFIDF
2. Hierarchical top-down two-wise K-Means clustering algorithm builds a hierarchy of clusters
   - The hierarchy is an artificial equivalent of hierarchical subject index (Yahoo like)
3. The leaf nodes of the hierarchy (bottom level) are used to visualize the documents
   - Each leaf is represented by characteristic keywords
   - Each hierarchical binary split splits recursively the rectangular area into two sub-areas
Tiling based visualization of 1700 IST project descriptions into 2 groups
Tiling based visualization of 1700 IST project descriptions into 3 groups
Tiling based visualization of 1700 IST project descriptions into 4 groups.
Tiling based visualization of 1700 IST project descriptions into 5 groups
Tiling visualization (up to 50 documents per group) of 1700 IST project descriptions (60 groups)
TextArc – visualization of word occurrences

http://www.textarc.org/
NewsMap – visualization of news articles

http://newsmap.jp/
Document Atlas – visualization of document collections and their structure

http://docatlas.ijs.si
Information Extraction

(slides borrowed from William Cohen’s Tutorial on IE)
Example: Extracting Job Openings from the Web

foodscience.com-Job2

JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.htm
OtherCompanyJobs: foodscience.com-Job1
Peter Norvig Robert Wilensky University of California, Berkeley Computer Science Thirteenth International Conference on Computational Linguistics, Volume 3
ResearchIndex

Abstract: This paper critically evaluates three recent abduction models: those of Charniak and Goldman (1989), Hobbs, Stickel, Martin, and Edwards (1988), and Ng and Mooney (1990). These three models share the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abduction approach, and some tentative solutions.

Context of citations to this paper: More

... (break slight modification of the one given in Ng and Mooney, 1990) The new definition remedies the anomaly reported in Norvig and Wilensky (1990) of occasionally preferring spurious interpretations of greater depth. Table 1: Empirical Results Comparing Coherence and...

... costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wilensky (1990). The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 1) a. Only literals...
What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...
What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
IE in Context

Create ontology

Spider

Filter by relevance

IE

Segment
Classify, Associate
Cluster

Train extraction models

Load DB

Database

Query, Search

Data mine

Document collection

Label training data
Typical approaches to IE

- Hand-built rules/models for extraction
  - ...usually extended regexp rules
  - ...GATE system from U. Sheffield (http://gate.ac.uk/)

- Machine learning used on manually labelled data:
  - Classification problem on sliding window
    - ...examples are taken from sliding window
    - ...models classify short segments of text such as title, name, institution, ...
    - ...limitation of sliding window because it does not take into account sequential nature of text
  - Training stochastic finite state machines (e.g. HMM)
    - ...probabilistic reconstruction of parsing sequence