Outline

- **Text-Mining**
  - How to deal with text data on various levels?

- **Link-Analysis**
  - How to analyze graphs in the Web context?

- **Semantic-Web**
  - How semantics fits into the picture?

- **Wrap-up**
  - …what did we learn and where to continue?
Text-Mining

How to deal with text data on various levels?

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...”
Which areas are active in Text Processing?

Why 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
- Concepts are difficult to visualize
- High dimensionality - tens or hundreds of thousands of features
Why Text is Easy? (M.Hearst 97)

- **Highly redundant data**
  - …most of the methods 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

---

Levels of Text Processing 1/4

- **Word Level**
  - Words Properties
  - Stop-Words
  - Stemming
  - Frequent N-Grams
  - Thesaurus (WordNet)
- Sentence Level
- Document Level
- Document-Collection Level
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
- Natural language dependent – examples:
  - English: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ALSO, ...
  - Slovenian: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ... 
  - Croatian: A, AH, AHA, ALI, AKO, BEZ, DA, IPAK, NE, NEG, ...
## Example of stop word removal

<table>
<thead>
<tr>
<th>Original text</th>
<th>After the stop-words removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Systems Asia Web - provides research, IS-related commercial</td>
<td>Information Systems Asia Web provides research IS-related</td>
</tr>
<tr>
<td>materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region.</td>
<td>commercial materials interaction research sponsorship interested corporations focus Asia Pacific region.</td>
</tr>
<tr>
<td>Survey of Information Retrieval - guide to IR, with an emphasis on web-based projects. Includes a glossary, and pointers to interesting papers.</td>
<td>Survey Information Retrieval guide IR emphasis web-based projects Includes glossary pointers interesting papers</td>
</tr>
</tbody>
</table>

---

## Stemming (I)

- 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 (II)

- For English – publicly available algorithms give good results
  - Most widely used is Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- In morphologically rich languages (e.g. Slovenian) there are 10-20 different forms correspond to the same word:
  - E.g. (“to laugh” in Slovenian): smej, smejal, smejala, smejale, smejali, smejal, smejati, smejejo, smejet, smejeve, smeješ, smejemo, smejiš, smeje, smejoč, smejš, smeje, smejva

Stemming (III): 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
### Stemming (IV): Rules automatically obtained for Slovenian language

- **Machine Learning** applied on Multext-East dictionary ([http://nl.ijs.si/ME/](http://nl.ijs.si/ME/))
- **Two example rules:**
  - Remove the ending “OM” if 3 last char is any of HOM, NOM, DOM, SOM, POM, BOM, FOM. For instance, ALAHOM, AMERICANOM, BENJAMINOM, BERLINOM, ALFREDOM, BEOGRADOM, DICKENSON, JEZUSOM, JOSIPOM, OLIMPOM,... but not ALEKSANDROM (ROM -> ER)
  - Replace CEM by EC. For instance, ARABCEM, BAVARCEM, BOVCEM, EVROPEJCEM, GORENJCEM, ... but not FRANCEM (remove EM)

---

### Phrases in the form of frequent N-Grams

- **Simple way for generating phrases are frequent n-grams:**
  - N-Gram is a sequence of n consecutive words (e.g. “machine learning” is 2-gram)
  - “Frequent n-grams” are the ones which appear in all observed documents MinFreq or more times
- **N-grams are interesting because of the simple and efficient dynamic programming algorithm:**
  - **Given:**
    - Set of documents (each document is a sequence of words),
    - MinFreq (minimal n-gram frequency),
    - MaxNGramSize (maximal n-gram length)
  - **for Len = 1 to MaxNgramSize do**
    - Generate candidate n-grams as sequences of words of size Len using frequent n-grams of length Len-1
    - Delete candidate n-grams with the frequency less then MinFreq
Generation of frequent n-grams for 50,000 documents from Yahoo Directory

Example document after stop word removal, stemming and generating frequent n-grams

Original text on the Yahoo Web page:
2. UK Only
3. Idomeneus - IR & DB repository - These pages mostly contain IR related resources such as test collections, stop lists, stemming algorithms, and links to other IR sites.
4. University of Glasgow - Information Retrieval Group - information on the resources and people in the Glasgow IR group.
5. Centre for Intelligent Information Retrieval (CIIR).
6. Information Systems Asia Web - provides research, IS-related commercial materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region.
7. Seminar on Cataloging Digital Documents
8. Survey of Information Retrieval - guide to IR, with an emphasis on web-based projects. Includes a glossary, and pointers to interesting papers.
9. University of Dortmund - Information Retrieval Group

Document represented by n-grams:
1. "REFERENCE LIBRARIES LIBRARY INFORMATION SCIENCE (#3 LIBRARY INFORMATION SCIENCE) INFORMATION RETRIEVAL (#2 INFORMATION RETRIEVAL)"
2. "UK"
3. "IR PAGES IR RELATED RESOURCES COLLECTIONS LISTS LINKS IR SITES"
4. "UNIVERSITY GLASGOW INFORMATION RETRIEVAL (#2 INFORMATION RETRIEVAL) GROUP INFORMATION RESOURCES (#2 INFORMATION RESOURCES) PEOPLE GLASGOW IR GROUP"
5. "CENTRE INFORMATION RETRIEVAL (#2 INFORMATION RETRIEVAL)"
6. "INFORMATION SYSTEMS ASIA WEB RESEARCH COMMERCIAL MATERIALS RESEARCH ASIA PACIFIC REGION"
7. "CATALOGING DIGITAL DOCUMENTS"
8. "INFORMATION RETRIEVAL (#2 INFORMATION RETRIEVAL) GUIDE IR EMPHASIS INCLUDES GLOSSARY InterESTING"
9. "UNIVERSITY INFORMATION RETRIEVAL (#2 INFORMATION RETRIEVAL) GROUP"
WordNet – database of lexical relations

- WordNet is 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>

WordNet – excerpt from the graph (borrowed slide)
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>

Levels of Text Processing 2/4

- Word Level
  - Words Properties
  - Stop-Words
  - Stemming
  - Frequent N-Grams
  - Thesaurus (WordNet)

- Sentence Level
- Document Level
- Document-Collection Level
Levels of Text Processing 3/4

- Word Level
- Sentence Level
- **Document Level**
  - Summarization
  - Text Segmentation
- Document-Collection Level

---

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 methods adopt linear weighting model – each text unit (sentence) is assessed by the following formula:
  - Weight(U) = LocationInTheText(U) + CuePhrase(U) + Statistics(U) + AdditionalPresence(U)
- ...lot of heuristics and tuning of parameters (also with ML)
- …output consists from topmost text units (sentences)
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 was evaluated on “Document Understanding Conference” test set of documents and their summaries
  - ...the approach extracts semantic network from a document and tries to extract relevant part of the semantic network to represent summary
  - Results achieved 70% recall of and 25% precision on extracted Subject-Predicate-Object triples
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

Training of summarization model

- Train a model which tells which **Subject-Predicate-Object** triple belongs into the summary
- Support Vector Machine (SVM) used for training
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.

Cracks appeared Tuesday in the U.N. trade embargo against Iraq as Eastern bloc nations sought to circumvent the economic noose around their country. Romania, denied that claim Tuesday, calling it “absolutely false and without foundation.” Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of

stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

States, Virgil Constantinescu, denied that claim Tuesday, calling it “absolutely false and without foundation.” Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.

Meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that the U.S. embargo on Iraq, President Bush, session of Congress as a non-military activity to avoid television news that Saddam Hussein will fail to make his case of stand up to aggression, and we will. . . . Bush, who noted that the U.S. military only lost in the South Asian disaster, said the U.S. would suffer a "major blow" to its reputation and credibility if it did not respond to Saddam Hussein’s aggression.
Text Segmentation

- **Problem**: divide text that has no given structure into segments with similar content
- **Example applications**:
  - topic tracking in news (spoken news)
  - identification of topics in large, unstructured text databases
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

Levels of Text Processing 4/4

- Word Level
- Sentence Level
- Document Level
- **Document-Collection Level**
  - Text Representation
  - Document Similarity
  - Feature Selection
  - Supervised learning
  - Semisupervised learning
  - Unsupervised learning
  - Visualization
  - Information Extraction
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”.

This representation is referred to also as “Bag-Of-Words” or “Vector-Space-Model”.

---

**Flat representation – vector space model**

- 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”.

---
Bag-of-words document representation

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:

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

- \(tf(w)\) – 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.
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 340,763 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.
Similarity between document vectors

- Each document is represented as a vector of weights $D = \langle x \rangle$
- Cosine similarity (dot product) is the most widely used similarity measure between two document vectors
  - …calculated cosine of the angle between document vectors
  - …efficient to calculate (sum of products of intersecting words)
  - …similarity value between 0 (different) and 1 (the same)

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

Feature Selection
Example of the problem

Data set

- Five Boolean features
- \( C = F_1 \lor F_2 \)
- \( F_3 = \neg F_2 \), \( F_5 = \neg F_4 \)
- Optimal subset: \( \{F_1, F_2\} \) or \( \{F_1, F_3\} \)
- Optimization in space of all feature subsets (2^5 possibilities)

(borrowed from tutorial on genomics [Yu 2004])

Search for feature subset

An example of search space ([John & Kohavi 1997])
Feature subset selection

- commonly used search strategies:
  - **forward selection**
    FSubset=(); greedily add features one at a time
  - **forward stepwise selection**
    FSubset=(); greedily add or remove features one at a time
  - **backward elimination**
    FSubset=AllFeatures; greedily remove features one at a time
  - **backward stepwise elimination**
    FSubset=AllFeatures; greedily add or remove features one at a time
  - **random mutation**
    FSubset=RandomFeatures;
    greedily add or remove randomly selected feature one at a time
    stop after a given number of iterations

Approaches to feature subset selection

- **Filters** - evaluation function independent of the learning algorithm
- **Wrappers** - evaluation using model selection based on the machine learning algorithm
- **Embedded approaches** - feature selection during learning
- **Simple Filters** - assume feature independence (used for problems with large number of features, eg. text classification)
Filtering

Evaluation independent of ML algorithm

Filters: Distribution-based [Koller & Sahami 1996]

- Idea: select a minimal subset of features that keeps class probability distribution close to the original distribution
  - $P(C|\text{FeatureSet})$ is close to $P(C|\text{AllFeatures})$

1. start with all the features
2. use backward elimination to eliminate a predefined number of features

- evaluation: the next feature to be deleted is obtained using Cross-entropy measure
### Filters: Relief [Kira & Rendell 1992]

Evaluation of a feature subset
1. represent examples using the feature subset
2. on a random subset of examples calculate average difference in distance from
   - the nearest example of the same class and the nearest example of the different class
   \[
   \delta(\text{Ex}_1, \text{Ex}_2) = \sum_{j=1}^{[\text{Subset}]} \text{diff}(\text{Ex}_i(F), \text{Ex}_j(F))
   \]

- \(F\) discrete \(\{0; \text{Ex}_1(F) = \text{Ex}_2(F)\} \quad \text{F cont.} \frac{|\text{Ex}_1(F) - \text{Ex}_2(F)|}{\text{max}(F) - \text{min}(F)}\)
- some extensions, empirical and theoretical analysis in [Robnik-Sikonja & Kononenko 2003]

### Filters: FOCUS [Almallim & Dietterich 1991]

Evaluation of a feature subset
1. represent examples using the feature subset
2. count conflicts in class value (two examples with the same feature values and different class value)

- Search: all the (promising) subsets of the same (increasing) size are evaluated until a sufficient (no conflicts) subset is found
- assumes existence of a small sufficient subset --> not appropriate for tasks with many features
- some extensions of the algorithm use heuristic search to avoid evaluating all the subsets of the same size
Illustration of FOCUS

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F4</th>
<th>F5</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Filters: Random [Liu & Setiono 1996]

Evaluation of a feature subset
1. represent examples using the feature subset
2. calculate the inconsistency rate (the average difference between the number of examples with equal feature values and the number of examples among them with the locally, most frequent class value)
3. select the smallest subset with inconsistency rate below the given threshold

- Search: random sampling to search the space of feature subsets
- evaluate the predetermined number of subsets
- noise handling by setting the threshold > 0
- if threshold = 0, then the same evaluation as in FOCUS
Filters: MDL-based [Pfahringer 1995]

Evaluation using Minimum Description Length
1. represent examples using the feature subset
2. calculate MDL of a simple decision table representing examples
   - Search: start with random feature subset and add or delete a feature, one at a time
   - performs at least as well as the wrapper approach applied on the simple decision tables and scales up better to large number of training examples

Wrapper

Evaluation uses the same ML algorithm that is used after the feature selection
Wrappers: Instance-based learning

Evaluation using instance-based learning
- represent examples using the feature subset
- estimate model quality using cross-validation

- Search [Aha & Bankert 1994]
  - start with random feature subset
  - use beam search with backward elimination
- Search [Skalak 1994]
  - start with random feature subset
  - use random mutation

Wrappers: Decision tree induction

Evaluation using decision tree induction
- represent examples using the feature subset
- estimate model quality using cross-validation

- Search [Bala et al 1995], [Cherkauer & Shavlik 1996]
  - using genetic algorithm
- Search [Caruana & Freitag 1994]
  - adding and removing features (backward stepwise elimination)
  - additionally, at each step removes all the features that were not used in the decision tree induced for the evaluation of the current feature subset
Metric-based model selection

Idea: poor models behave differently on training and other data
Evaluation using machine learning algorithm
- represent examples using the feature subset
- generate model using some ML algorithm
- estimate model quality comparing the performance of two models on training and on unlabeled data, chose the largest subset that satisfies triangular inequality with all the smaller subsets

\[ d_i(f_j, f_{ji}) \leq d_i(f_j, P_{ji}) + d_i(f_{ji}, P_{ji}) \ldots 0 \leq j < i \]

- Combine metric and cross-validation [Bengio & Chapados 2003]
  - based on their disagreement on testing examples (higher disagreement means lower trust to cross-validation)
  - Intuition: cross-validation provides good results but has high variance and should benefit from a combination with model selection having with lower variance

Embedded

Feature selection as integral part of model generation
Embedded

- at each iteration of the incremental optimization of the model
  - use a fast gradient-based heuristic to find the most promising feature [Perkins et al 2003]
- Idea: features that are relevant to the concept should affect the generalization error bound of non-linear SVM more than irrelevant features
  - use backward elimination based on the criteria derived from generalization error bounds of the SVM theory (the weight vector norm or, using upper bounds of the leave-one-out error) [Rakotomamonjy 2003]

Embedded: in filters [Cardie 1993]

Use embedded feature selection as preprocessing
- evaluation and search using the process embedded in decision tree induction
- the final feature subset contains only the features that appear in the induced decision tree
- used for learning using Nearest Neighbor algorithm
Why feature selection on text?

- Text in bag-of-words representation could be understood as a table of data records where:
  - Number of records in the table corresponds to the number of documents in the corpus
  - Each document has $w$ variables where $w$ is the size of corpus vocabulary (in the range of 100,000 words for bigger corpora)
  - ...most of the values in the table are zero, meaning that the table is sparse
- Most of the classical analysis methods (from machine learning and statistics) are not able to deal with large number of variables and with sparse representation:
  - ...feature selection helps in reducing the number of variables exploiting high redundancy property within the text
Feature subset selection on text data – commonly used methods

- Simple filtering using some scoring measure to evaluate individual feature
  - supervised measures:
    - information gain, cross entropy for text (information gain on only one feature value), mutual information for text
  - supervised measures for binary class
    - odds ratio (target class vs. the rest), bi-normal separation
  - unsupervised measures:
    - term frequency, document frequency
- Simple filtering using embedded approach to score the features
  - scoring measure equal to weights in the normal of hyperplane of linear SVM trained on all the features [Brank et al 2002]
  - learning using linear SVM, Perceptron, Naïve Bayes

Scoring individual feature

- InformationGain: \[ \sum_{F \in \mathcal{F}} P(F) \sum_{C \in \text{pos, neg}} P(C|F) \log \frac{P(C|F)}{P(C)} \]
- CrossEntropyTxt: \[ P(W) \sum_{C \in \text{pos, neg}} P(C|W) \log \frac{P(C|W)}{P(C)} \]
- MutualInfoTxt: \[ \sum_{C \in \text{pos, neg}} P(C) \log \frac{P(W|C)}{P(W)} \]
- OddsRatio: \[ \log \frac{P(W|pos) \times (1 - P(W|neg))}{(1 - P(W|pos)) \times P(W|neg)} \]
- Frequency: \[ Freq(W) \]
- Bi-NormalSeparation: \[ F^{-1}(P(W|pos)) - F^{-1}(P(W|neg)) \]

F - Normal distribution cumulative probability function
Influence of feature selection on the classification performance

- Some ML algorithms are more sensitive to the feature subset than other
  - Naïve Bayes on document categorization sensitive to the feature subset
  - Linear SVM has embedded weighting of features that partially compensates for feature selection

Example of the best features

| Odds Ratio | feature score | [P(F|pos), P(F|neg)] |
|------------|---------------|-----------------------|
| IR         | 5.28          | [0.075, 0.0004]       |
| INFORMATION RETRIEVAL | 5.13 | [0.075, 0.0007] |
| RETRIEVAL  | 4.77          | [0.03, 0.0003]        |
| ASIA       | 4.32          | [0.03, 0.0004]        |
| PACIFIC    | 4.02          | [0.015, 0.0003]       |
| INTERESTING| 4.02          | [0.015, 0.0003]       |
| EMPHASIS   | 4.02          | [0.015, 0.0003]       |
| GROUP      | 3.64          | [0.045, 0.0012]       |
| MASSACHUSETTS | 3.46 | [0.015, ...] |
| COMMERCIAL | 3.46          | [0.015, 0.0005]       |
| REGION     | 3.1           | [0.015, 0.0007]       |

| Information Gain | feature score | [P(F|pos), P(F|neg)] |
|------------------|---------------|----------------------|
| LIBRARY          | 0.46          | [0.015, 0.091]       |
| PUBLIC           | 0.23          | [0, 0.034]           |
| PUBLIC LIBRARY   | 0.21          | [0, 0.029]           |
| UNIVERSITY       | 0.21          | [0.045, 0.028]       |
| LIBRARIES        | 0.197         | [0.015, 0.026]       |
| INFORMATION      | 0.17          | [0.119, 0.021]       |
| REFERENCES       | 0.117         | [0.015, 0.012]       |
| RESOURCES        | 0.11          | [0.029, 0.0102]      |
| COUNTY           | 0.096         | [0, 0.0089]          |
| INTERNET         | 0.091         | [0, 0.00826]         |
| LINKS            | 0.091         | [0.015, 0.00819]     |
| SERVICES         | 0.089         | [0, 0.0079]          |
Supervised Learning

Document categorization

Machine learning

labeled documents

unlabeled document

document category
(label)
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)
  - structured (e.g., Yahoo, DMOz, Medline)

---

Document categorization

- **Machine learning**
- **labeled documents**
- **unlabeled document**
- **document category (label)**
### Algorithms for learning document classifiers

- **Popular algorithms for text categorization:**
  - Support Vector Machines
  - Logistic Regression
  - Perceptron algorithm
  - Naive Bayesian classifier
  - Winnow algorithm
  - Nearest Neighbour
  - ....

---

### Perceptron learning algorithm

**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 $\sum(d_w * w_j) >= 0$ 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
      - $w_{new} = w_j + \text{sign(true-class)} * \text{Beta}$ (input parameter Beta>0)
      - // where sign(positive) = 1 and sign(negative) = -1
Measuring success – Model quality estimation

\[
\begin{align*}
\text{Precision}(M,\text{targetC}) &= P(\text{targetC}|\text{targetC}) \\
\text{Recall}(M,\text{targetC}) &= P(\text{targetC}|\text{targetC}) \\
\text{Accuracy}(M) &= \sum_i P(C_i) \times \text{Precision}(M,C_i) \\
F_\beta(M,\text{targetC}) &= \frac{(1+\beta^2) \text{Precision}(M,\text{targetC}) \times \text{Recall}(M,\text{targetC})}{\beta^2 \text{Precision}(M,\text{targetC}) + \text{Recall}(M,\text{targetC})}
\end{align*}
\]

- Classification accuracy
- Break-even point (precision=recall)
- F-measure (precision, recall = sensitivity)

Reuters dataset – Categorization to flat categories

- Documents classified by editors into one or more categories
- Publicly available set of Reuter news mainly from 1987:
  - 120 categories giving the document content, such as: earn, acquire, corn, rice, jobs, oilseeds, gold, coffee, housing, income,...
- …from 2000 is available new dataset of 830,000 Reuters documents available for research
Distribution of documents (Reuters-21578)

Top 20 categories of Reuter news in 1987-91

Example of Perceptron model for Reuters category “Acquisition”

<table>
<thead>
<tr>
<th>Feature</th>
<th>Positive Class Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAKE</td>
<td>11.5</td>
</tr>
<tr>
<td>MERGER</td>
<td>9.5</td>
</tr>
<tr>
<td>TAKEOVER</td>
<td>9</td>
</tr>
<tr>
<td>ACQUIRE</td>
<td>9</td>
</tr>
<tr>
<td>ACQUIRED</td>
<td>8</td>
</tr>
<tr>
<td>COMPLETES</td>
<td>7.5</td>
</tr>
<tr>
<td>OWNERSHIP</td>
<td>7.5</td>
</tr>
<tr>
<td>SALE</td>
<td>7.5</td>
</tr>
<tr>
<td>OWNERSHIP</td>
<td>7.5</td>
</tr>
<tr>
<td>BUYOUT</td>
<td>7</td>
</tr>
<tr>
<td>ACQUISITION</td>
<td>6.5</td>
</tr>
<tr>
<td>UNDISCLOSED</td>
<td>6.5</td>
</tr>
<tr>
<td>BUYS</td>
<td>6.5</td>
</tr>
<tr>
<td>ASSETS</td>
<td>6</td>
</tr>
<tr>
<td>BID</td>
<td>6</td>
</tr>
<tr>
<td>BP</td>
<td>6</td>
</tr>
<tr>
<td>DIVISION</td>
<td>5.5</td>
</tr>
</tbody>
</table>
SVM, Perceptron & Winnow
text categorization performance on
Reuters-21578 with different representations

Comparison of algorithms

Comparison on using SVM on stemmed 1-grams
with related results
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 (we use English)
- 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
CALL FOR PAPERS

Fourth Computational Natural Language Learning Workshop

CoNLL-2000

Lisboa, September 14, 2000

http://www.uva.ar.es/coell00/CoNLL-2000

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

The meeting will be held in conjunction with IJCAI-2000, the International Conference on Artificial Intelligence (http://www.ijcai.org/2000) and the Language in a Logic workshop (http://www.uva.ar.es/coell00/CoNLL-2000) in Lisboa on Tuesday, September 14, 2000. We will have a single task competition evaluating systems. There will be one winner with the best system and the LLL workshop on topics of common interest. Previous CoNLL meetings were held in Madrid (1994), and Rome.

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

- Computational models of human language acquisition
- Computational models of the human cognitive processes of language comprehension (speech processing, phonology, morphology, syntax, semantics, text generation, language engineering and applications)
- Symbolic language models (rule induction and decision tree learning, Logic Learning, Inductive Logic Programming, Inductive Learning, Transformative Inductive Learning)
- Statistical language models (Markov Networks, Minimum Error Rate)
- Statistical methods (Maximum Entropy, Maximum Entropy Markov Models, SVMs, HMMs) (Graphical Models, Conditional Random Fields, CRFs)
- Interactive Learning
- Active learning, experimental methods, testing learning
- Computational Learning Theory analysis of language learning
- Empirical and theoretical comparisons of language learning methods
- Models of cohesion and analogy in natural language

A special session of the workshop will be devoted to a special task: the identification of discourse markers (cohesion) with machine learning methods, and data mining.
For each content category generate a separate classifier that predicts probability for a new document to belong to its category.
Considering promising categories only
(classification by Naive Bayes)

\[ P(C \mid Doc) = \frac{P(C) \prod_{W \in Doc} P(W \mid C)^{Freq(W, Doc)}}{\sum_i P(C_i) \prod_{W \in Doc} P(W_i \mid C_i)^{Freq(W_i, Doc)}} \]

- Document is represented as a set of word sequences \( W \)
- Each classifier has two distributions: \( P(W\mid pos) \), \( P(W\mid neg) \)
- Promising category:
  - calculated \( P(pos\mid Doc) \) is high meaning that the classifier has \( P(W\mid pos)>0 \) for at least some \( W \) from the document (otherwise, the prior probability is returned, \( P.neg \) is about 0.90)

Summary of experimental results

<table>
<thead>
<tr>
<th>Domain</th>
<th>probability</th>
<th>rank</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertain.</td>
<td>0.96</td>
<td>16</td>
<td>0.44</td>
<td>0.80</td>
</tr>
<tr>
<td>Arts</td>
<td>0.99</td>
<td>10</td>
<td>0.40</td>
<td>0.83</td>
</tr>
<tr>
<td>Computers</td>
<td>0.98</td>
<td>12</td>
<td>0.40</td>
<td>0.84</td>
</tr>
<tr>
<td>Education</td>
<td>0.99</td>
<td>9</td>
<td>0.57</td>
<td>0.65</td>
</tr>
<tr>
<td>Reference</td>
<td>0.99</td>
<td>3</td>
<td>0.51</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Semisupervised Learning

Active Learning

- We use this method whenever hand-labeled data are rare or expensive to obtain.
- Interactive method
- Requests only labeling of "interesting" objects.
- Much less human work needed for the same result compared to arbitrary labeling examples.

![Diagram of Active Learning](image)
Some approaches to Active Learning

- **Uncertainty sampling** (efficient)
  - select example closest to the decision hyperplane (or the one with classification probability closest to $P=0.5$) (Tong & Koller 2000 Stanford)

- **Maximum margin ratio change**
  - select example with the largest predicted impact on the margin size if selected (Tong & Koller 2000 Stanford)

- **Monte Carlo Estimation of Error Reduction**
  - select example that reinforces our current beliefs (Roy & McCallum 2001, CMU)

- **Random sampling** as baseline

Experimental evaluation (using F1-measure) of the four listed approaches shown on three categories from Reuters-2000 dataset

- average over 10 random samples of 5000 training (out of 500k) and 10k testing (out of 300k) examples
- the last two methods are rather time consuming, thus we run them for including the first 50 unlabelled examples
- experiments show that active learning is especially useful for unbalanced data

Category with very unbalanced class distribution having 2.7% of positive examples

Uncertainty seems to outperform MarginRatio
Illustration of Active learning

- starting with one labeled example from each class (red and blue)
- select one example for labeling (green circle)
- request label and add re-generate the model using the extended labeled data

Illustration of linear SVM model using
- arbitrary selection of unlabeled examples (random)
- active learning selecting the most uncertain examples (closest to the decision hyperplane)
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:
  - K-Means clustering (flat, hierarchical)
  - Agglomerative hierarchical clustering
  - EM (Gaussian Mixture)
  - ...
K-Means clustering algorithm

- **Given:**
  - set of documents (e.g. TFIDF vectors),
  - distance measure (e.g. cosine)
  - \( K \) (number of groups)
- **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)
  - For each group calculate new centroid (group mass point, average document in the group)

Example of hierarchical clustering (bisection k-means)
**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**

<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>

**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>Dim 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dim 2</td>
<td>-</td>
<td>0.46</td>
<td>0.84</td>
<td>0.30</td>
<td>1.00</td>
<td>0.35</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>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 0.9</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td>0.5 -0.2</td>
<td>-0.6 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d5</td>
<td>0.7 0.2</td>
<td>-0.3 0.9</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d6</td>
<td>0.1 -0.5</td>
<td>-0.9 0.9</td>
<td>0.7 1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Correlation matrix**

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

…typical way of doing text visualization

- By having text in the sparse vector Bag-of-Words representation we usually perform some kind of **clustering algorithm** identify structure which is then mapped into 2D or 3D space (e.g. using MDS)
- …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 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:

- Quantum
- Smart
- Security
- Month
- Card
- Learning

- SMEs
- Service
- Month
- Platform
- Services
- Network

- Data
- Speech
- Learning
- System
- NM
- Embedded

- Commerce
- Training
- Mobile
- Multimedia
- Knowledge
- Learning

Tiling based visualization of 1700 IST project descriptions into 5 groups:

- Quantum
- Smart
- Security
- Month
- Card
- Learning

- SMEs
- Service
- Month
- Business

- Health
- IST
- Agent
- UMTS
- Network

- Commerce
- Training
- Mobile
- Multimedia
- Knowledge
- Learning
WebSOM

- Self-Organizing Maps for Internet Exploration
  - ...algorithm that automatically organizes the documents onto a two-dimensional grid so that related documents appear close to each other
  - ... based on Kohonen's Self-Organizing Maps
WebSOM visualization

Explanation of the symbols on the map
- acorn - comp.sys.acorn.hardware
- amiga - comp.sys.amiga.hardware
- books - tech, sci, fiction, rec, arts, books
- cdfom - comp.publish.cdfom.hardware
- compilers - comp.compilers
- fuzzy - comp.ai.fuzzy
- genetic - comp.ai.genetic
- hp - comp.sys.hp.hardware
- humor - rec.humor
- lang.atlant - comp.lang.atlant
- lang.ml - comp.lang.ml
- linux - comp.os.linux.hardware
- lisp - comp.lang.lisp
- mac - comp.sys.mac.hardware
- mac.storage - comp.sys.mac.hardware.storage
- movies - movies
- music - music
- os - comp.os.windows.nt.setup, hardware
- pc.chips - comp.sys.ibm.pc.hardware.chips
- pc.comms - comp.sys.ibm.pc.hardware.comms
- pc.storage - comp.sys.ibm.pc.hardware.storage
- pc.video - comp.sys.ibm.pc.hardware.video
- philosophy - philosophy
- plant - blot, biology, plant
- prolog - comp.lang.prolog
- sc.lang - sci, lang
- smalltalk - comp.lang.smalltalk

ThemeScape

- Graphically displays images based on word similarities and themes in text
- Themes within the document spaces appear on the computer screen as a relief map of natural terrain
  - The mountains indicate where themes are dominant - valleys indicate weak themes
  - Themes close in content will be close visually based on the many relationships within the text spaces
  - Algorithm is based on K-means clustering

http://www.pnl.gov/infoviz/technologies.html
ThemeScape Document visualization

ThemeRiver

topic stream visualization

- The ThemeRiver visualization helps users identify time-related patterns, trends, and relationships across a large collection of documents.
- The themes in the collection are represented by a "river" that flows left to right through time.
- The theme currents narrow or widen to indicate changes in individual theme strength at any point in time.

http://www.pnl.gov/infoviz/technologies.html
Kartoo.com – visualization of search results

http://kartoo.com/

TextArc – visualization of word occurrences

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

http://www.marumushi.com/apps/newsmap/newsmap.cfm

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
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…

What is “Information Extraction”

As a family of techniques: Information Extraction = segmentation + classification + clustering + association

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
aka “named entity extraction”
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation
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...
Typical approaches to IE

- Hand-built rules/models for extraction
  - …usually extended regexp rules
- 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

Link-Analysis

How to analyze graphs in the Web context?
What is Link Analysis?

- Link Analysis is exploring associations between the objects
  - …most characteristic for the area is graph representation of the data
  - Category of graphs which attract recently the most interest are the ones which are generated by some social process (social networks) – this would include web

- Synonyms for Link Analysis or at least very related areas are Graph Mining, Network Analysis, Social Network Analysis

- In the next slides we’ll present some of the typical definitions, ideas and algorithms

What is Power Law?

- Power law describes relations between the objects in the network
  - …it is very characteristic for the networks generated within some kind of social process
  - …it describes scale invariance found in many natural phenomena (including physics, biology, sociology, economy and linguistics)

- In Link Analysis we usually deal with power law distributed graphs
Power-Law on the Web

- In the context of Web the power-law appears in many cases:
  - Web pages sizes
  - Web page connectivity
  - Web connected components’ size
  - Web page access statistics
  - Web Browsing behavior
- Formally, power law describing web page degrees are:

\[
\begin{align*}
\Pr(\text{out-degree is } k) & \propto 1/k^\alpha_{\text{out}} \\
\Pr(\text{in-degree is } k) & \propto 1/k^\alpha_{\text{in}}
\end{align*}
\]

(This property has been preserved as the Web has grown)
Small World Networks

- Empirical observation for the Web-Graph is that the diameter of the Web-Graph is small relative to the size of the network
  - …this property is called “Small World”
  - …formally, small-world networks have diameter exponentially smaller than the size
- By simulation it was shown that for the Web-size of 1B pages the diameter is approx. 19 steps
  - …empirical studies confirmed the findings
Example of Small World: project collaboration network

- The network represents collaboration between institutions on projects funded by European Union
  - ...there are 7886 organizations collaborating on 2786 projects
  - ...in the network, each node is an organization, two organizations are connected if they collaborate on at least one project

Small world properties of the collaboration network:

- **Main connected part** of the network contains 94% of the nodes
- **Max distance** between any two organizations is 7 steps ...
  meaning that any organization can be reached in up to 7 steps from any other organization
- **Average distance** between any two organizations is 3.15 steps (with standard deviation 0.38)
- 38% (2770) of organizations have avg. distance 3 or less

Connectedness of the most connected institution

- 1856 collaborations
- avg. distance is 1.95
- max. distance is 4
Connectedness of semi connected institution

- 179 collaborations
- avg. distance is 2.42
- max. distance is 4

Connectedness of min. connected institution

- 8 collaborations
- max. distance is 7
In November 1999 large scale study using AltaVista crawls in the size of over 200M nodes and 1.5B links reported “bow tie” structure of web links

- we suspect, because of the scale free nature of the Web, this structure is still preserved
Modeling the Web Growth

- Links/Edges in the Web-Graph are not created at random
  - ...probability that a new page gets attached to one of the more popular pages is higher than to a one of the less popular pages
  - Intuition: “rich gets richer” or “winners takes all”
  - Simple algorithm “Preferential Attachment Model” (Barabassi, Albert) efficiently simulates Web-Growth

“Preferential Attachment Model” Algorithm

- \( M_0 \) vertices (pages) at time 0
- At each time step new vertex (page) is generated with \( m \leq M_0 \) edges to \( m \) random vertices
  - ...probability for selection a vertex for the edge is proportional to its degree
- ...after \( t \) time steps, the network has \( M_0 + t \) vertices (pages) and \( mt \) edges
  - ...probability that a vertex has connectivity \( k \) follows the power-law
Estimating importance of the web pages

- Two main approaches, both based on eigenvector identification on the adjacency matrix
  - Hubs and Authorities (HITS)
  - PageRank – used by Google

Hubs and Authorities

- Intuition behind HITS is that each web page has two natures:
  - ...being good content page (authority weight)
  - ...being good hub (hub weight)
- ...and the idea behind the algorithm:
  - ...good authority page is pointed to by good hub pages
  - ...good hub page is pointing to good authority pages
Hubs and Authorities

(Kleinberg 1998)

“Hubs and authorities exhibit what could be called a mutually reinforcing relationship”

Iterative relaxation:

\[
\begin{align*}
\text{Hub}(p) &= \sum_{q: p \rightarrow q} \text{Authority}(q) \\
\text{Authority}(p) &= \sum_{q: q \rightarrow p} \text{Hub}(q)
\end{align*}
\]

PageRank

- PageRank was developed by the founders of the Google in 1998
  - …its basic intuition is to calculate primal eigenvector of the graph adjacency matrix
  - …each page gets a value which nicely corresponds to the importance of the node within the network
  - PageRank can be computed effectively by an iterative procedure
Example Calculation

Average PR: 1.000

Semantic-Web

How semantics fits into the picture?
What is Semantic Web? (informal)

- Informal statements:
  - “…if the ordinary web is for mainly computer-to-human communication, then the semantic web aims primarily at computer-to-computer communication.
  - The idea is to establish infrastructure for dealing with common vocabularies.
  - The goal is to overcome surface syntax representation of the data and deal with “semantic” of the data.
    - …as an example, one should be able to make a “semantic link” from a database column with the name “ZIP-Code” and a GUI form with a “ZIP” field since they actually mean the same – they both describe the same abstract concept.
  - Semantic Web is mainly about integration and standards!

What is Semantic Web? (formal)

- Formal statement (from http://www.w3.org/2001/sw/):
  - “The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.”
  - “It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners.”
What is the link between Text-Mining, Link Analysis and Semantic Web?

- **Text-Mining, Link-Analysis** and other analytic techniques deal mainly with extracting and aggregating the information from raw data
  - ...they maximize the quality of extracted information
- **Semantic Web**, on the other hand, deals mainly with the integration and representation of the given data
  - ...it maximizes reusability of the given information
- **Both areas** are very much complementary and necessary for operational information engineering

Ontologies – central objects in SW

- Ontologies are central formal objects within Semantic Web
  - Ontologies have origin in philosophy, but within computer science they represent a data model that represents a domain and is used to reason about the objects in that domain and the relations between them
  - ...their main aim is to describe and represent an area of knowledge in a formal way
  - Most of the Semantic Web standards/languages (XML, RDF, OWL) are concerned with some level of ontological representation of the knowledge
**What is an ontology?**

<table>
<thead>
<tr>
<th>Abstract model of some domain</th>
<th>Consensual knowledge</th>
<th>explicit specification,</th>
<th>machine processable</th>
<th>Formal,</th>
</tr>
</thead>
<tbody>
<tr>
<td>concepts, properties, relations, functions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frank van Harmelen 2003: http://seminars.ijs.si/sek4

**Which elements represent an ontology?**

- An ontology typically consists of the following elements:
  - **Instances** – the basic or “ground level” objects
  - **Classes** – sets, collections, or types of objects
  - **Attributes** – properties, features, characteristics, or parameters that objects can have and share
  - **Relations** – ways that objects can be related to one another

- Analogies between ontologies and relational databases:
  - **Instances** correspond to **records**
  - **Classes** correspond to **tables**
  - **Attributes** correspond to **record fields**
  - **Relations** correspond to **relations between the tables**
Which levels Semantic Web is dealing with?

- The famous "Semantic Web Layer Cake" shows representation levels and related technologies

- Stack of Semantic Web Languages
  - XML (eXtended Markup Language)
    - Surface syntax, no semantics
  - XML Schema
    - Describes structure of XML documents
  - RDF (Resource Description Framework)
    - Datamodel for "relations" between "things"
  - RDF Schema
    - RDF Vocabulary Definition Language
  - OWL (Web Ontology Language)
    - A more expressive Vocabulary Definition Language
Bluffer’s guide to RDF (1/2)

- **Object -> Attribute -> Value** triples

- objects are **web-resources**
- Value is again an Object:
  - triples can be **linked**
  - data-model = **graph**

Bluffer’s guide to RDF (2/2)

- Every identifier is a URL
  = world-wide unique naming!
- Has XML syntax

```xml
<rdf:Description rdf:about="#pers05">
  <authorOf>ISBN...</authorOf>
</rdf:Description>
```

- Any statement can be an object
  - …graphs can be **nested**
OWL – Web Ontology Language

- The OWL (Web Ontology Language) is a markup language for publishing and sharing data using ontologies on the Web
  - …it is standardized at W3C (http://www.w3.org/TR/owl-guide/)
  - One of the main goals was to achieve high machine interpretability
  - …and higher expresivity then languages such as RDF and RDF Schema

OWL Layers

- **OWL Lite:**
  - Classification hierarchy
  - Simple constraints
- **OWL DL:**
  - Maximal expressiveness
  - While maintaining tractability
  - Standard formalisation
- **OWL Full:**
  - Very high expressiveness
  - Loosing tractability
  - Non-standard formalisation
  - All syntactic freedom of RDF (self-modifying)
Cyc …a little bit of historical context

- Older AI-ers probably still remember Cyc
  - …one of the boldest attempts in AI history to encode common sense knowledge in one KB
  - The project started in 1984 at Stanford
  - …in 1994 the company Cycorp was established
  - Till 2006 $78M were spend into the KB
  - Cyc gets available: OpenCyc & ResearchCyc
  - In 2005 Cyc KB gets opened and available for research
  - In 2006 Cyc-Europe was established (in Ljubljana)
…part of Cyc Ontology on Human Beings

Structure of Cyc Ontology

Knowledge Base Layers

Upper Ontology

Core Theories

Domain-Specific Theories

Facts (Database)
Structure of Cyc Ontology

Knowledge Base Layers

Upper Ontology: Abstract Concepts

Upper Ontology

Core Theories

Domain-Specific Theories

Facts (Database)

For all events a and b, a causes b implies a precedes b
Structure of Cyc Ontology

For all events $a$ and $b$, $a$ causes $b$ implies $a$ precedes $b$.

For any mammal $m$ and any anthrax bacteria $a$, $m$'s being exposed to $a$ causes $m$ to be infected by $a$.

John is a person infected by anthrax.
An example of Psychoanalyst’s Cyc taxonomic context

```
#Psychoanalyst (lexical representation: “psychoanalyst”, “psychoanalysts”)
| specialization-of #MedicalCareProfessional |
| specialization-of #HealthProfessional |
| specialization-of #Professional-Adult |
| specialization-of #Professional |
| specialization-of #Psychologist |
| specialization-of #Scientist |
| specialization-of #Researcher |
| specialization-of #PersonWithOccupation |
| specialization-of #Person |
| specialization-of #HomoSapiens |
| specialization-of #BiologicalSpecies |
| specialization-of #BiologicalTaxon |
| specialization-of #SomeSampleKindsOfMammal-Biology-Topic |
| specialization-of #AdultAnimal |
| specialization-of #Animal |
| specialization-of #SolidTangibleThing |
| specialization-of #SolidTangibleThing |
| specialization-of #StatesOfMatter-Material-Topic |
| specialization-of (#GraduateFn #University) |
| specialization-of (#Graduate #DegreeGrantingHigherEducationInstitution) |
```

Cyc Ontology Samples

- Next slides describe bits of Cyc Ontology on
  - Temporal Relations
  - Senses of “x is a physical part of y”
  - Senses of “x is physically in y”
  - Events and their performers (role types)
  - Organizations
  - Propositional Attitudes
  - Biology
  - Materials
  - Devices
  - Weather
Temporal relations

37 Relations Between Temporal Things

- #$temporalBoundsIntersect
- #$temporarilyIntersects
- #$startsAfterStartingOf
- #$endsAfterEndingOf
- #$startingDate
- #$temporallyContains
- #$temporallyCooriginating
- #$temporalBoundsContain
- #$temporalBoundsIdentical
- #$startsDuring
- #$overlapsStart
- #$sstartingPoint
- #$sstartingWith
- #$after

Senses of ‘Part’ relation

- #$parts
- #$intangibleParts
- #$subInformation
- #$subEvents
- #$physicalDecompositions
- #$physicalPortions
- #$physicalParts
- #$externalParts
- #$internalParts
- #$anatomicalParts
- #$constituents
- #$functionalPart
Senses of ‘In’ relation (1/3)

- Can the inner object leave by passing between members of the outer group?
  - Yes -- Try #Sin-Among

Senses of ‘In’ relation (2/3)

- Does part of the inner object stick out of the container?
  - None of it. -- Try #Sin-ContCompletely
  - Yes -- Try #Sin-ContPartially
  - No -- Try #Sin-ContClosed

- If the container were turned around could the contained object fall out?
  - Yes -- Try #Sin-ContOpen
Senses of ‘In’ relation (3/3)

Is it attached to the inside of the outer object?
- Yes -- Try #$connectedToInside

Can it be removed by pulling, if enough force is used, without damaging either object?
- No -- Try #$in-Snugly or #$screwedIn

Does the inner object stick into the outer object?
- Yes -- Try #$sticksInto

Event Types

- #$PhysicalStateChangeEvent
- #$TemperatureChangingProcess
- #$BiologicalDevelopmentEvent
- #$ShapeChangeEvent
- #$MovementEvent
- #$ChangingDeviceState
- #$GivingSomething
- #$DiscoveryEvent
- #$Cracking
- #$Carving
- #$Buying
- #$Thinking
- #$Mixing
- #$Singing
- #$CuttingNails
- #$PumpingFluid

... 11,000 more
A Few Transportation Event Types

- #$TransportationEvent
- #$ControllingATransportationDevice
- #$TransportWithMotorizedLandVehicle
- ($#SteeringFn #$RoadVehicle)
- #$TransporterCrashEvent
- #$VehicleAccident
- #$CarAccident
- #$Colliding
- #$IncurringDamage
- #$TippingOver
- #$Navigating
- #$EnteringAVehicle

Relations Between Events and Participants

- #$performedBy
- #$causes-Event
- #$objectPlaced
- #$objectOfStateChange
- #$outputsCreated
- #$inputsDestroyed
- #$assistingAgent
- #$beneficiary
- #$fromLocation
- #$toLocation
- #$deviceUsed
- #$driverActor
- #$damages
- #$vehicle
- #$providerOfMotiveForce
- #$transportees

...over 400 more.
Representation of a Sample Event

Here are some roles for Attack874

isa: TerroristAttack.
performedBy: JihadGroup.
deviceUsed: Bomb8388.
eventOccursAt: CityOfLondonEngland.
victim: Person9399.
victim: Person52666.
assistingAgent: AlQaeda.
objectsDestroyed: Structure2990.
objectsDestroyed: Vehicle523452.

Samples of Organization Roles

- #$governingBody
- #$parentCompany
- #$subOrgs-Command
- #$subOrgs-Permanent
- #$subOrgs-Temporary
- #$physicalQuarters
- #$hasHQinCountry
- #$officeInCountry
- #$memberTypes
- #$organizationHead
- #$PolicyFn
- #$mainProductType

+ inherited roles from generalisations
(e.g., #$startingTime, #$alsoKnownAs).
Emotions

Types of Emotions:
- #$Adulation
- #$Abhorrence
- #$Relaxed-Feeling
- #$Gratitude
- #$Anticipation-Feeling
- …over 120 of these

Predicates for Defining and Attributing Emotions:
- #$contraryFeelings
- #$appropriateEmotion
- #$actionExpressesFeeling
- #$feelsTowardsObject
- #$feelsTowardsPersonType

Propositional Attitudes

- #$goals
- #$intends
- #$desires
- #$hopes
- #$expects
- #$beliefs

- #$opinions
- #$knows
- #$rememberedProp
- #$perceivesThat
- #$seesThat
- #$tastesThat
Devices

- Over 4000 Specializations of #$PhysicalDevice
  - #$ClothesWasher
  - #$NuclearAircraftCarrier

- Vocabulary for Describing Device Functions
  - #$primaryFunction-DeviceType

- Device-specific Predicates
  - #$gunCaliber
  - #$speedOf

- Device States (40+)
  - #$DeviceOn
  - #$CockedState

Vehicular Transport Devices

- Over 800 Specializations of #$RoadVehicle
  - #$AcuraCar
  - #$SportUtilityVehicle
  - #$Humvee

- Over 100 Specializations of #$AutoPart
  - #$AutomobileTire
  - #$ShockAbsorber
  - #$Windshield

- Five Facets of #$RoadVehicle
  - #$RoadVehicleByChassisType
  - #$RoadVehicleTypeByBodyStyle
  - #$RoadVehicleTypeByModel
  - #$RoadVehicleTypeByPowerSource
  - #$RoadVehicleTypeByUse

- Specialized Predicates
  - #$highwayFuelConsumption
  - #$vehicleLoadClass
  - #$trafficableForVehicle
  - #$vehicle
Weather

- Weather Attributes
  - #$ClearWeather
  - #$Visibility
  - (#$LowAmountFn #$Raininess)
- Weather Objects
  - #$CloudInSky
  - #$SnowMob
- Weather Events
  - #$TornadoAsEvent
  - #$SnowProcess

“If it’s raining, carry an umbrella”

- the performer is a human being,
- the performer is sane,
- the performer can carry an umbrella; thus:
  - the performer is not a baby, not unconscious, not dead,
  - the performer is going to go outdoors now/soon,
- their actions permit them a free hand (e.g., not wheelbarrowing)
- their actions wouldn’t be unduly hampered by it (e.g., marathon-running)
- the wind outside is not too fierce (e.g., hurricane strength)
- the time period of the action is after the invention of the umbrella
- the culture is one that uses umbrellas as a rain- (not just sun-)protection device.
- the performer has easy access to an umbrella; thus:
  - not too destitute, not someone who lives where it practically never rains,
  - not at the office/theater/... caught without an umbrella
- the performer is going to be unsheltered for some period of time
- the more waterproof their clothing, the gentler the rain, and
- the warmer the air, the longer that time period
- the performer will not be wet anyway (e.g., swimming)
The beautiful world of Web X.X versions
(...)a trial to put all of them on one slide

<table>
<thead>
<tr>
<th>Description</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web 1.0</strong></td>
<td>Static HTML pages</td>
</tr>
<tr>
<td>(web as we first learned it)</td>
<td>HTML, HTTP</td>
</tr>
<tr>
<td><strong>Web 1.5</strong></td>
<td>Dynamic HTML content</td>
</tr>
<tr>
<td>(web as we know it)</td>
<td>Client side (JavaScript, DHTML, Flash, ...), server side (CGI, PHP, Perl, ASP/.NET, JSP, ...)</td>
</tr>
<tr>
<td><strong>Web 2.0</strong></td>
<td>Integration on all levels, collaboration, sharing</td>
</tr>
<tr>
<td>vocabularies</td>
<td>weblogs, social bookmarking, social tagging, wikis, podcasts, RSS feeds, many-to-many publishing, web services, ...</td>
</tr>
<tr>
<td>(web as it is being sold)</td>
<td>URI, XML, RDF, OWL, ...</td>
</tr>
<tr>
<td><strong>Web 3.0</strong></td>
<td>...adding meaning to semantics - AI dream revival</td>
</tr>
<tr>
<td>(web as we would need it)</td>
<td>Closest area of a research would be &quot;common sense reasoning&quot; and the &quot;Cyc system&quot;</td>
</tr>
</tbody>
</table>

Web 2.0 –is there any new quality?

- IMHO, with “Web 2.0” the Web community became **really aware** of the importance of the global collaborative work
  - ...next step in globalization of the Web
  - **Bottom-up** “social networking” seems to nicely complement the traditional **top-down** schema design approaches

Visualization of Web 2.0 typical vocabulary
Web 2.0 – the current hype

Google search volume of “Web 2.0” vs. “semantic web”

![Google Trends graph]

What about Web 4.0? 😊

- Citation from some Intel blog:
  - “…Web 4.0 is the impending state at which all information converges into a great ball of benevolent self-aware light, and solves every problem from world peace to …”

- Ultimate stage in web development…
  - …will prevent Web 5.0 to happen since everything will be resolved already by Web 4.0.
Wrap-up

...what did we learn and where to continue?

References to some TM&LA Books
References to some SW Books

References to Conferences

- Information Retrieval: SIGIR, ECIR
- Machine Learning/Data Mining: ICML, ECML/PKDD, KDD, ICDM, SDM
- Computational Linguistics: ACL, EACL, NAACL
- Semantic Web: ISWC, ESWS
References to some of the TM & LA & SW workshops (available online)


Some of the Products

- Authonomy
- ClearForest
- Megaputer
- SAS/Enterprise-Miner
- SPSS - Clementine
- Oracle - ConText
- IBM - Intelligent Miner for Text
- Microsoft – Analysis Services
**Major Databases & Text-Mining**

- **Oracle** – includes some functionality within the database engine (e.g. classification with SVM, clustering, …)
- **IBM DB2** – text mining appears as a database extender accessible through several SQL functions
  - …a lot of functionality is included in UIMA environment
- **Microsoft SQL Server** – text processing is available as a preprocessing stage in Data-Transformation Services module

**Final Remarks**

- In the future we can expect stronger integration and **bigger overlap** between TM, IR, NLP and SW…
- …the technology and solutions will try to **capture deeper semantics** within the text,
- …**integration of various** data sources (where text is just one of the modalities) is becoming increasingly important.