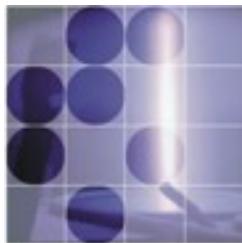


How to Represent Text? ...from Characters to Logic

Marko Grobelnik (marko.grobelnik@ijs.si)
Jozef Stefan Institute (<http://www.ijs.si/>)
Slovenia, Europe



Berkley, Aug 8th 2011

Outline

- ▶ Some Initial thoughts
 - ...quick example why representation matters
- ▶ How we represent Text?
 - Big picture
 - Levels of representation:
 - Lexical
 - Syntactic
 - Semantic
- ▶ Further references
 - ...events, books, videos

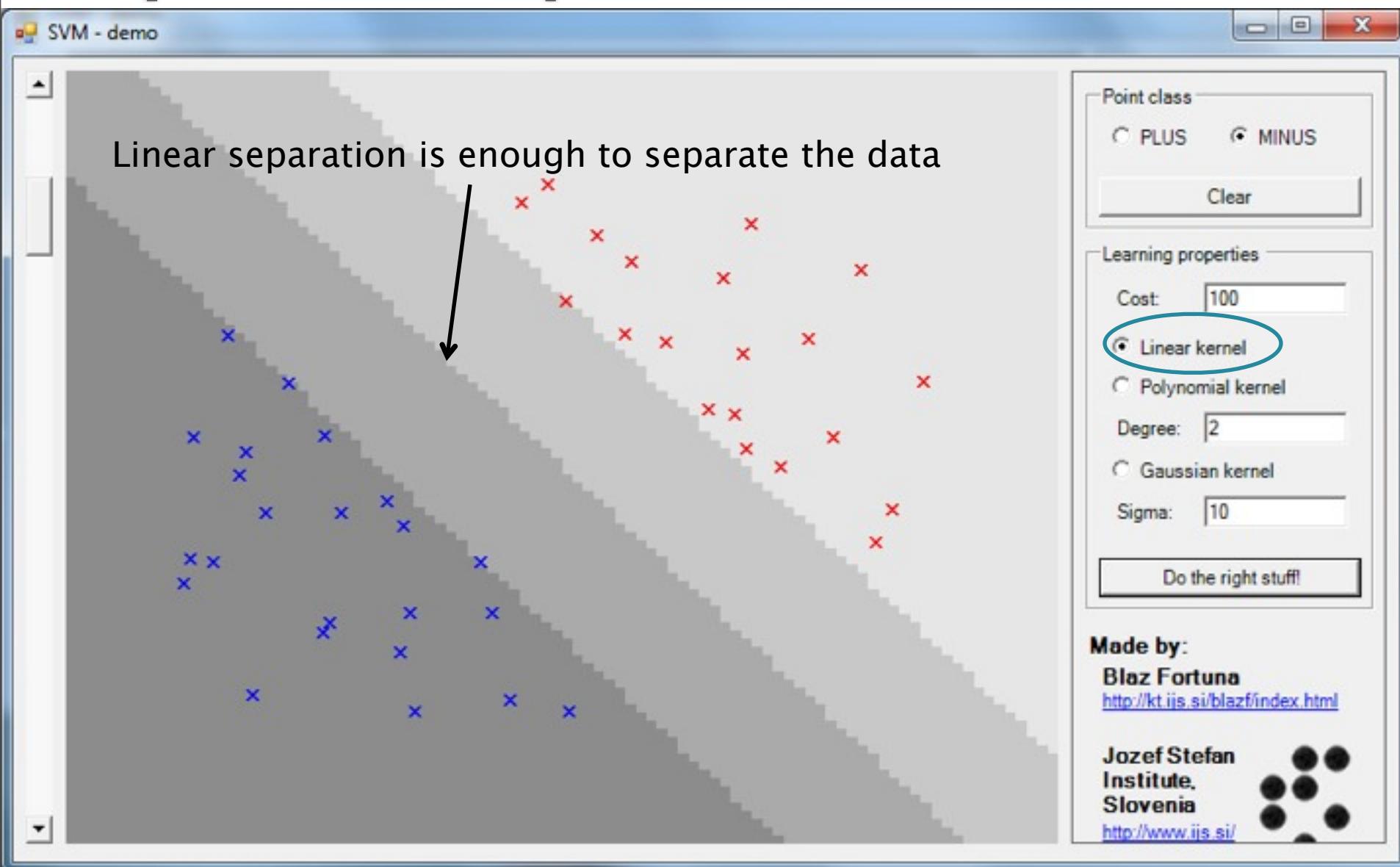
Some Initial thoughts

- ▶ Two major steps in Machine learning:
 - Choosing representation (feature engineering)
 - Modeling (statistics+optimization)
- ▶ ...typically people do modeling well and often ignore data representation
- ▶ But...
 - ...**good representation with bad algorithm** gives typically better results than **good algorithm with bad representation**

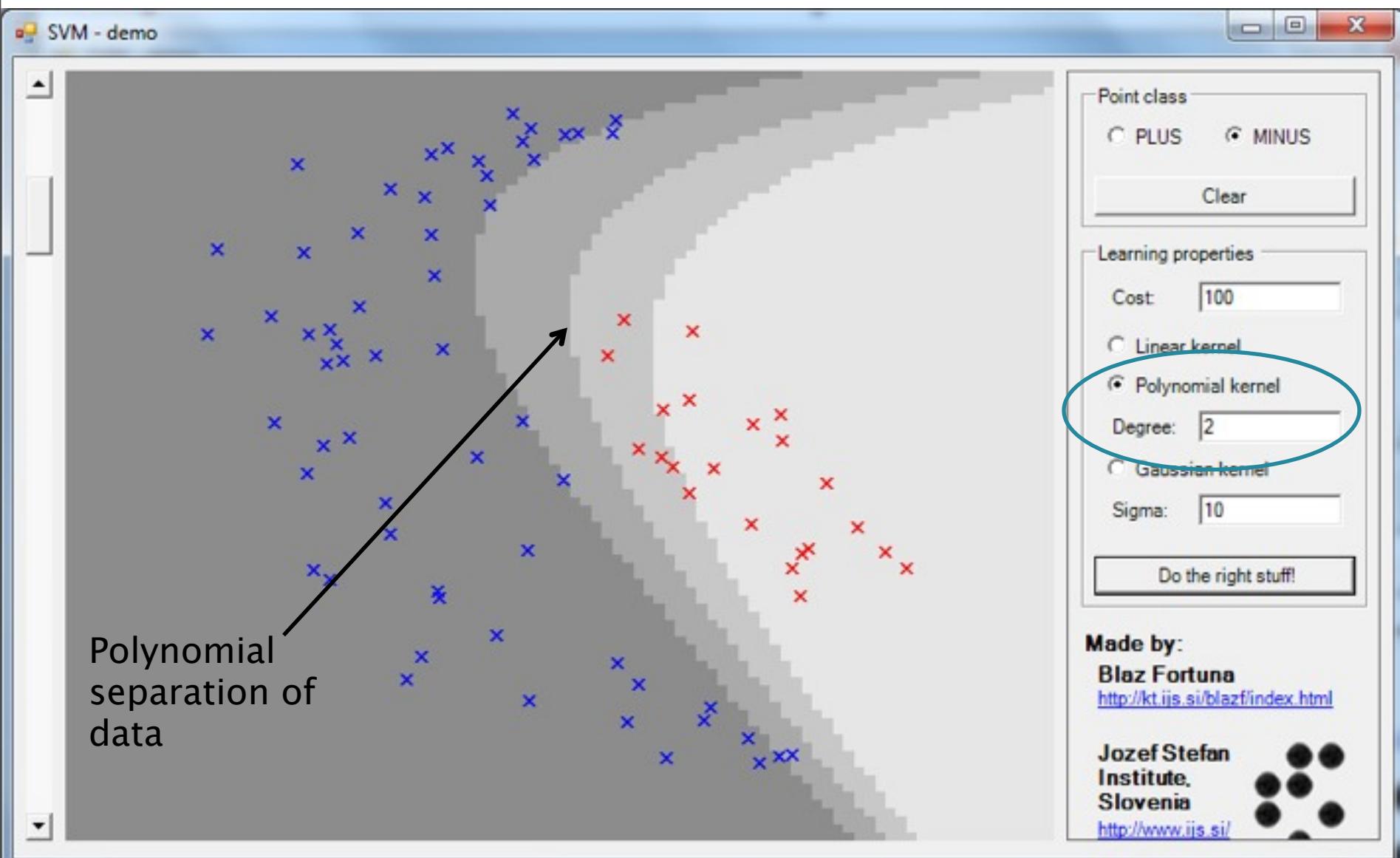
Quick example: Why representation matters?

- ▶ Selection of kernels when using SVM (Support Vector Machine) equals selecting the right representation for data

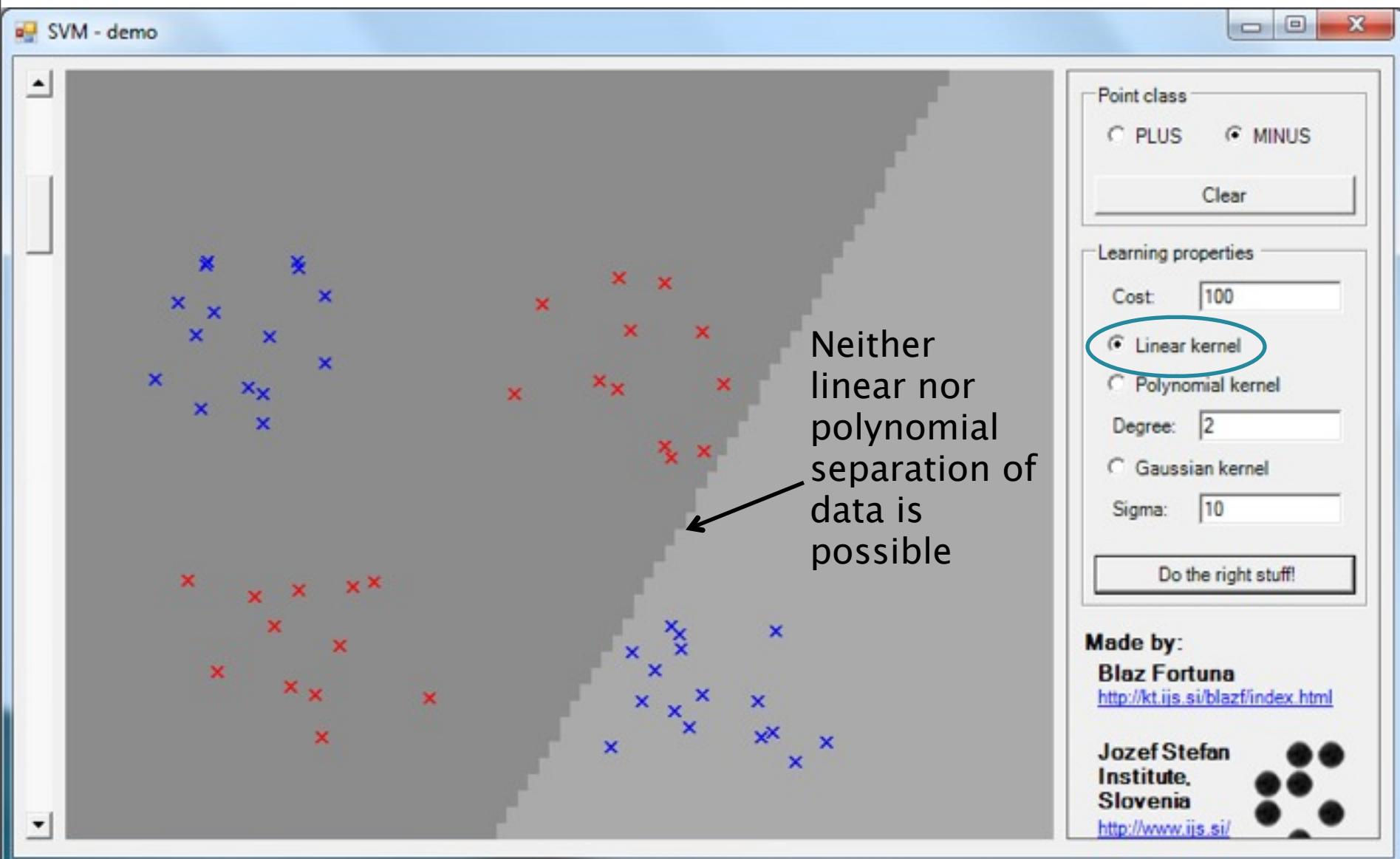
Easy decision problems require simple data representation



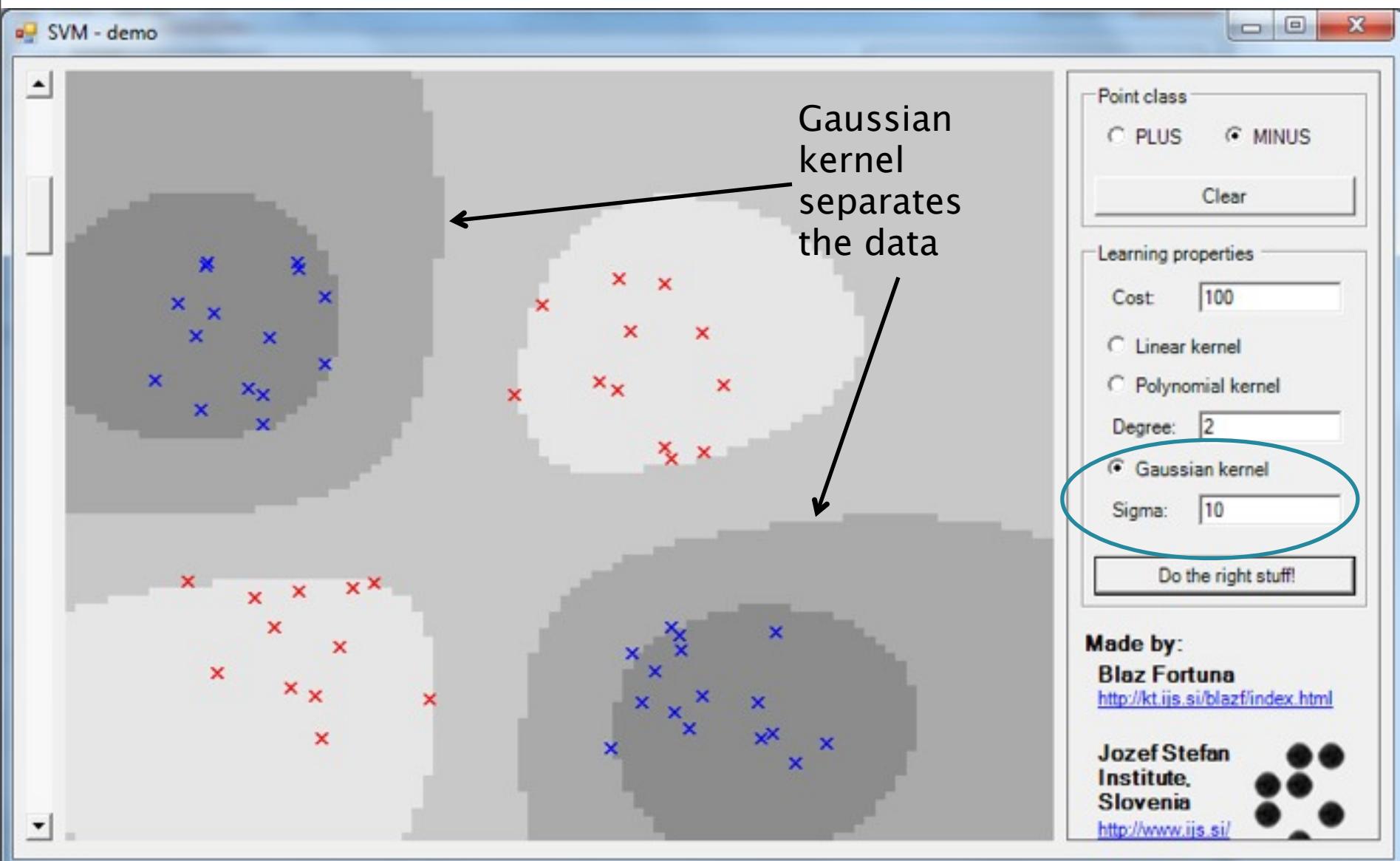
Harder decision problems require better data representation



Hard decision problems require sophisticated data representation



Hard decision problems require sophisticated data representation

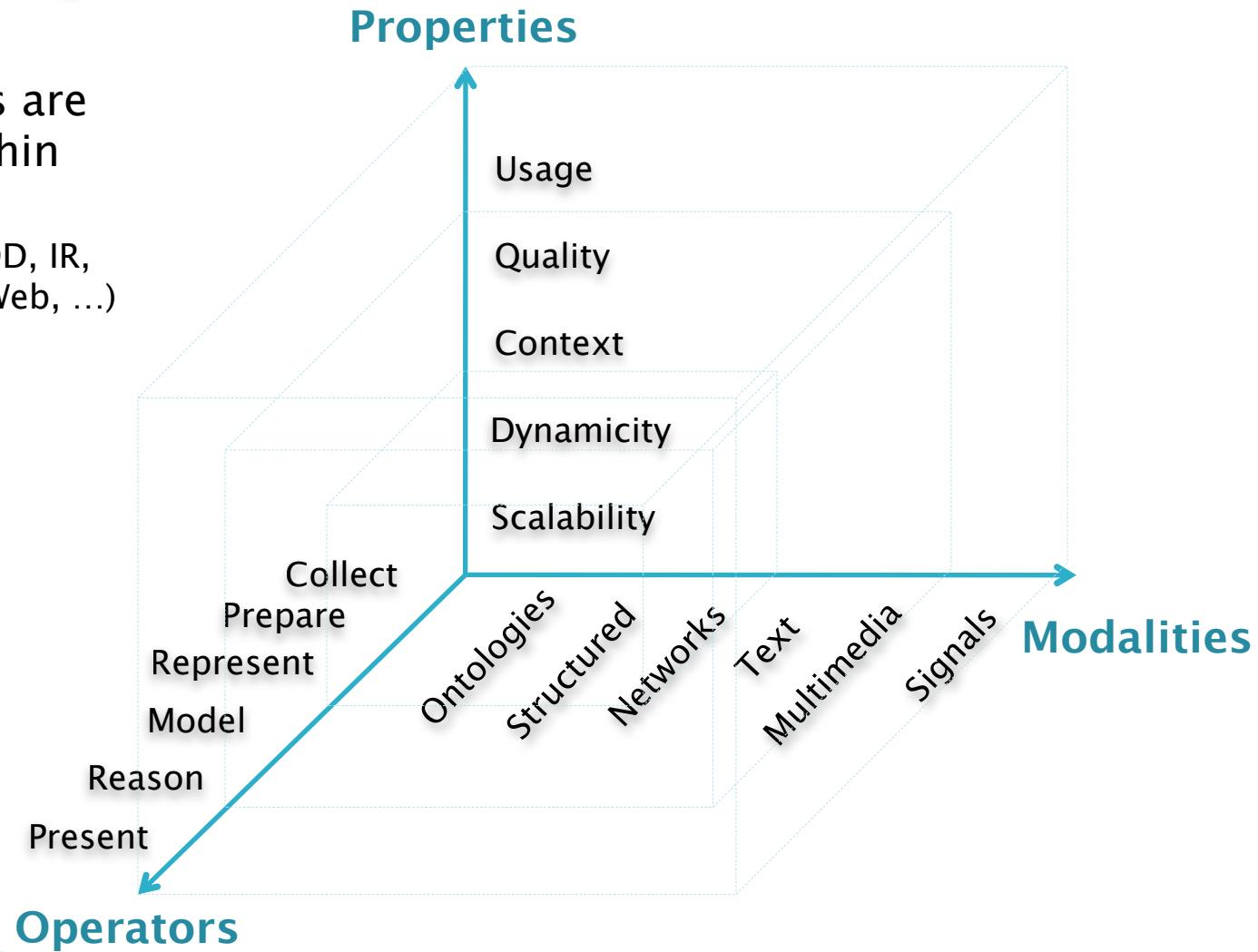


How we represent Text?

The Big picture

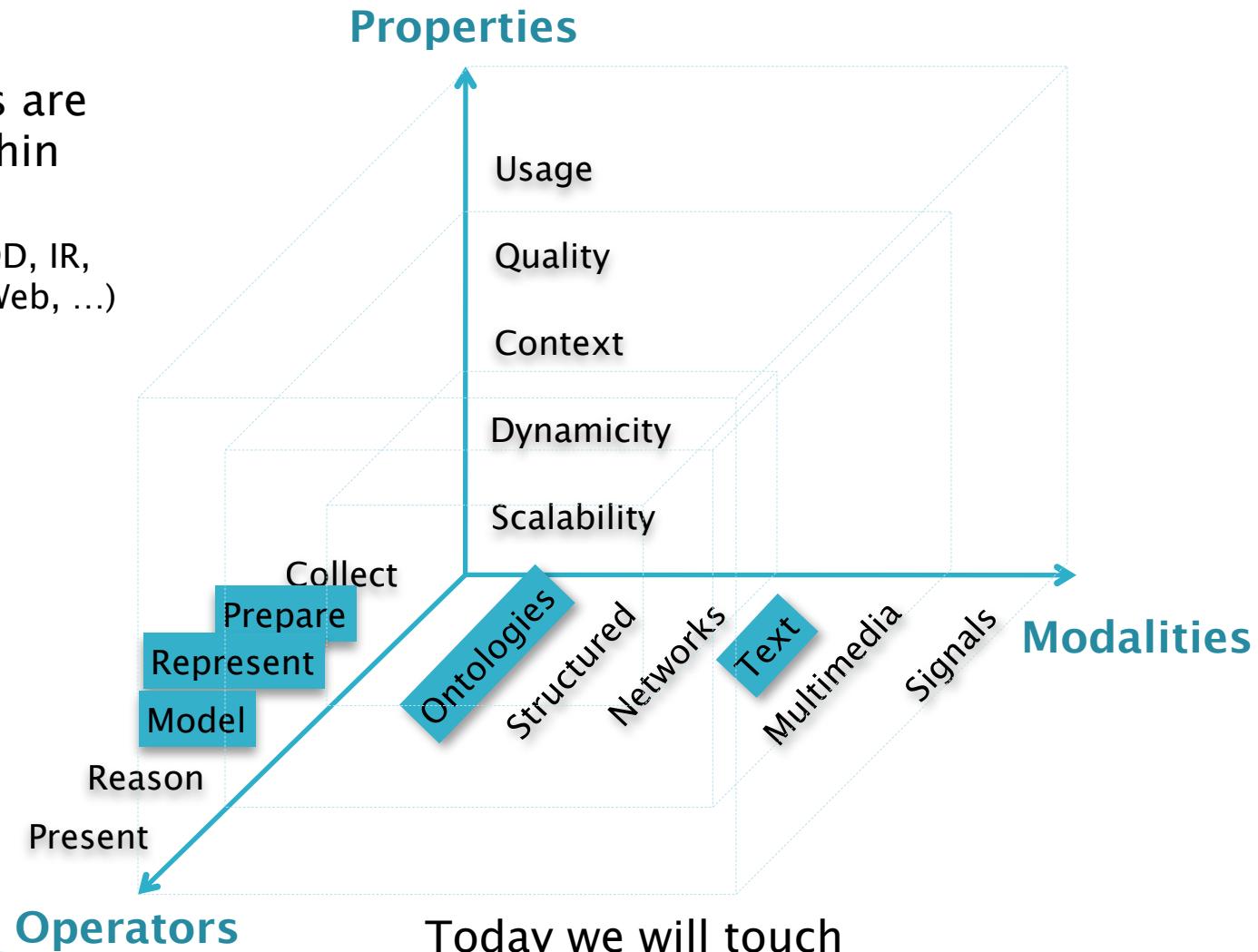
How we process data?

- ▶ Research areas are sub-cubes within the data cube
 - (such as ML, KDD, IR, SNA, NLP, SemWeb, ...)



What we do with data?

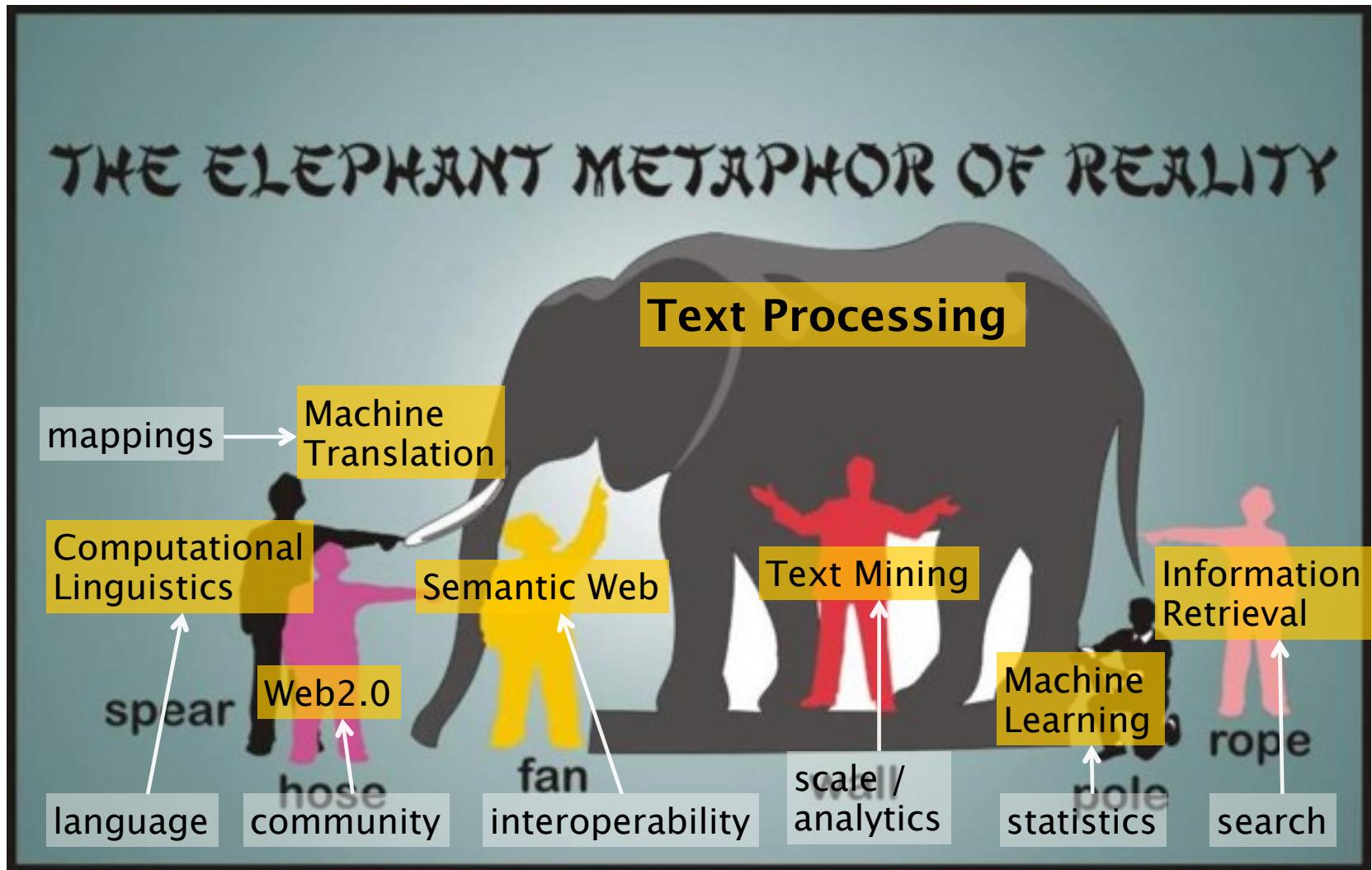
- ▶ Research areas are sub-cubes within the data cube
 - (such as ML, KDD, IR, SNA, NLP, SemWeb, ...)



Key paradigms when

- ▶ Three key scientific paradigms
 - **Top-down approaches (model driven)**
 - (Traditional NLP, KRR, Semantic Web)
 - **Bottom-up approaches (data driven)**
 - (Machine Learning, Data Mining)
 - **Collaborative approaches (socially driven)**
 - (Web2.0, Social Computing)

How different research areas approach text?



How do we represent text?

Levels of representation

Levels of text representations

- ▶ Character (character n-grams and sequences)
 - ▶ Words (stop-words, stemming, lemmatization)
 - ▶ Phrases (word n-grams, proximity features) **Named entity extraction** (names of people, places, organizations)
 - ▶ Part-of-speech tags
 - ▶ Taxonomies / thesauri
 - ▶ Vector-space model
 - ▶ Language models
 - ▶ Full-parsing
 - ▶ Cross-modality
 - ▶ Collaborative tagging / web2.0
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories
- Lexical**
- Syntactic**
- Semantic**
- Text categorization, Clustering, Search , Summarization, ...
 - Spam filtering, Machine translation
 - Data integration
 - Unifying semantics of data
 - Reasoning, Semantic Search

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
- ▶ Linked Data
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Lexical

Syntactic

Semantic

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
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Lexical

Syntactic

Semantic

Character level representation

- ▶ Character level representation of a text consists from contiguous sequences of characters...
 - ...a document is represented by a frequency distribution of sequences
 - ...each character sequence of length 1, 2, 3, ... represent a feature with its frequency

"the quick red"

trigrams

the	qui	k_r
he_	uic	_re
e_q	ick	red
qu	ck	

<http://en.wikipedia.org/wiki/Trigram>

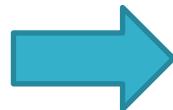
Good and bad sides of character level representation (n-grams)

- ▶ Representation has several strengths:
 - ...it is **robust** since avoids language morphology
 - (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
 - It is used as a basis for “string kernels” in combination with SVM for capturing complex character sequence patterns
- ▶ ...for deeper semantic tasks, the representation is too weak

Language identification

- Given a text document, the task is to identify a language of the document (from a predefined list of languages)
 - ...the key insight is that each language has characteristic “signature” of character n-grams
 - Demo: <http://www.lingua-systems.com/language-identifier/lid-library/identify-language.html>

The most common
letter bigrams in the
English language



th	1.52%	en	0.55%	ng	0.18%
he	1.28%	ed	0.53%	of	0.16%
in	0.94%	to	0.52%	al	0.09%
er	0.94%	it	0.50%	de	0.09%
an	0.82%	ou	0.50%	se	0.08%
re	0.68%	ea	0.47%	le	0.08%
nd	0.63%	hi	0.46%	sa	0.06%
at	0.59%	is	0.46%	si	0.05%
on	0.57%	or	0.43%	ar	0.04%
nt	0.56%	ti	0.34%	ve	0.04%
ha	0.56%	as	0.33%	ra	0.04%
es	0.56%	te	0.27%	ld	0.02%
st	0.55%	et	0.19%	ur	0.02%

<http://en.wikipedia.org/wiki/Bigram>

Character level normalization

▶ Hassle which we usually avoid:

- Since we have plenty of character encodings in use, it is sometimes nontrivial to identify a character and represent it in canonical form
- ...e.g. in Unicode the same word could be written in many ways – canonization:

Source	:	NFD	NFC
Å 00C5	:	A ő 0041 030A	Å 00C5
Ô 00F4	:	O ^ 006F 0302	Ô 00F4

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

Lexical

-
- ▶ Vector-space model
 - ▶ Language models
 - ▶ Full-parsing
 - ▶ Cross-modality
-
- ▶ Collaborative tagging / Web2.0
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Syntactic

Semantic

Word level

- ▶ Word tokens are 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

Key semantic word Properties

- ▶ Relations among word surface forms and their senses:
 - **Homonymy**: 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, ...

Stemming and lemmatization

- ▶ 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 transformation of a word into its stem:
 - **universe, university, universities, university's, universal -> univers**
- ▶ Lemmatization is transformation of a word into its normalized form
 - **universe -> universe,**
 - **university, universities, university's -> university,**
 - **universal -> universal**
- ▶ ...stemming provides an inexpensive mechanism to

Stemming

- ▶ 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 -> hesitation
 - 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 (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
- ▶ Linked Data
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Lexical

Syntactic

Semantic

Phrase level

- ▶ Instead of having just single words we can deal with sequences of words (phrases)
 - E.g. “artificial intelligence”, “text mining”, “word for windows”
- ▶ We use two types of phrases:
 - Frequent contiguous word sequences
 - Frequent non-contiguous word sequences
 - ...both types of phrases could be identified by simple dynamic programming algorithm
- ▶ The main effect of using phrases is to more precisely identify sense

Google n-gram corpus

- ▶ In 2006 Google announced availability of n-gram corpus:
 - <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html#links>
 - 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

Examples of Google n-grams

ceramics collectables collectibles 55
ceramics collectables fine 130
ceramics collected by 52
ceramics collectible pottery 50
ceramics collectibles cooking 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

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

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
- ▶ Linked Data
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Lexical

Syntactic

Semantic

Part-of-Speech

- ▶ 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”
- ▶ Part-of-Speech taggers are usually learned by HMM algorithm on manually tagged data

Part-of-Speech Tags

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

Part-of-Speech examples

verb
Stop!

noun	verb
John	works.

noun	verb	verb
John	is	working.

pronoun	verb	noun
She	loves	animals.

noun	verb	adjective	noun
Animals	like	kind	people.

noun	verb	noun	adverb
Tara	speaks	English	well.

noun	verb	adjective	noun
Tara	speaks	good	English.

pronoun	verb	preposition	adjective	noun	adverb
She	ran	to	the	station	quickly.

pron.	verb	adj.	noun	conjunction	pron.	verb	pron.
She	likes	big	snakes	but	I	hate	them.

Here is a sentence that contains every part of speech:

interjection	pron.	conj.	adj.	noun	verb	prep.	noun	adverb
Well,	she	and	young	John	walk	to	school	slowly.

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**
-

Lexical

- ▶ Vector–space model
 - ▶ Language models
 - ▶ Full–parsing
 - ▶ Cross–modality
-
- ▶ Collaborative tagging / Web2.0
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Syntactic

Semantic

Taxonomies/thesaurus level

- ▶ Thesaurus has as 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 also exists in several languages
 - WordNet: <http://wordnet.princeton.edu/>

WordNet – database of lexical

- ▶ 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

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

WordNet relations

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

Relation	Definition	Example
<i>Hypernym</i>	From lower to higher concepts	breakfast → meal
<i>Hyponym</i>	From concepts to subordinates	meal → lunch
<i>Has-Member</i>	From groups to their members	faculty → professor
<i>Member-Of</i>	From members to their groups	copilot → crew
<i>Has-Part</i>	From wholes to parts	table → leg
<i>Part-Of</i>	From parts to wholes	course → meal
<i>Antonym</i>	Opposites	leader → follower

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

Lexical

-
- ▶ Vector-space model

- ▶ Language models
- ▶ Full–parsing
- ▶ Cross–modality

Syntactic

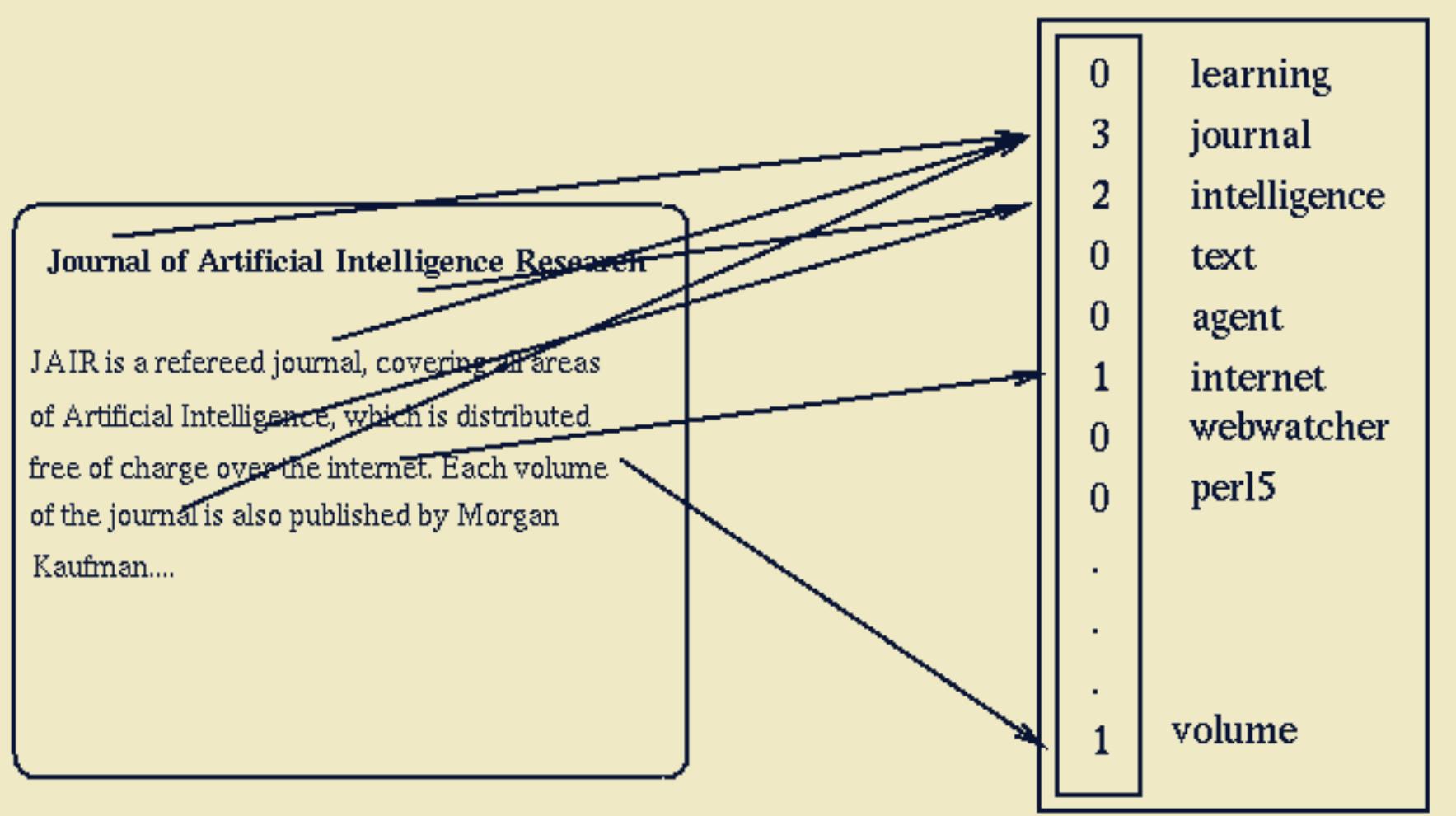
-
- ▶ Collaborative tagging / Web2.0
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Semantic

Vector-space model level

- ▶ The most common way to represent 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 language structure doesn’t harm efficiency of solving many relevant problems
 - This representation is referred to also as “**Bag-Of-Words**
 - Typical tasks on vector-space-model are classification, clustering, visualization etc.

Bag-of-Words document representation



Bag-of-Words Words

- ▶ 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

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]

Original text

Bag-of-Words
representation
(high dimensional
sparse vector)

Similarity between BoW

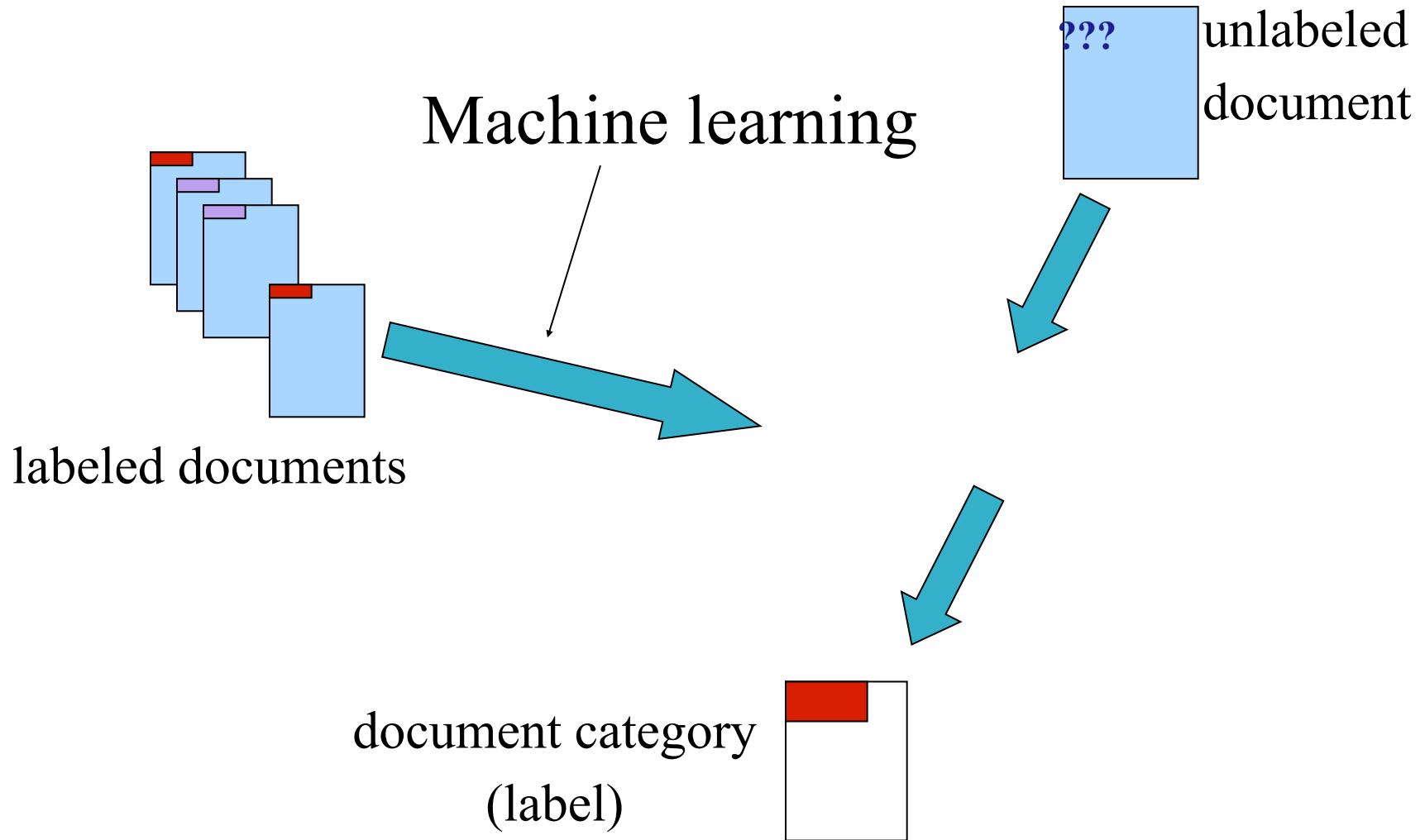
- ▶ The key of Bag-of-Words representation is to calculate topical similarity between documents fast
- ▶ Each document is represented as a vector of weights $D = \langle x \rangle$
- ▶ Cosine similarity (dot product) is the most widely used similarity measure between two document vectors
 - ...calculates 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)

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2} \sqrt{\sum_k x_k^2}}$$

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)

Document categorization



Algorithms for learning document classifiers

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

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** $\sum(d_i * w_i) \geq 0$ **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
 - $w_{t+1} = w_t + \text{sign}(\text{true-class}) * \text{Beta}$ (input parameter Beta>0)
 - // where sign(positive) = 1 and sign(negative) = -1

Measuring success – Model quality estimation

$$Precision(M, targetC) = P(targetC | \overline{targetC})$$

The truth, and

$$Recall(M, targetC) = P(\overline{targetC} | targetC)$$

..the whole truth

$$Accuracy(M) = \sum_i P(\overline{C}_i) \times Precision(M, C_i)$$

$$F_{\hat{\alpha}}(M, targetC) = \frac{(1 + \hat{\alpha}^2) Precision(M, targetC) \times Recall(M, targetC)}{\hat{\alpha}^2 Precision(M, targetC) + Recall(M, targetC)}$$

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

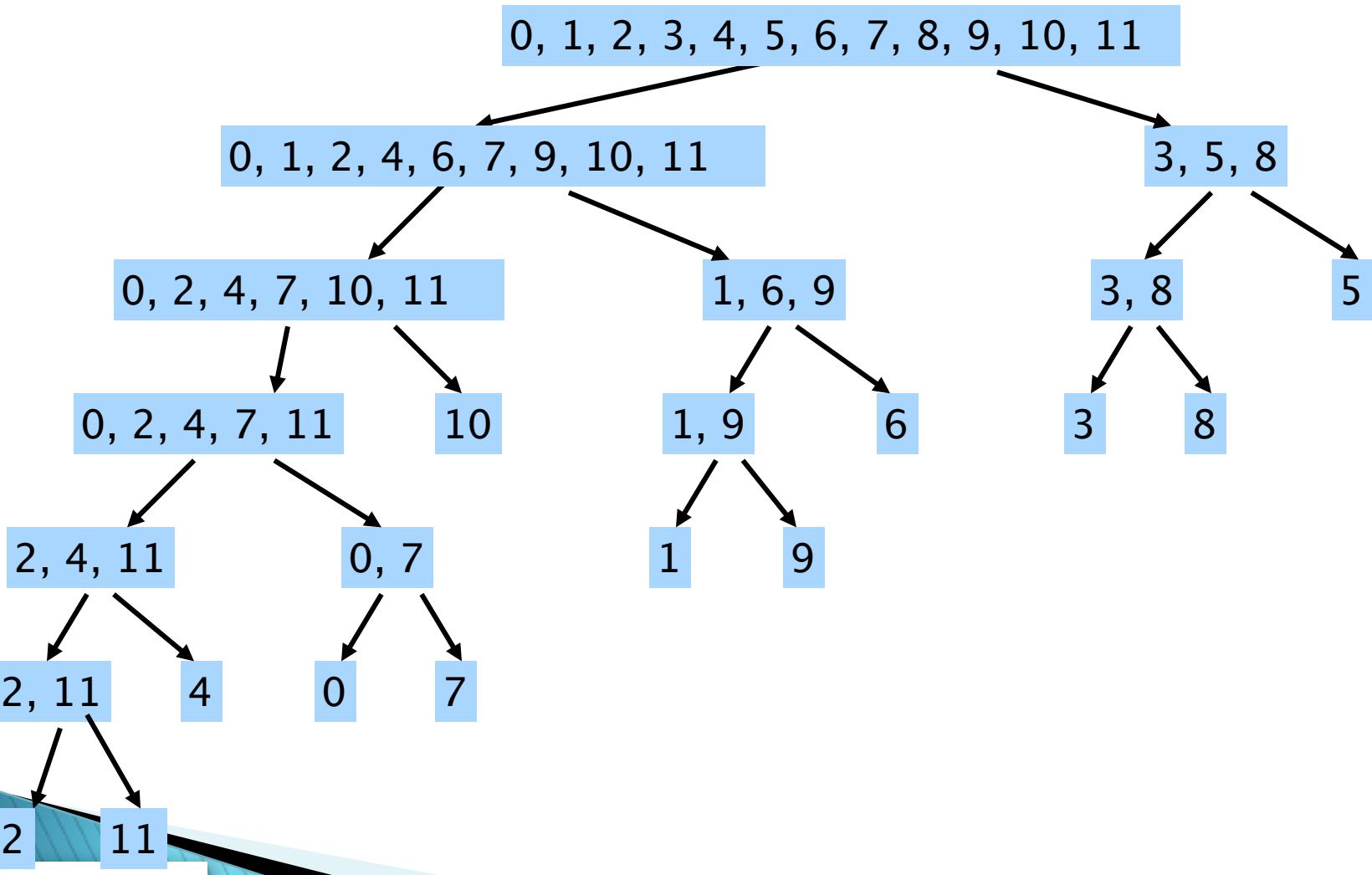
Document Clustering Task

- ▶ Clustering is a process of finding natural groups in the data in a 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 (bisecting 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

	d1	d2	d3	d4	d5	d6
cosmonaut	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

Original document-term matrix

Rescaled document matrix,
Reduced into two dimensions

	d1	d2	d3	d4	d5	d6
Dim1	-1.62	-0.60	-0.04	-0.97	-0.71	-0.26
Dim2	-0.46	-0.84	-0.30	1.00	0.35	0.65

High correlation although
d2 and d3 don't share
any word

Correlation matrix

	d1	d2	d3	d4	d5	d6
d1	1.00					
d2	0.8	1.00				
d3	0.4	0.9	1.00			
d4	0.5	-0.2	-0.6	1.00		
d5	0.7	0.2	-0.3	0.9	1.00	
d6	0.1	-0.5	-0.9	0.9	0.7	1.00

Document Classification into large taxonomies

DMoz (Open Directory Project) <http://dmoz.org>

- ▶ Largest handcrafted taxonomy on the Web
 - 4,892,414 sites
 - 91,778 editors
 - over 1,007,212 categories
- ▶ Data available for download
 - <http://www.dmoz.org/rdf.html>

The screenshot shows the DMOZ homepage within a Firefox browser. The title bar reads "Firefox" and "dmoz - Open Directory Project". The address bar shows the URL "http://www.dmoz.org/". The page itself has a green header with the "dmoz open directory project" logo and "In partnership with AOL Search.". Below the header is a search bar with a "Search" button and a link to "advanced" search options. The main content area is organized into three columns of category links:

- Arts**: Movies, Television, Music...
- Business**: Jobs, Real Estate, Investing...
- Computers**: Internet, Software, Hardware...
- Games**: Video Games, RPGs, Gambling...
- Health**: Fitness, Medicine, Alternative...
- Home**: Family, Consumers, Cooking...
- Kids and Teens**: Arts, School Time, Teen Life...
- News**: Media, Newspapers, Weather...
- Recreation**: Travel, Food, Outdoors, Humor...
- Reference**: Maps, Education, Libraries...
- Regional**: US, Canada, UK, Europe...
- Science**: Biology, Psychology, Physics...
- Shopping**: Clothing, Food, Gifts...
- Society**: People, Religion, Issues...
- Sports**: Baseball, Soccer, Basketball...
- World**: Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Svenska

At the bottom of the page, there's a "Become an Editor" button and the text "Help build the largest human-edited directory of the web". A copyright notice "Copyright © 2011 Netscape" and a footer statistic "4,892,414 sites - 91,778 editors - over 1,007,212 categories" are also present. A stylized green lizard logo is located in the bottom right corner.

Classification of a query into DMoz

The image shows two Mozilla Firefox windows side-by-side. The left window is titled "Classification into DMoz Top_Science" and contains a form for inputting a URL and text, setting categories (25), keyword threshold (0.75), and context. The right window is titled "Classification - Mozilla Firefox" and displays the results of the classification, listing keywords and their scores along with a list of categories and their scores. A large blue arrow points from the left window to the right window, indicating the flow of data from input to output.

Classification into DMoz Top_Science

URL:

Text:
format|

Categories: 25

Keyword Treshold (0=Non, 1=All): 0.75

Context:

XML Format:

Plain-Text Mime-Type:

Done

Classification - Mozilla Firefox

Results of Classification into DMoz Top_Science

Keywords:

- Science (0.183)
- Math (0.183)
- Number_Theory (0.182)
- Diophantine_Equations (0.154)
- Fermat's_Last_Theorem (0.115)

Categories:

1	<input type="checkbox"/>	0.504	Top/Science/Math/Number_Theory/Diophantine_Equations/Fermat's_Last_Theorem
2	<input type="checkbox"/>	0.341	Top/Science/Math/Number_Theory/Diophantine_Equations
3	<input type="checkbox"/>	0.117	Top/Science/Math/Number_Theory/Tables
4	<input type="checkbox"/>	0.113	Top/Science/Math/Number_Theory/History
5	<input type="checkbox"/>	0.074	Top/Science/Math/Number_Theory/Factoring
6	<input type="checkbox"/>	0.064	Top/Science/Math/Number_Theory/Factoring/Tables
7	<input type="checkbox"/>	0.053	Top/Science/Math/Number_Theory
8	<input type="checkbox"/>	0.053	Top/Science/Math/Number_Theory/Prime_Numbers/Primality_Tests/Pseudoprimes
9	<input type="checkbox"/>	0.037	Top/Science/Math/History/People
10	<input type="checkbox"/>	0.035	Top/Science/Math/Number_Theory/Prime_Numbers/Mersenne
11	<input type="checkbox"/>	0.027	Top/Science/Math/Number_Theory/Publications/Books

Done

Classification of a document into DMoz

The screenshot shows two Mozilla Firefox windows side-by-side. The left window is titled 'Classification into DMoz Top_Science' and contains a text input field with a large amount of scientific research text, several input fields for 'Categories', 'Keyword Threshold', 'Context', and 'XML Format', and two buttons for 'Submit' and 'Reset'. A blue arrow points from the right side of this window towards the second window. The right window is titled 'Classification - Mozilla Firefox' and displays the results of the classification. It has a title bar 'Results of Classification into DMoz Top_Science'. Below it are two tables: 'Keywords' and 'Categories'. The 'Keywords' table lists various scientific terms with their scores: Science (0.043), Institutions (0.069), Research_Centres (0.048), Social_Sciences (0.043), Institutes (0.041), Biology (0.039), Oceanography (0.018), Earth_Sciences (0.016), Research_Institutes (0.016), Regional (0.016), Research (0.014), Europe (0.014), Organisations (0.011), Math (0.011), Ecology (0.008), Environment (0.012), Associations (0.010), Economics (0.010), Agriculture (0.010), Oracle (0.010), and Plasma (0.010). The 'Categories' table lists ten categories with their scores: Top_Science_Institutions_Research_Institutes (0.181), Top_Science_Biology_Institutions_Research_Centres (0.144), Top_Science_Institutions_Regional (0.113), Top_Science_Environment_Organisations_Research_Institutes (0.109), Top_Science_Math_Research_Institutes (0.291), Top_Science_Institutions_Associations (0.289), Top_Science_Math_Research (0.281), Top_Science_Social_Sciences_Economics_Institutes (0.278), Top_Science_Agriculture_Research_Centres (0.276), and Top_Science_Institutions_Regional_Europe (0.270).

Keywords	Score
Science	(0.043)
Institutions	(0.069)
Research_Centres	(0.048)
Social_Sciences	(0.043)
Institutes	(0.041)
Biology	(0.039)
Oceanography	(0.018)
Earth_Sciences	(0.016)
Research_Institutes	(0.016)
Regional	(0.016)
Research	(0.014)
Europe	(0.014)
Organisations	(0.011)
Math	(0.011)
Ecology	(0.008)
Environment	(0.012)
Associations	(0.010)
Economics	(0.010)
Agriculture	(0.010)
Oracle	(0.010)
Plasma	(0.010)

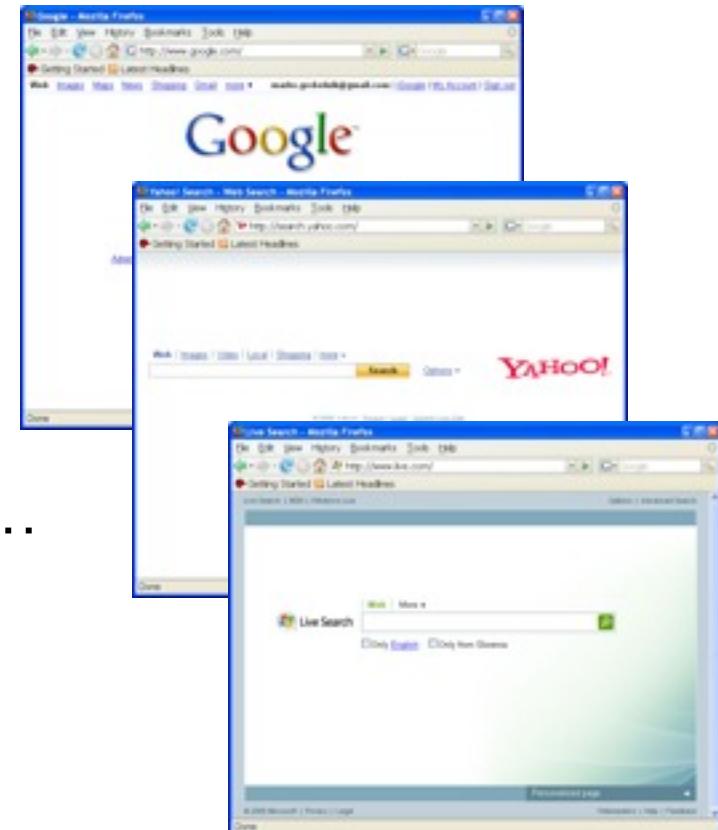
Categories	Score
Top_Science_Institutions_Research_Institutes	0.181
Top_Science_Biology_Institutions_Research_Centres	0.144
Top_Science_Institutions_Regional	0.113
Top_Science_Environment_Organisations_Research_Institutes	0.109
Top_Science_Math_Research_Institutes	0.291
Top_Science_Institutions_Associations	0.289
Top_Science_Math_Research	0.281
Top_Science_Social_Sciences_Economics_Institutes	0.278
Top_Science_Agriculture_Research_Centres	0.276
Top_Science_Institutions_Regional_Europe	0.270

Visual & Contextual Search

(WWW2008)

Contextualized search

- ▶ What is the most common tasks where we manipulate text in everyday life?
 - “Internet search”!
- ▶ ...but – how smart is search technology today?
 - ...not too smart!
 - It is sophisticated, but not smart...



Example: searching for

- ▶ Query “jaguar” has many meanings...
- ▶ ...but the first page of search engines doesn’t provide us with many answers
- ▶ ...there are 84M more results

A screenshot of a Mozilla Firefox browser window showing a Google search results page for the query "jaguar". The search bar contains "jaguar". The results page shows 10 out of approximately 84,200,000 results. The first result is for "Jaguar" (Cars), followed by "Jaguar UK - Jaguar Cars" (listing links for XF, TEST DRIVE, Brochure, Dealer, eNewsletter, SITEMAP, COMPANY, Privacy Policy, Accessibility Statement, Contact Us, TERMS & CONDITIONS, and a Cached version), "Jaguar US - Home" (listing links for USA official website, Build Your Jaguar, Request Brochure, Get Email Updates, Locate a Dealer, Search Your Profile Site Map, Contact Us, Privacy, and a Cached version), "Jaguar - Wikipedia, the free encyclopedia" (listing links for Panthera onca, pronunciation, New World mammal, Felidae family, and one ..., and a Cached version), and "Jaguar Cars" (listing links for English, Français, www.jaguar.ca/, and a Cached version). A blue circle highlights the number of results (84,200,000) at the top of the page, and a black arrow points from the bottom-left towards this circled number.

Context sensitive search with <http://searchpoint.ijs.si>

Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point

The screenshot shows a Windows Internet Explorer window titled "Jaguar - SearchPoint - Windows Internet Explorer". The address bar contains "Http://searchpoint.ijs.si/Result.aspx". The page displays a search results list for the query "jaguar". Each result includes a rank, a title, a brief description, and a URL. The results are color-coded by category. A conceptual map on the right side of the page shows various topics like Mammalia, Vehicles, Sports, and NFL, with arrows indicating relationships between them. A red dot on the map points to the word "Jaguar" in the search bar. The search results are annotated with arrows pointing from the labels on the left to specific parts of the page: "Query" points to the search bar, "Conceptual map" points to the map, "Search Point" points to the red dot on the map, and "Dynamic contextual ranking based on the search point" points to the ranked search results.

(9) [Jaguar](#)
General information and facts from Big Cats Online.
<http://www.abf90.dial.pipex.com/jaguar.htm>

(59) [Jaguar, Jaguar Profile, Facts, Information, Photos, Pictures](#)
Get jaguar profile, facts, information, photos, pictures, sounds, habitats, reports, news, and more from National Geographic.
<http://animals.nationalgeographic.com/animals/mammals/Jaguar.html>

(9) [Jaguar - Wikipedia, the free encyclopedia](#)
The jaguar (*Panthera onca*) is a New World mammal of the Felidae family, and one of four "big cats" in the *Panthera* genus, along with the tiger, ...
<http://en.wikipedia.org/wiki/Jaguar>

(11) [Jaguar](#)
Jaguar Facts, Jaguar Photos and Jaguars in the news at the world's largest big cat rescue and sanctuary.
<http://www.bigcatrescue.org/jaguar.htm>

(1) [Jaguar](#)
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets.
<http://www.jaguar.com/>

(32) [Jaguar](#)
Contains extensive information about the Jaguar. Information includes habitat, body size, and life span.
<http://www.abf90.dial.pipex.com/bco/jaguar.htm>

(2) [Jaguar UK - Jaguar Cars](#)
Jaguar & Ownership. Highlights. Gallery. Models & Pricing. Design Your XK. TEST DRIVE. Brochure. Dealer. eNewsletter ...
<http://www.jaguar.co.uk/>

(17) [Jaguar Enthusiasts' Club](#)
World's largest audited membership. UK-based. JEC's site has extensive resources available for the enthusiast, including information about their Sections, ...
<http://www.jec.org.uk/>

(20) [San Diego Zoo's Animal Bytes: Jaguar](#)
Get fun and interesting jaguar facts in an easy-to-read style from the San Diego Zoo's Animal

"Jožef Stefan" Institute

News reporting bias

(Fortuna, Galleguillos, Cristianini 2008)

News Reporting Bias example

UK SOLDIERS CLEARED IN IRAQI DEATH – SEVEN BRITISH SOLDIERS WERE ACQUITTED ON THURSDAY OF CHARGES OF BEATING AN INNOCENT IRAQI TEENAGER TO DEATH WITH RIFLE BUTTS. A JUDGE AT A SPECIALLY CONVENED MILITARY COURT IN EASTERN ENGLAND ORDERED THE ADJUDICATING PANEL TO RETURN 'NOT GUILTY' VERDICTS AGAINST THE SEVEN BECAUSE HE DID NOT BELIEVE THERE WAS SUFFICIENT EVIDENCE AGAINST THEM, THE MINISTRY OF DEFENCE SAID. . . .

BRITISH MURDERERS IN IRAQ ACQUITTED – THE JUDGE AT A COURT-MARTIAL ON THURSDAY DISMISSED MURDER CHARGES AGAINST SEVEN SOLDIERS, FROM THE 3RD BATTALION, THE PARACHUTE REGIMENT, WHO'RE ACCUSED OF MURDERING IRAQI TEENAGER; CLAIMING THERE'S INSUFFICIENT EVIDENCE TO SECURE A CONVICTION, THE ASSOCIATED PRESS REPORTED THURSDAY. . . .

Experimental setup

- ▶ Time period: March 31st 2005 – April 14th 2006
- ▶ Size of collections:

Source	No. of news
Al Jazeera	2142
CNN	6840
Detroit News	2929
International Herald Tribune	9641

- ▶ Number of discovered matches:

	AJ	CNN	DN	IHT
AJ	–	816	447	834
CNN	816	–	1103	2437
DN	447	1103	–	895
IHT	834	2437	895	–

Prediction of news source

- ▶ **The task:** given a pair of news articles describing the same event, can we predict the news source for each?
- ▶ In this experiment we focused on CNN and Al Jazeera.
- ▶ SVM linear classifier was used for prediction
 - Evaluation was done using 10-fold cross-validation
 - Significance of results was tested against random matches

Detecting News Reporting

- ▶ The task:
 - Given a news story, are we able to say from which news source it came?
- ▶ We compared **CNN** and **Aljazeera** reports about the same events from the war in Iraq
 - ...300 aligned articles describing the same story from both sources
- ▶ The same topics are expressed in both sources with the following keywords:
 - CNN with:
 - **Insurgents**, Troops, Baghdad, Iran, **Militant**, Police, **Suicide**, **Terrorist**, United, National, Hussein, **Alleged**, Israeli, Syria, Terrorism...
 - Aljazeera with:
 - Attacks, Claims, **Rebels**, Withdrawing, Report, **Fighters**, President, **Resistance**, Occupation, Injured, Army, Demanded, Hit, Muslim, ...

News Visualization

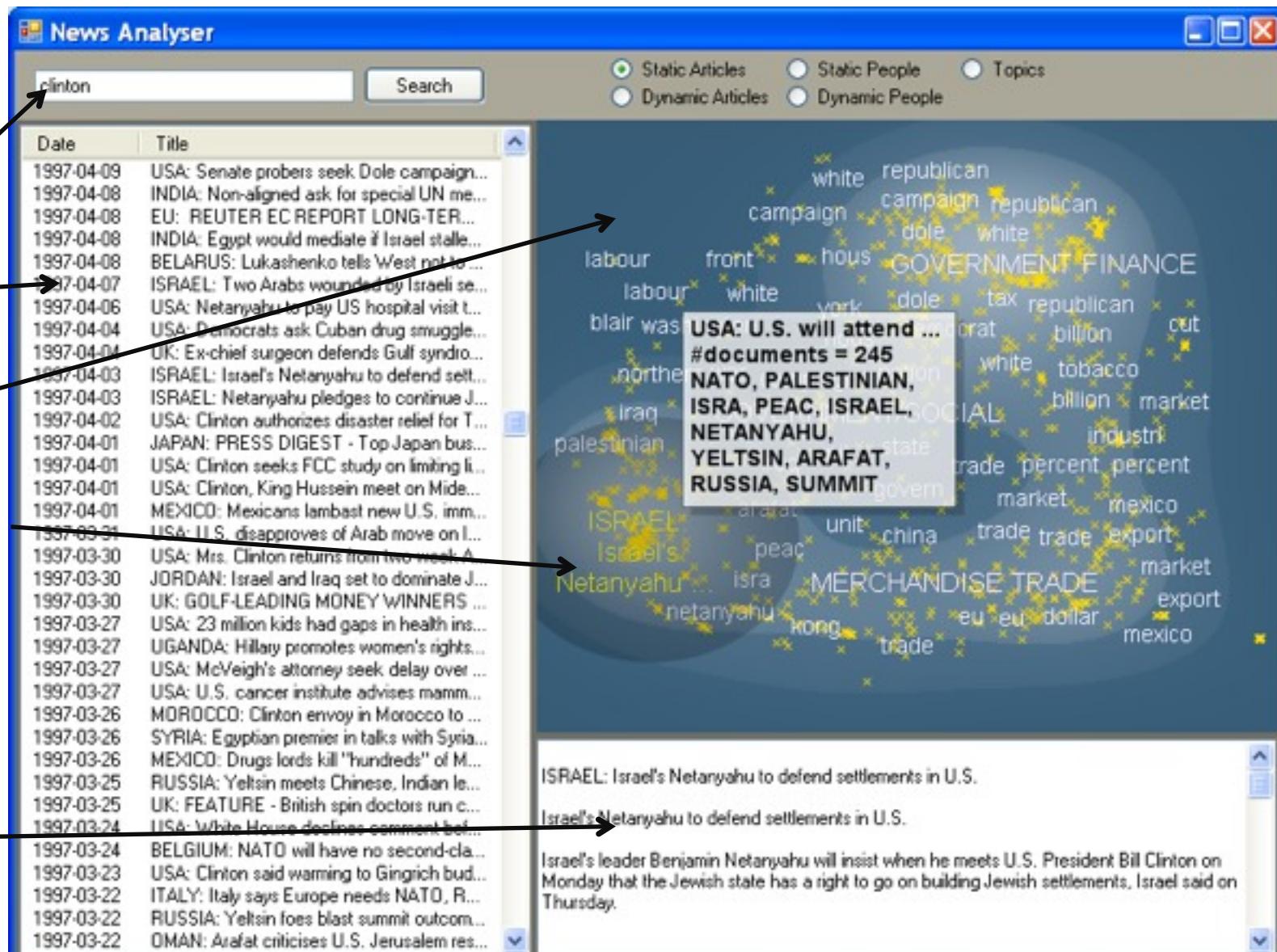
Topic landscape of the query “Clinton” from Reuters news 1996–1997

Query
Search Results

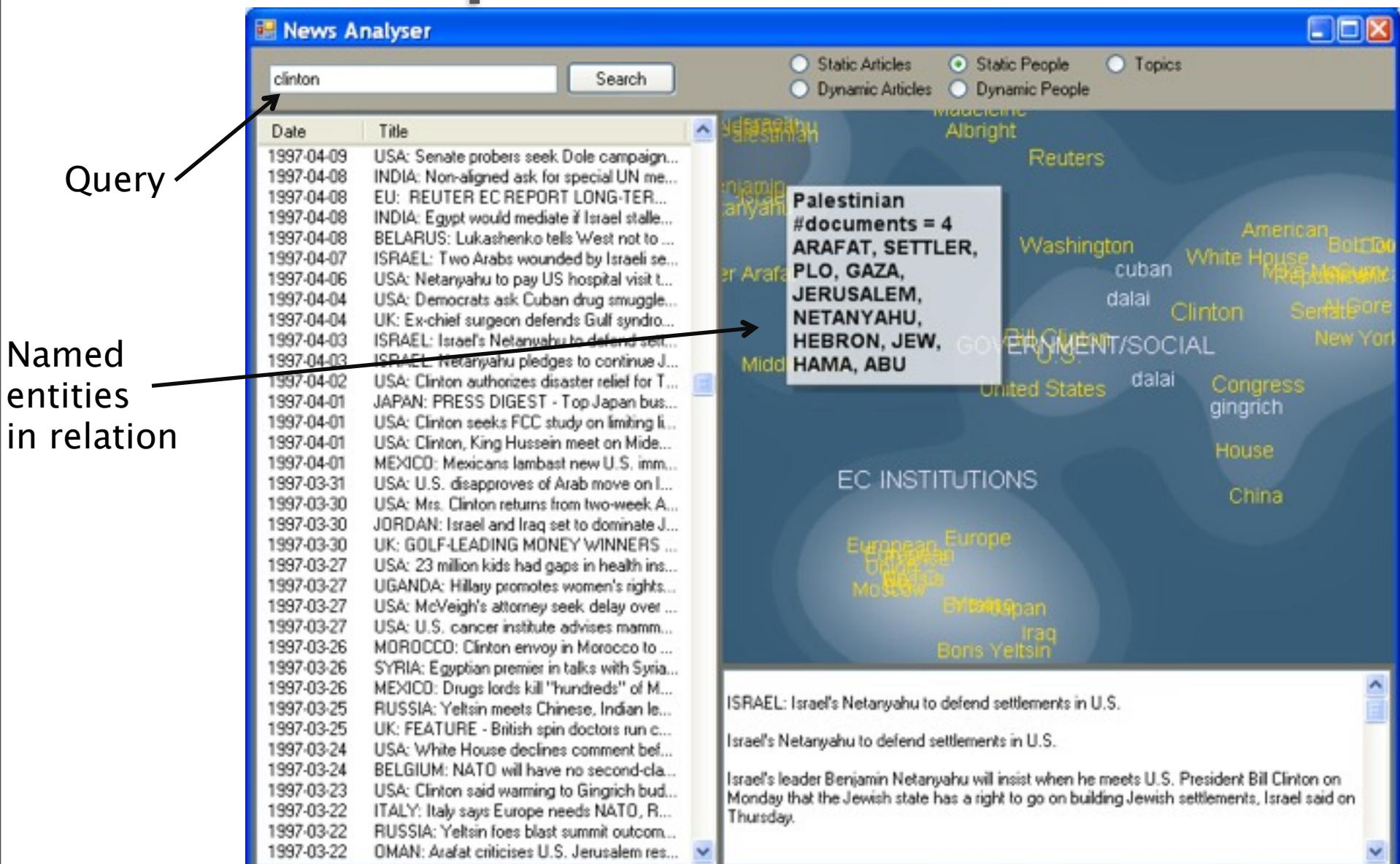
Topic Map

Selected group of news

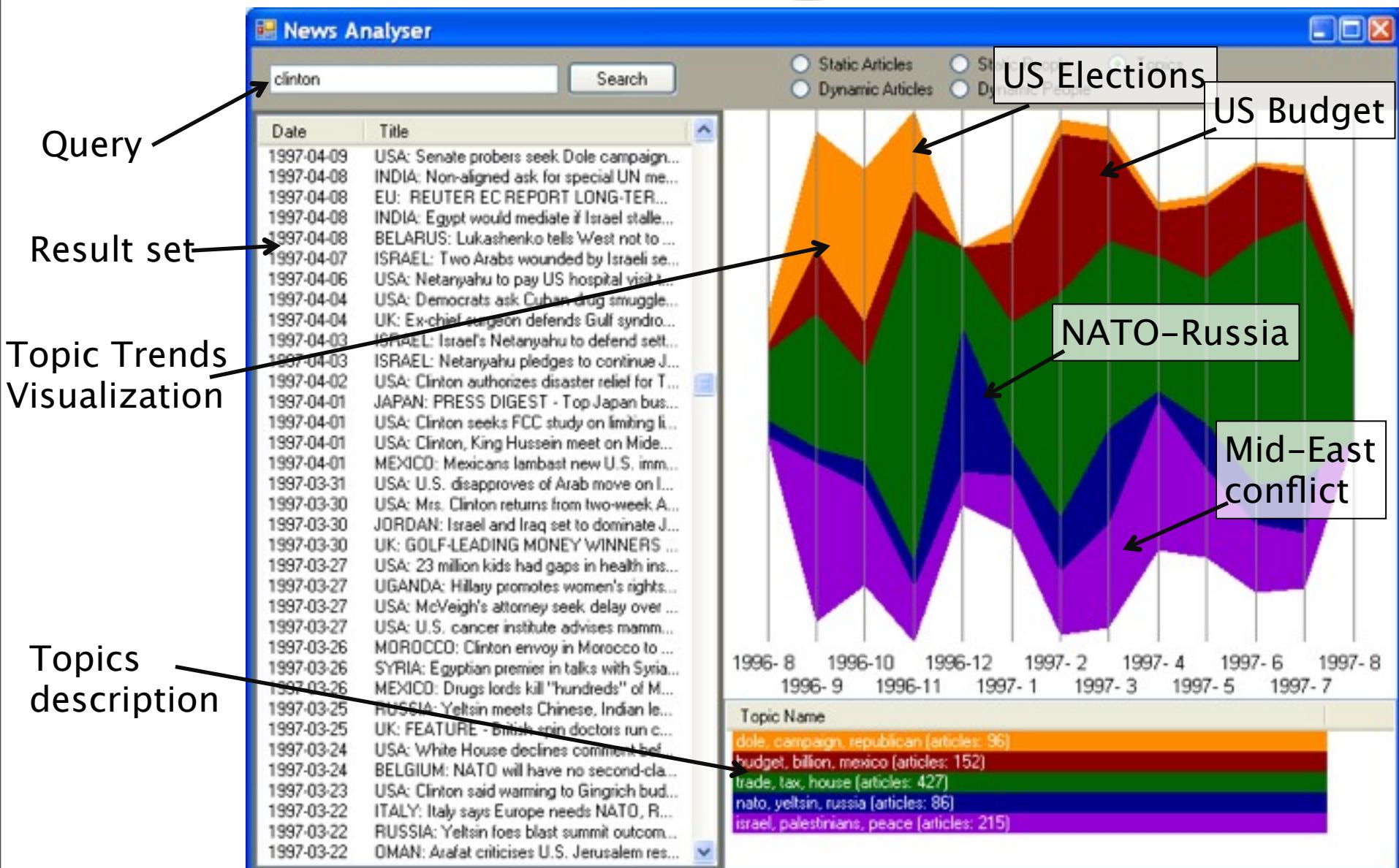
Selected story



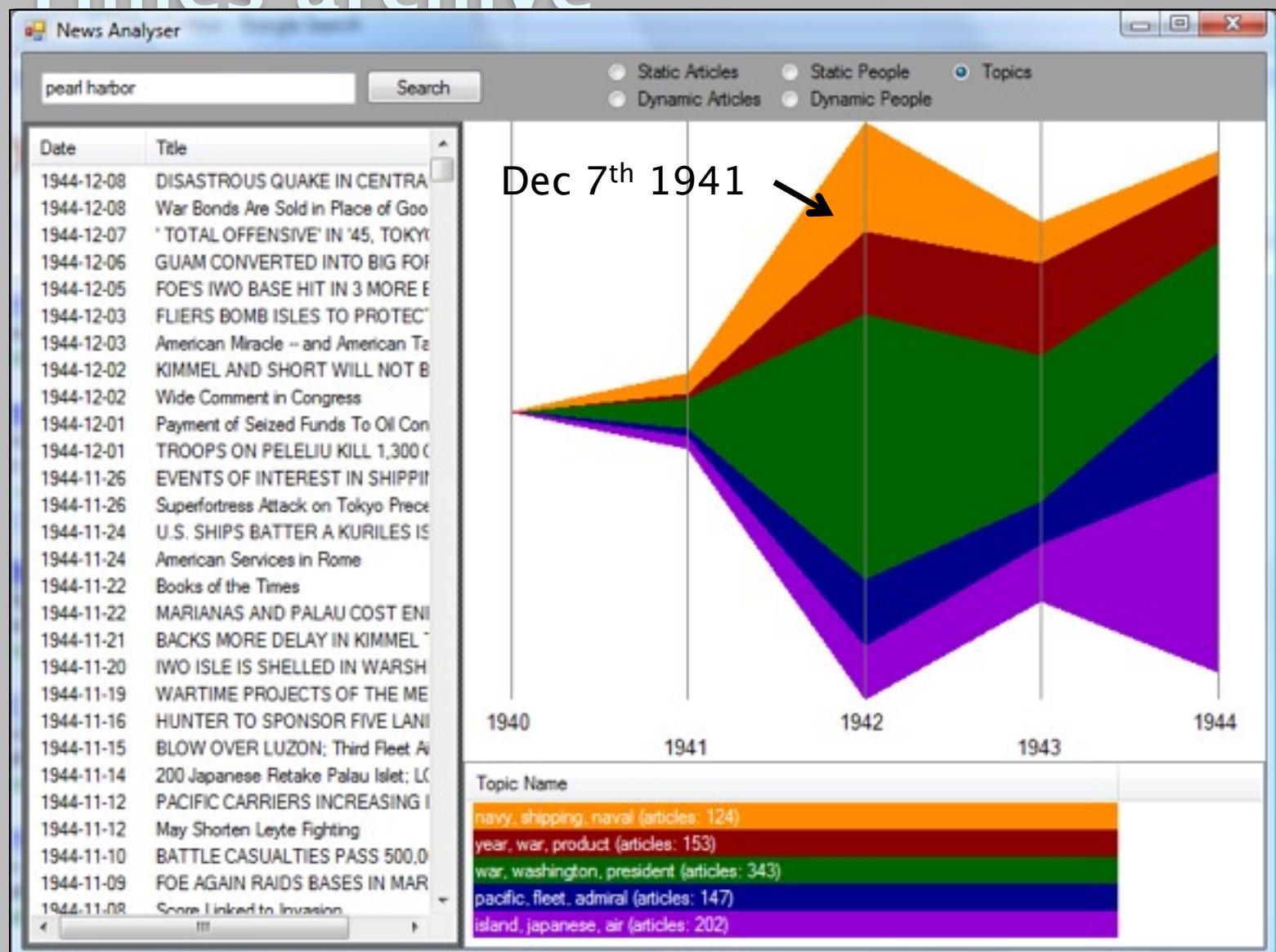
Visualization of social relationships between “Clinton”



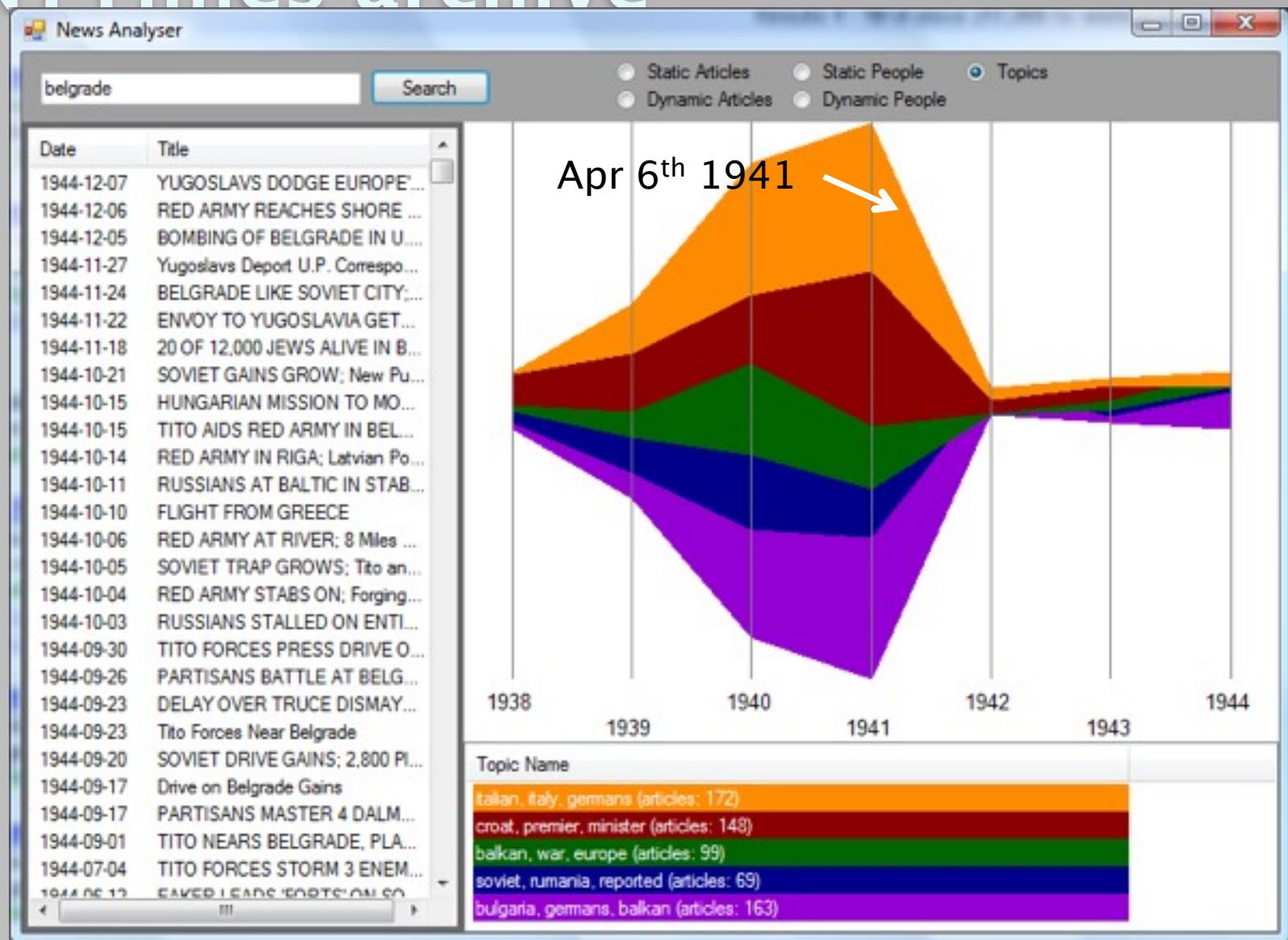
Topic Trends Tracking of the documents including “Clinton”



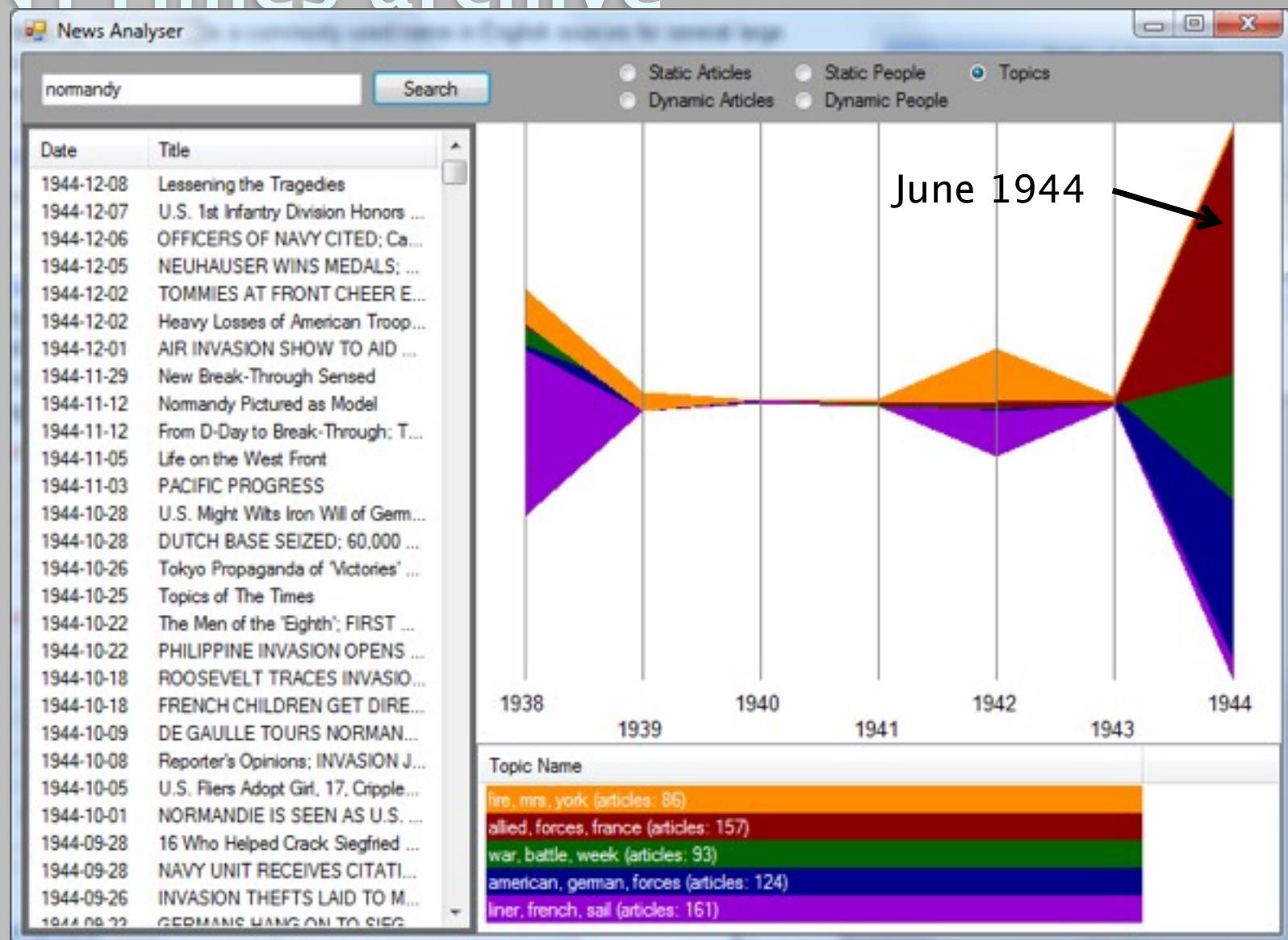
WW2 query “Pearl Harbor” into NYTimes archive



WW2 query “Belgrade” into NYTimes archive



WW2 query “Normandy” into NYTimes archive



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

Lexical

- ▶ Vector–space model
 - ▶ Language models
 - ▶ Full–parsing
 - ▶ Cross–modality
-
- ▶ Collaborative tagging / Web2.0
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

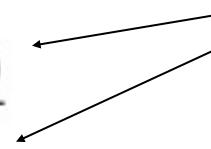
Syntactic

Semantic

Language model level

- ▶ Language modeling is about determining probability of a sequence of words

- The task typically gets reduced to estimating probabilities of a next word given two previous words (trigram model):

$$P(w_i|w_{i-2}w_{i-1}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}$$


Frequencies
of word
sequences

- It has many applications including speech recognition, OCR, handwriting recognition, machine translation and spelling correction

Context aware auto-complete

(Tadej Stajner 2011)

Context-aware prediction for document authoring

- **Situation:**
 - An enterprise has many professional profiles
 - Each professional profile develops its own sub-language
- **Goal:**
 - Assist document authoring by finding most likely completions of a text fragment
 - Ranking consists of sequence probability and sequence length

Context-aware prediction for document authoring

- Sentence:

'THE SOLUTION THAT IS PROVIDED ...'

- Language model of Technical Design and Implementation Material
 - 0.2505: THE SOLUTION THAT IS PROVIDED **AS PART OF THE COMPILATION UNIT**
 - 0.1352: THE SOLUTION THAT IS PROVIDED **FOR BUILDING AND HOSTING WEB**
- Language model of Proposal material
 - 0.4727: THE SOLUTION THAT IS PROVIDED **TO IDENTIFY RULES**
 - 0.2957: THE SOLUTION THAT IS PROVIDED **TO HELP BUILDING**
 - 0.0691: THE SOLUTION THAT IS PROVIDED **FOR THE CROSS - FUNCTIONAL**

Context-aware prediction for document authoring

- Sentence:
'THE MARKETING STRATEGY FOR THE SOLUTION ...'
- Language model of Demo/Prototype
 - **0.0214: THE MARKETING STRATEGY FOR THE SOLUTION AND WATCH THE TEAM PRESENT A DEMO OF THE DATA QUALITY COCKPIT**
- Language model of Market Intelligence
 - **0.7883: THE MARKETING STRATEGY FOR THE SOLUTION AND ASSESSES THE POTENTIAL OPPORTUNITIES**
 - **0.2262: THE MARKETING STRATEGY FOR THE SOLUTION TO MEET THE CLIENT CHALLENGES**

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
- ▶ Linked Data
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

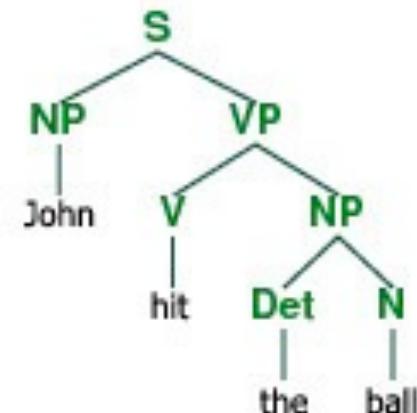
Lexical

Syntactic

Semantic

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



Text Enrichment

WWW2009-SemSearch

Text enrichment with [Enrycher.ijs.si](http://enrycher.ijs.si)

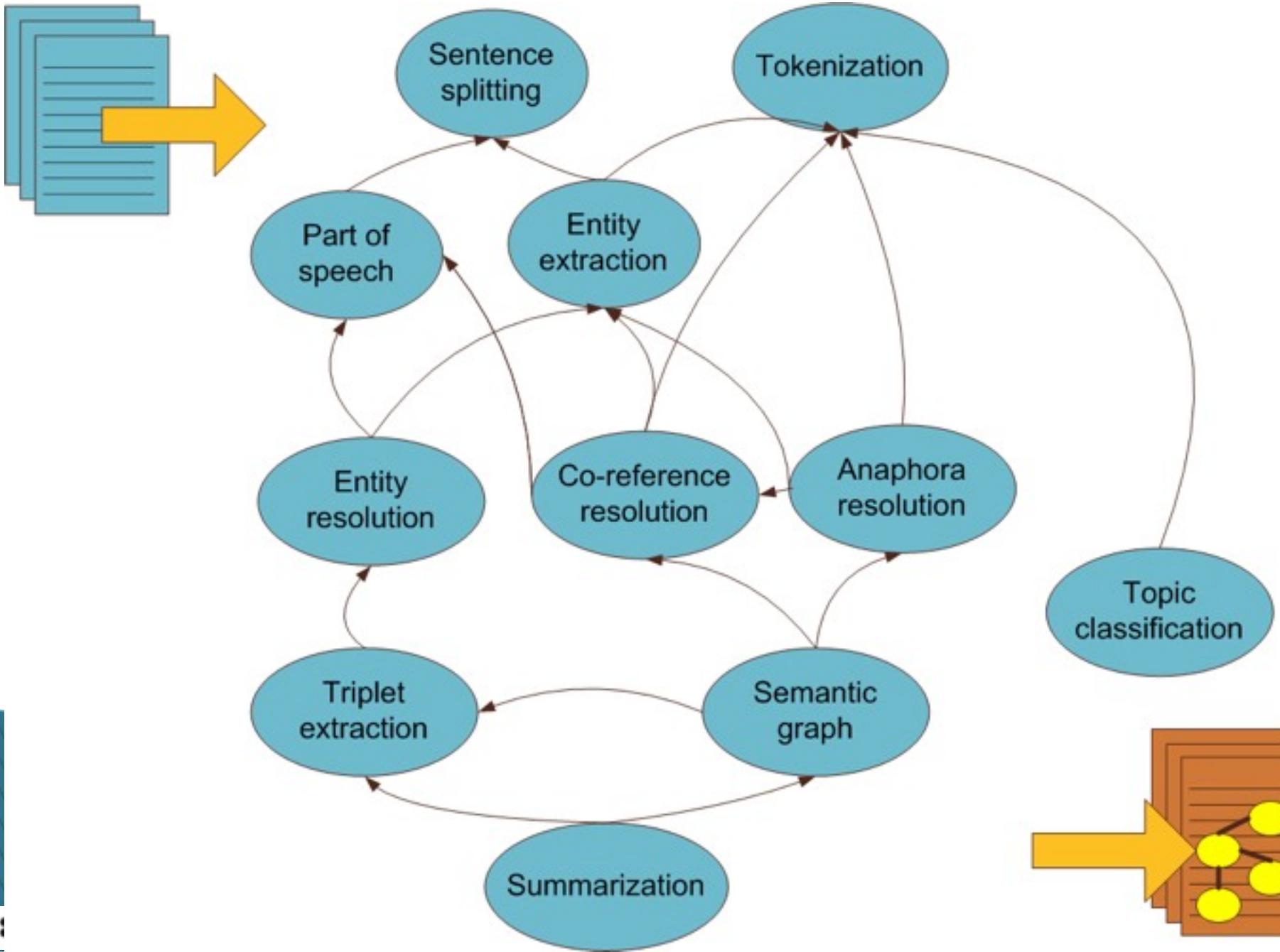
The screenshot shows the Enrycher web application. At the top, there's a navigation bar with links for home, about, api, and contact. Below that is a search bar labeled "try out enrycher!" with a placeholder "example:" and a list of suggestions. The main content area displays a text snippet about a soccer match between Slovenia and Ireland, followed by an "Interesting statements" section with a tree diagram of enriched words like "Michael Essien", "Ireland", "soccer", etc., and their relationships. At the bottom, there are buttons for "HTML response", "XML response", and "enrycher", along with a "Done" button.



This screenshot shows the "enrycher result" page. It features a large word cloud visualization where words are represented as bubbles of varying sizes and colors (green, blue, yellow) connected by lines, indicating their semantic relationships. Below the visualization is a section titled "Interesting statements" containing a list of enriched text fragments. At the bottom, there are buttons for "TODAY'S CONTENT (444 items)", "RSS FEED (240 items)", "SEARCH", and "Done".



This screenshot shows a Microsoft Internet Explorer window displaying the raw XML response from the Enrycher API. The XML structure includes various nodes such as "text", "enrichedText", "entity", "mention", and "relation". The content of the XML corresponds to the enriched text and word cloud shown in the previous screenshots.



Knowledge based summarization

AAAI 2005

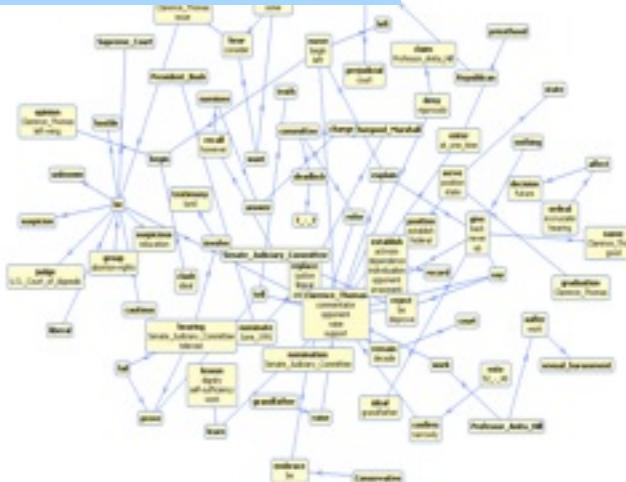
Original document

Cracks appear in U.N. Trade Embargo Against Iraq.

Hussein sought to recruit the economic power around his country, Japan, forcing the sanctions. Hoping to defuse criticism that it is not doing its share to be affected by the U.N. embargo on Iraq, President Bush on Tuesday night promised a "Saddam Hussein will fall" to make his conquest of Kuwait permanent. "America's military may remain there; Saudi Arabia's desert indefinitely." I cannot predict just where that will lead," he said. "I can only say that we will do what we have to do to follow his address to Congress with a televised message for the Iraqi people, whom had offered Bush time on Iraqi TV. The Philippines and Namibia, the first of the countries to break ranks, are sending their own tankers to get it," said no to the Iraqi leader. Saddam's tanks were never far from the water, so many believe that they are now being used to prevent ships from getting to Iraq. But according to a State Department survey, Cuba and Romania have struck oil wells in Iraq and companies elsewhere are trying to continue trade with Baghdad in defiance of U.N. sanctions. Romania denies the allegation. The report, made available to The Associated Press, said some Eastern European countries also are trying to maintain their military sales to Iraq. A well-informed source in Tehran told The Associated Press that Iran has agreed to an Iranian oil deal to buy up to 100,000 barrels of refined oil per month. That would come after the United States and Britain or Bahrain on the reported food-for-oil deal. But the source, who requested anonymity, said the deal will not affect Iraqi Foreign Minister Tariq Aziz's visit Sunday to Tehran, the first by a senior Iraqi official since the 1980-88 gulf war. After the visit, the two countries announced they would resume diplomatic relations. Well-informed oil industry sources in the region, contacted by AP, said that although Iraq is a major oil exporter itself, it currently has to import about 150,000 barrels of refined oil a day for domestic use because of damages suffered in the gulf war. Also in similar lines, ABC News reported that following Aziz's visit, Iran is prepared to buy more than all the refined oil available for the dollar. In addition, the United States' James A. Baker III, meanwhile, met in Moscow with Soviet Foreign Minister Eduard Shevardnadze, two days after the U.S.-Soviet summit that produced a joint demand that Iraq withdraw from Kuwait. During the summit, Bush encouraged Mikhail Gorbachev to withdraw 190 Soviet military specialists from Iraq, where they remain to fulfill contracts. Shevardnadze told the Soviet press yesterday the specialists had been withdrawn. "A war is a war, justified or not, 5,800 Soviet citizens in Iraq," his speech, "but we must not go out to fight the war of the invaders of our country." "Our position is not to change, and it will not change. America and the world will not be blackmailed," the president added. "Vital issues of principle are at stake. Saddam Hussein is literally trying to wipe a country off the face of the Earth." In other developments, a U.S. diplomat in Baghdad said Tuesday up to 800 Americans and Britons will fly out of Iraq equipped with their families, mostly women and children, leaving them behind. Saddam has said he is keeping foreign men as human shields against attacks. On Monday, a plane of 164 Westerners was based in Basra from Iraq. Estimates of the number of foreigners in Iraq range from 100,000 to 150,000. A soldier can rape a father's daughter in front of him and he can't do anything about it," the State Department said Iraq had told U.S. officials that American male citizens in Iraq will be allowed to leave the country if they want to. "The American public needs to understand how many men the Iraqi move could affect." A Pentagon spokesman said American forces will be sent to the region to help detect and identify individuals near its borders with Turkey and Syria. He said there will be little indication hostilities are imminent. Defense Secretary Dick Cheney said the cost of the U.S. military buildup in the Middle East was rising above the \$1 billion-a-month estimate generally used by government officials. He said the total cost, if no shooting war breaks out, could total \$1.5 billion in the next fiscal year beginning Oct. 1, "a significant increase" in help from Arab nations and other U.S. allies. In addition, the U.S. has been accused of ordering its forces to the Persian Gulf in response to a \$1 billion deployment, Jordan, hit hardest by the U.N. prohibition on trade with Iraq. "The pressure from abroad is getting so strong," said Hirooysu Horio, an official with the Ministry of International Trade and Industry. Local news reports said the aid would be extended through the World Bank and International Monetary Fund, and \$600 million would be set aside as part of the U.S.-Soviet agreement. On Tuesday, Treasury Secretary Nicholas Brady visited Tokyo to seek a waiver seeking \$10.5 billion to help Egypt, Jordan and Turkey. The U.S. already has promised \$3 billion to help those countries meet their losses after vehicles and prefabricated housing for non-military uses. But critics in the United States have said Japan moves because its economy depends heavily on oil from the Middle East. Japan imports 99 percent of its oil. Japan's constitution bars the use of force in settling international disputes and Japanese law restricts the right to Japanese territory, except in self-defense. On Monday, Saddam offered development projects if they would send their tankers to pick it up. The Iraqi president, however, on Tuesday, the Philippines' president, Ferdinand Marcos, said it had already failed to meet oil requirements, and Iraq said it would not "sell its sovereignty" for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of freedom as a "propaganda plot."

Venezuela, an OPEC member, has led a drive among oil-producing nations to boost production to make up the shortfall caused by the U.N. oil embargo. Oil from the world market. Their oil makes up 20 percent of the world's oil reserves. Only Saudi Arabia has higher reserves. But according to the State Department, U.N. sanctions have not been lifted. It was last week, five weeks ago, that Constance said Tuesday.

Linguistic processing and Creation of semantic graph



Summarization via semantic graphs

Automatically generated document summary

Cracks appeared in the U.N. trade embargo against Iraq. According to a State Department survey, Cuba and Romania have struck oil wells in Iraq and companies elsewhere are trying to continue trade with Baghdad in defiance of the sanctions. Iran has agreed to an Iranian oil deal to buy up to 100,000 barrels of refined oil per month. That would come after the United States and Britain or Bahrain on the reported food-for-oil deal. But the source, who requested anonymity, said the deal will not affect Iraqi Foreign Minister Tariq Aziz's visit Sunday to Tehran, the first by a senior Iraqi official since the 1980-88 gulf war. After the visit, the two countries announced they would resume diplomatic relations. Well-informed oil industry sources in the region, contacted by AP, said that although Iraq is a major oil exporter itself, it currently has to import about 150,000 barrels of refined oil a day for domestic use because of damages suffered in the gulf war. Also in similar lines, ABC News reported that following Aziz's visit, Iran is prepared to buy more than all the refined oil available for the dollar. In addition, the United States' James A. Baker III, meanwhile, met in Moscow with Soviet Foreign Minister Eduard Shevardnadze, two days after the U.S.-Soviet summit that produced a joint demand that Iraq withdraw from Kuwait. During the summit, Bush encouraged Mikhail Gorbachev to withdraw 190 Soviet military specialists from Iraq, where they remain to fulfill contracts. Shevardnadze told the Soviet press yesterday the specialists had been withdrawn. "A war is a war, justified or not, 5,800 Soviet citizens in Iraq," his speech, "but we must not go out to fight the war of the invaders of our country." "Our position is not to change, and it will not change. America and the world will not be blackmailed," the president added. "Vital issues of principle are at stake. Saddam Hussein is literally trying to wipe a country off the face of the Earth." In other developments, a U.S. diplomat in Baghdad said Tuesday up to 800 Americans and Britons will fly out of Iraq equipped with their families, mostly women and children, leaving them behind. Saddam has said he is keeping foreign men as human shields against attacks. On Monday, a plane of 164 Westerners was based in Basra from Iraq. Estimates of the number of foreigners in Iraq range from 100,000 to 150,000. A soldier can rape a father's daughter in front of him and he can't do anything about it," the State Department said Iraq had told U.S. officials that American male citizens in Iraq will be allowed to leave the country if they want to. "The American public needs to understand how many men the Iraqi move could affect." A Pentagon spokesman said American forces will be sent to the region to help detect and identify individuals near its borders with Turkey and Syria. He said there will be little indication hostilities are imminent. Defense Secretary Dick Cheney said the cost of the U.S. military buildup in the Middle East was rising above the \$1 billion-a-month estimate generally used by government officials. He said the total cost, if no shooting war breaks out, could total \$1.5 billion in the next fiscal year beginning Oct. 1, "a significant increase" in help from Arab nations and other U.S. allies. In addition, the U.S. has been accused of ordering its forces to the Persian Gulf in response to a \$1 billion deployment, Jordan, hit hardest by the U.N. prohibition on trade with Iraq. "The pressure from abroad is getting so strong," said Hirooysu Horio, an official with the Ministry of International Trade and Industry. Local news reports said the aid would be extended through the World Bank and International Monetary Fund, and \$600 million would be set aside as part of the U.S.-Soviet agreement. On Tuesday, Treasury Secretary Nicholas Brady visited Tokyo to seek a waiver seeking \$10.5 billion to help Egypt, Jordan and Turkey. The U.S. already has promised \$3 billion to help those countries meet their losses after vehicles and prefabricated housing for non-military uses. But critics in the United States have said Japan moves because its economy depends heavily on oil from the Middle East. Japan imports 99 percent of its oil. Japan's constitution bars the use of force in settling international disputes and Japanese law restricts the right to Japanese territory, except in self-defense. On Monday, Saddam offered development projects if they would send their tankers to pick it up. The Iraqi president, however, on Tuesday, the Philippines' president, Ferdinand Marcos, said it had already failed to meet oil requirements, and Iraq said it would not "sell its sovereignty" for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of freedom as a "propaganda plot."

Venezuela, an OPEC member, has led a drive among oil-producing nations to boost production to make up the shortfall caused by the U.N. oil embargo. Oil from the world market. Their oil makes up 20 percent of the world's oil reserves. Only Saudi Arabia has higher reserves. But according to the State Department, U.N. sanctions have not been lifted. It was last week, five weeks ago, that Constance said Tuesday.

Sentence extraction rule

Select a sub-graph that would characterize extracted summaries



Learn the sub-graph selection model

Detailed Summarization

Linguistic analysis of the text

- Deep parsing of sentences

Refinement of the text parse

- Named-entity consolidation

Determine that 'George Bush' = 'Bush'
= 'U.S. president'

- Anaphora resolution

Link pronouns with name-entities

Extract Subject-Predicate-Object triples

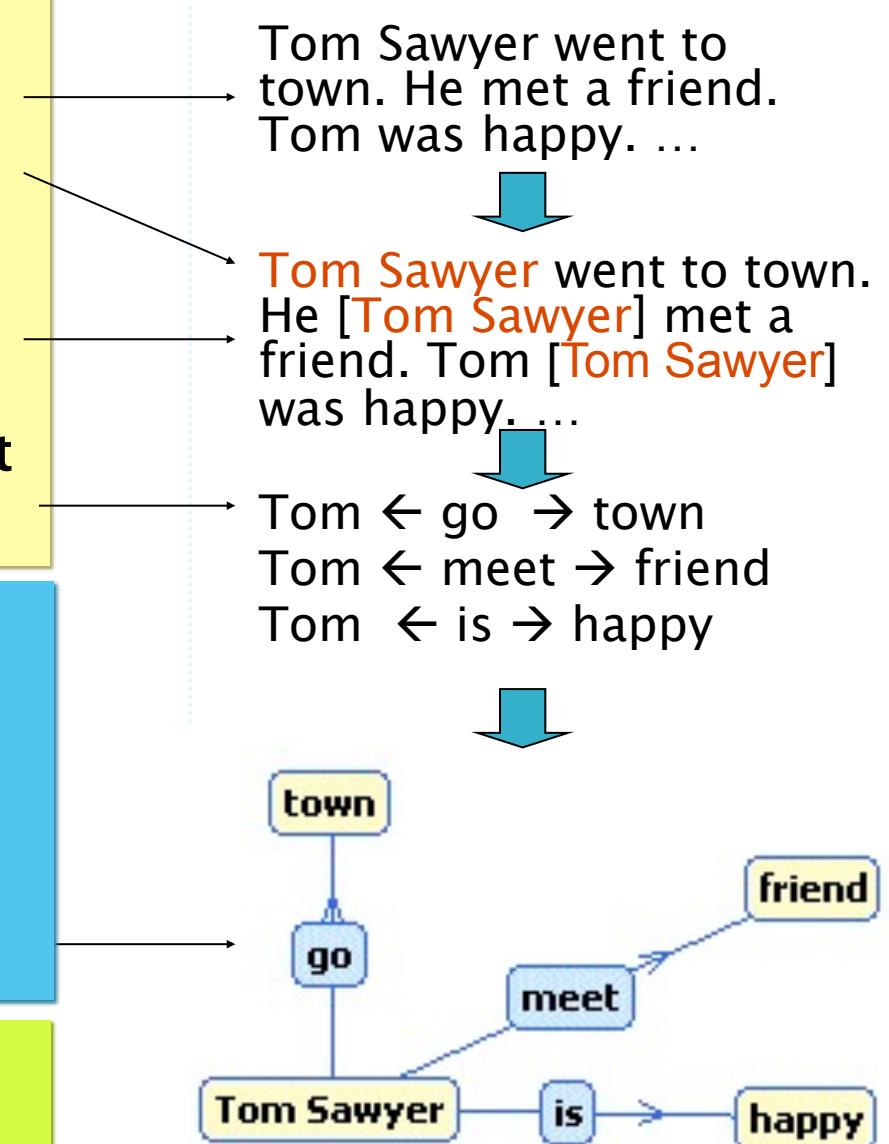
Compose a **graph** from triples

Describe each triple with a set of features for learning

Learn a model to classify triples into the summary

Generate a **summary graph**

Use summary graph to generate textual document summary



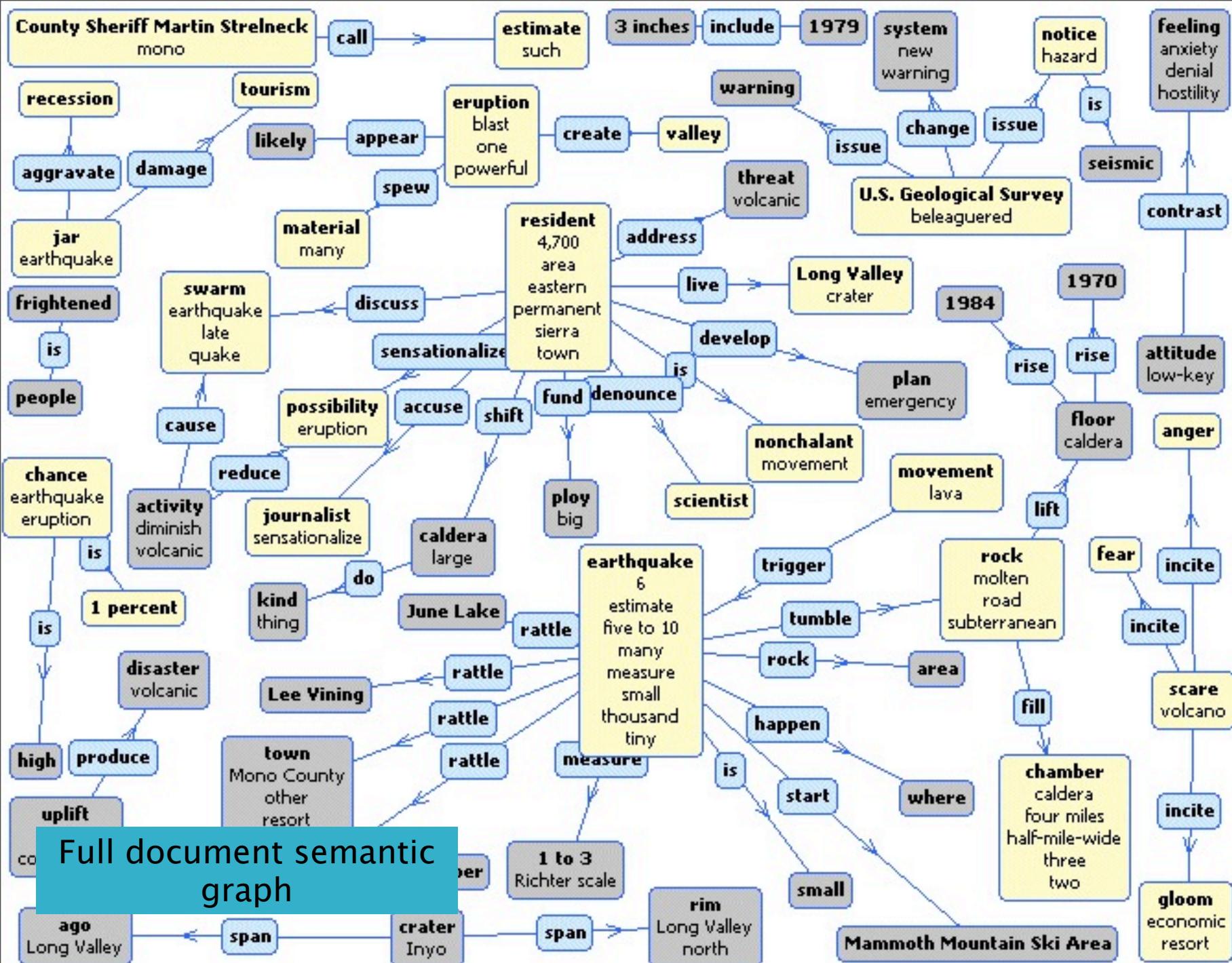
Example of automatic summary

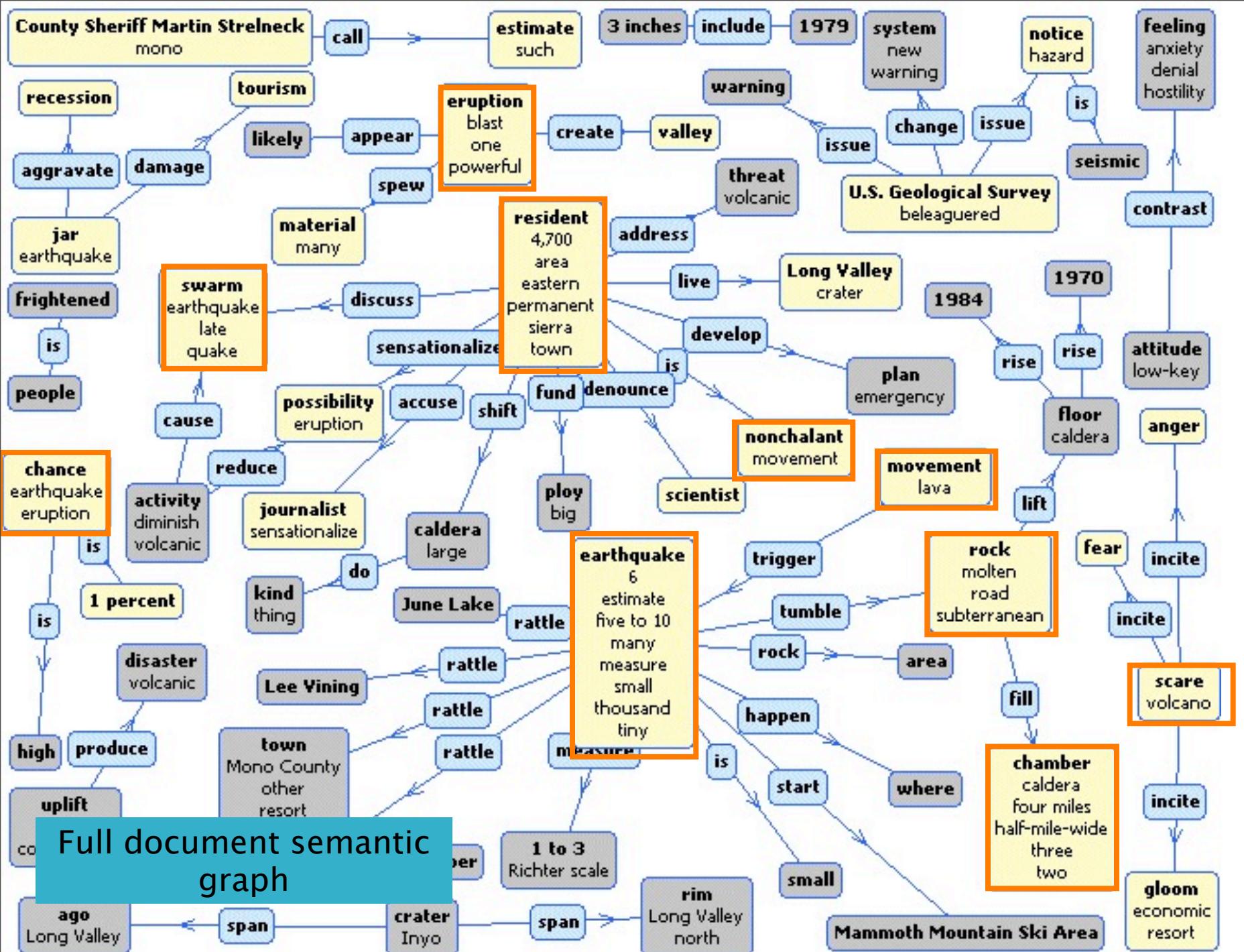
Cracks Appear in U.N. Trade Embargo Against Iraq.

Cracks appeared Tuesday in the U.N. trade embargo against Iraq as Saddam Hussein sought to circumvent the economic noose around his country. Japan, meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that it is not doing its share, Bush met with Congress and a nationwide radio and television audience that ``Saddam Hussein will fail'' to make his conquest of Kuwait permanent. ``America must stand up to aggression, and we will," said Bush, who added that the U.S. military may remain in the Saudi Arabian desert indefinitely. ``I cannot predict just how long it will take to convince Iraq to withdraw from Kuwait," Bush said. More than 150,000 U.S. troops have been sent to the Persian Gulf region to deter a possible Iraqi invasion of Saudi Arabia. Bush's aides said the president would follow his address to Congress with a televised message for the Iraqi people, declaring the world is united against their government's invasion of Kuwait. Saddam had offered Bush time on Iraqi TV. The Philippines and Namibia, the first of the developing nations to respond to an offer Monday by Saddam of free oil — in exchange for sending their own tankers to get it — said no to the Iraqi leader. Saddam's offer was seen as a none-too-subtle attempt to bypass the U.N. embargo, in effect since four days after Iraq's Aug. 2 invasion of Kuwait, by getting poor countries to dock their tankers in Iraq. But according to a State Department survey, Cuba and Romania have struck oil deals with Iraq and companies elsewhere are trying to continue trade with Baghdad, all in defiance of U.N. sanctions. Romania denies the allegation. The report, made available to The Associated Press, said some Eastern European countries also are trying to exchange food and medicine for reported food-for-oil deal.

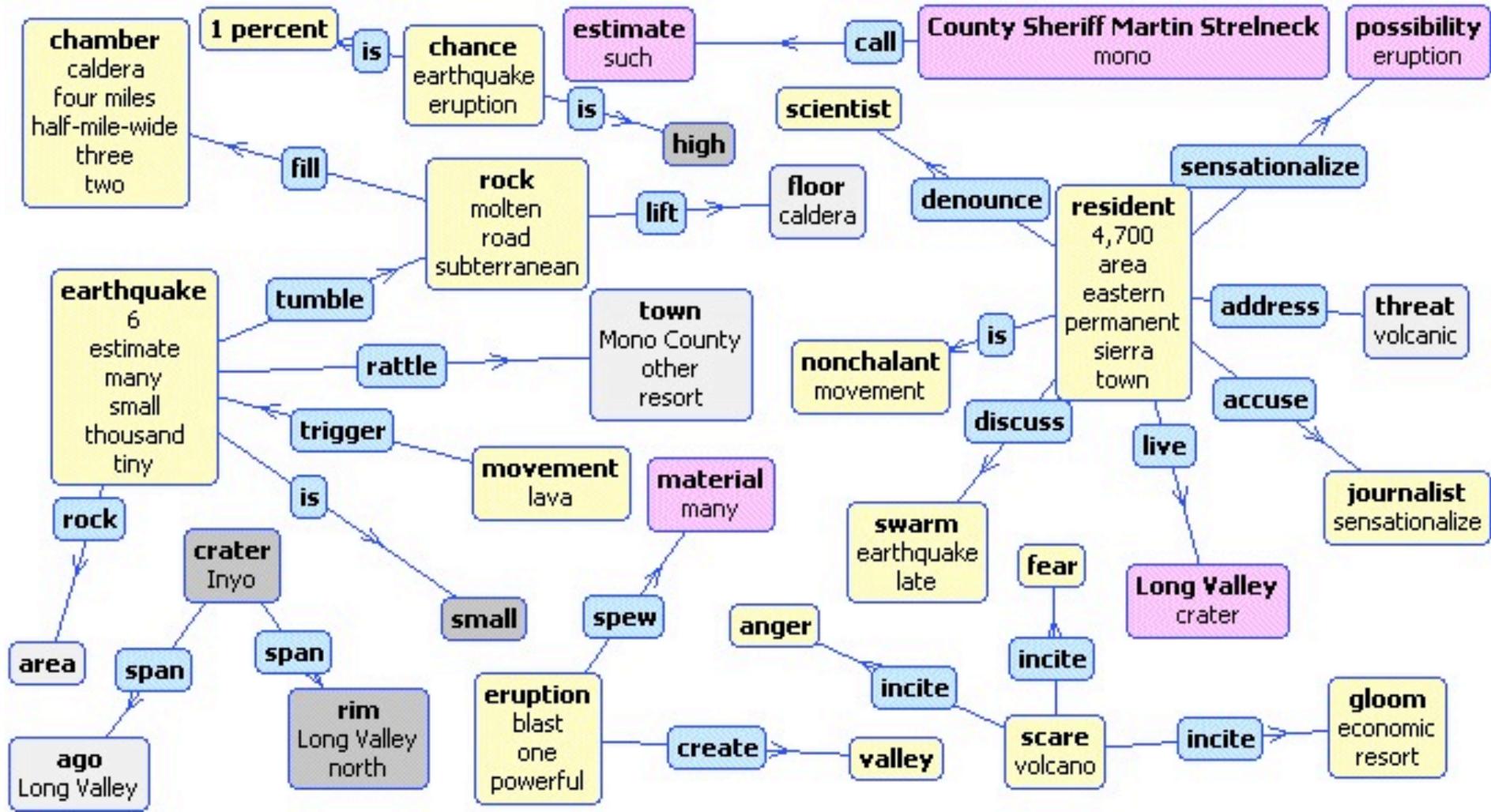
- Eight years after a volcano scare incited fear, anger and economic gloom in Sierra resorts, residents are nonchalant about renewed underground lava movement that is triggering thousands of tiny earthquakes.
- The resort town's 4,700 permanent residents live in Long Valley, a 19-mile-long, 9-mile-wide volcanic crater known as a caldera.
- The Earth's crust is being stretched apart in the region, allowing molten rock to fill half-mile-wide chambers under the caldera.
- The valley was created 730,000 years ago by one of Earth's most powerful eruptions, a blast that spewed 600 times more material than the May 1980 eruption of Mount St. Helens in Washington state.
- Despite the current activity, the probability of a major earthquake or a volcanic eruption in the area is ``less than 1 percent each year," said David Hill, the U.S. Geological Survey geophysicist in charge of research at Long Valley. Mono County Sheriff Martin Strelneck called such estimates ``a scientific guessing game," and said area residents rarely discuss the latest swarm of earthquakes, which started in May 1989.
- As a result, the Geological Survey issued a ``notice of potential volcanic hazard" for Long Valley in May 1982.
- That warning, coupled with jarring earthquakes, damaged tourism and aggravated a recession in the once-booming real estate market.

Human extracted
summary

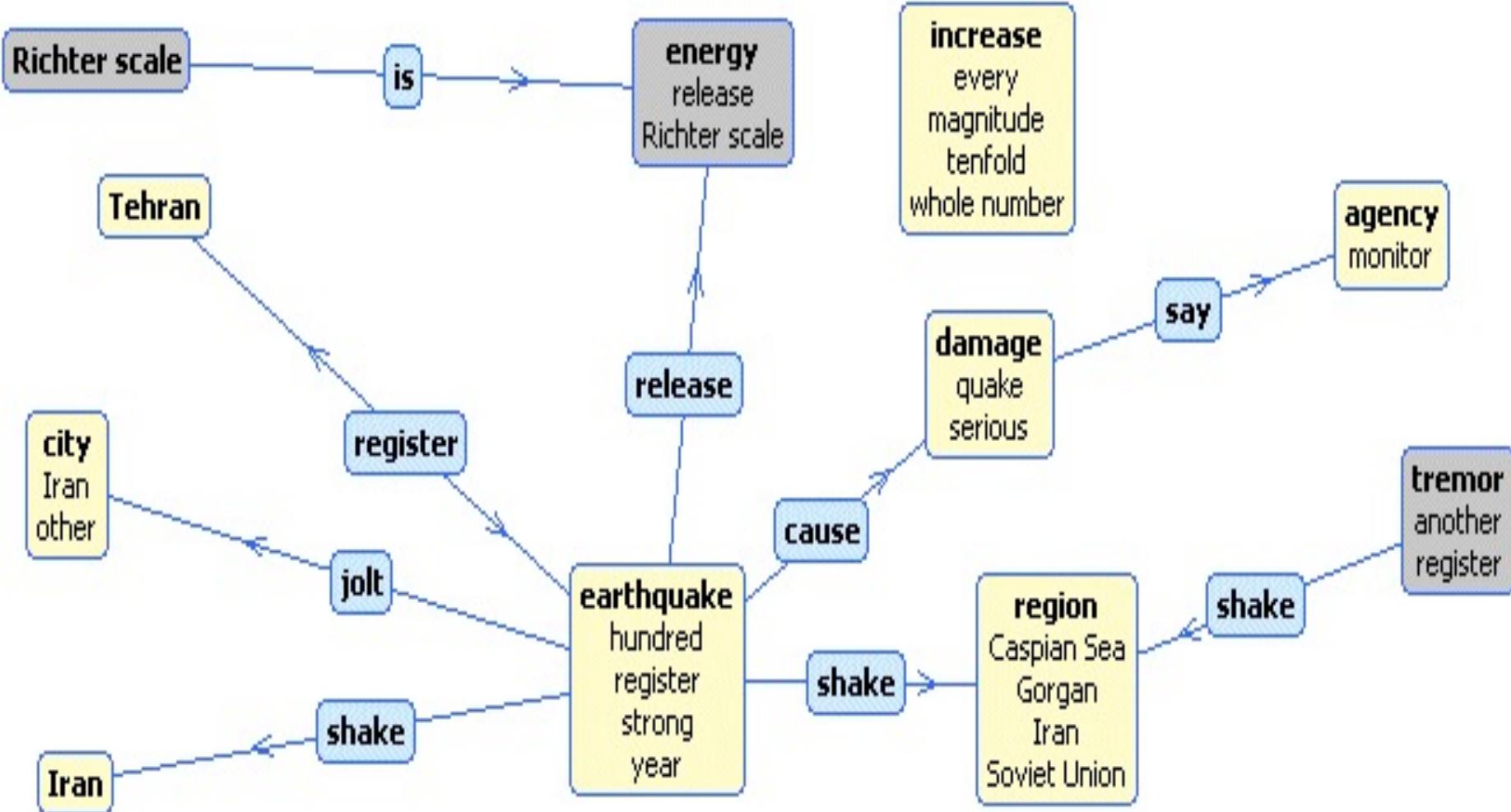


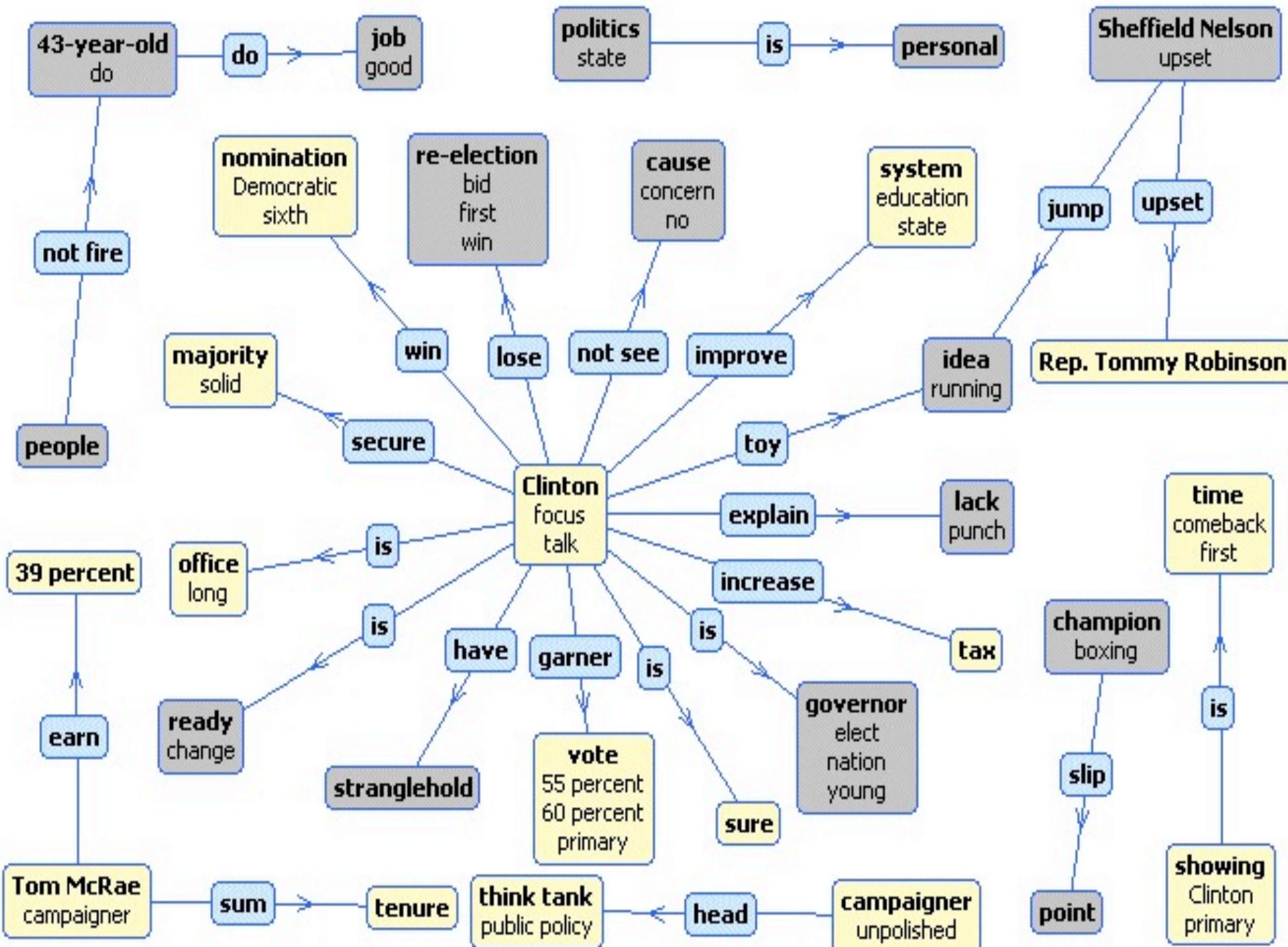


Automatically generated summary graph



More examples





Question Answering

<http://AnswerArt.net>

Asking a question in natural language

AnswerArt - Mozilla Firefox

Ble Edit View History Bookmarks Tools Help

http://answerart.net/ Google

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

AnswerArt

answer Art

Art of finding answers in document collection

who declared a war?

Ask

We found that

Eddie Murphy played the following characters: miles, League, season, remainder, basketball, man, Chandler, Jarrett, drums, negotiator, Ed, prince, dragon, Clarence.

Try this:

Who declared war?
Did Bavaria declare war?
What is an embargo?
Who is Lenin?
Where do birds fly?

Find related information by asking

What did Eddie Murphy perform?
Who performed stand-up?
What did Eddie Murphy do stand-up?
Who is Eddie Murphy?
Who performed something?
Who did something to Eddie Murphy?

Try this:

Who needs some sunshine?
What did Clinton say?
Where did an earthquake happen?
What do people drink?
Was Clinton visiting Venezuela?

Make sense of a document contents in a second.

AnswerArt - Document Overview example

Try this:

Italian film director Ferreri dies in Paris.
Russian parliament opposes efforts to bury Lenin.
American Air jet turned back by bird in engine.
Bavaria declares war on canned beer.
Argentine weekly official crop progress.

About AnswerArt - Contact - Terms of Use - Interactive Document Analysis

Done

...result set is set of mentions of potential answers

AnswerArt - who declared a war? - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://answerart.net/qa/?q=who+declared+a+war%3F

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

AnswerArt - who declar...

who declared a war?

Ask

We found that

the following	declared	war
Israel	declaring, declares, declared	war, Middle East war
France	declared	war, Gulf War
plan	announced	Gulf War
Republican Army	announced, declare, declares	war, guerrilla war
president	declared	war
state	declaring	war
government	declares, declaring	war
We	declared	war
President Boris Yeltsin	declared	war

Related documents

Israel EGYPT: Arafat says Israel declares war on Palestinians. Israel is declaring war against the Palestinian people.

Israel EGYPT: Arafat says Israel declares war on Palestinians. Arafat says Israel declares war on Palestinians.

Israel ISRAEL: Arafat aide says Israel has declared war. Arafat aide says Israel has declared war.

Israel EGYPT: Moslem cleric says force last option for Jerusalem. Israel annexed East Jerusalem and declared all of the city its eternal and indivisible capital after the 1967 Middle East war.

Israel ISRAEL: Israel upbeat, PLO glum after U.S.-brokered talks. Israel annexed East Jerusalem and declared all of the city its eternal and indivisible capital after the 1967 Middle East war.

Done

Document browser with summarization

AnswerArt - Document Overview - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://answerart.net/docview/?view=1&id=19970802|810803newsML.xml

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

AnswerArt - Document ...

DOCUMENT OVERVIEW

Timeline:

```
graph TD; A[autumn] --> B[bombing]; B --> C[west Jerusalem]; C --> D[talks]; D --> E[attack]; E --> F[declared]; F --> G[Saturday Israel]; G --> H[declared]; H --> I[Palestinians]
```

President Yasser Arafat --> J[told]; J --> K[reporters]

FACTS:

- Israel declared war
- Saturday Israel declared war
- Saturday Israel declared Palestinians
- west Jerusalem took risks
- suicide bombing west Jerusalem
- President Yasser Arafat told reporters
- President Yasser Arafat told talks

Show Document Overview

Summary:

Palestinian President Yasser Arafat said on Saturday Israel had declared war on Palestinians rather than "terrorism" by a series of measures it took in the wake of a double suicide bombing in west Jerusalem. "The Israeli government used this situation and this incident (the suicide bombing) to declare war on the Palestinian people, then National Authority and their leadership instead of declaring it against terrorism," Arafat told reporters after talks with Egyptian President Hosni Mubarak.

Arafat arrived in Egypt earlier on Saturday and went straight into talks with Mubarak on the crisis in the Middle East peace talks, Egyptian state television said.

"I told President Mubarak about all the dangerous measures announced by Israel after the terrorist operation that happened recently in Jerusalem," Arafat said. He has since returned to Gaza.

Arafat described as "collective punishment" the measures taken by Israel since the suicide bombing and said Mubarak told him he would "begin to ease the suffering of the Palestinian people," but did not give details.

Egypt has been trying to jumbo-start Israeli-Palestinian peace talks which collapsed in March after Israel began to build a new Jewish settlement in Arab East Jerusalem.

Arafat, who met Mubarak in the Egyptian leader's private residence in Burg al-Arab, near the Mediterranean city of Alexandria, was accompanied by senior Palestinian officials Saad...

Raise document overview by dragging the bottom right corner.

Or click this icon to decrease.

Done

The screenshot shows a Mozilla Firefox window displaying a document from AnswerArt. The main area is titled 'DOCUMENT OVERVIEW' and features a timeline diagram with nodes like 'autumn', 'bombing', 'west Jerusalem', 'talks', 'attack', 'declared', 'Saturday Israel', and 'Palestinians'. It also shows a sequence involving 'President Yasser Arafat' and 'reporters'. To the right, a sidebar lists 'FACTS' such as 'Israel declared war' and 'Suicide bombing west Jerusalem'. Below the timeline is a 'Summary' section with text from Palestinian President Yasser Arafat. A green box at the bottom right contains a summary of the text above, with instructions on how to resize it. The bottom left corner shows the logo of the 'Jožef Stefan Institute'.

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
-

Lexical

- ▶ Vector–space model
 - ▶ Language models
 - ▶ Full–parsing
 - ▶ Cross–modality
-
- ▶ Collaborative tagging / Web2.0
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Syntactic

Semantic

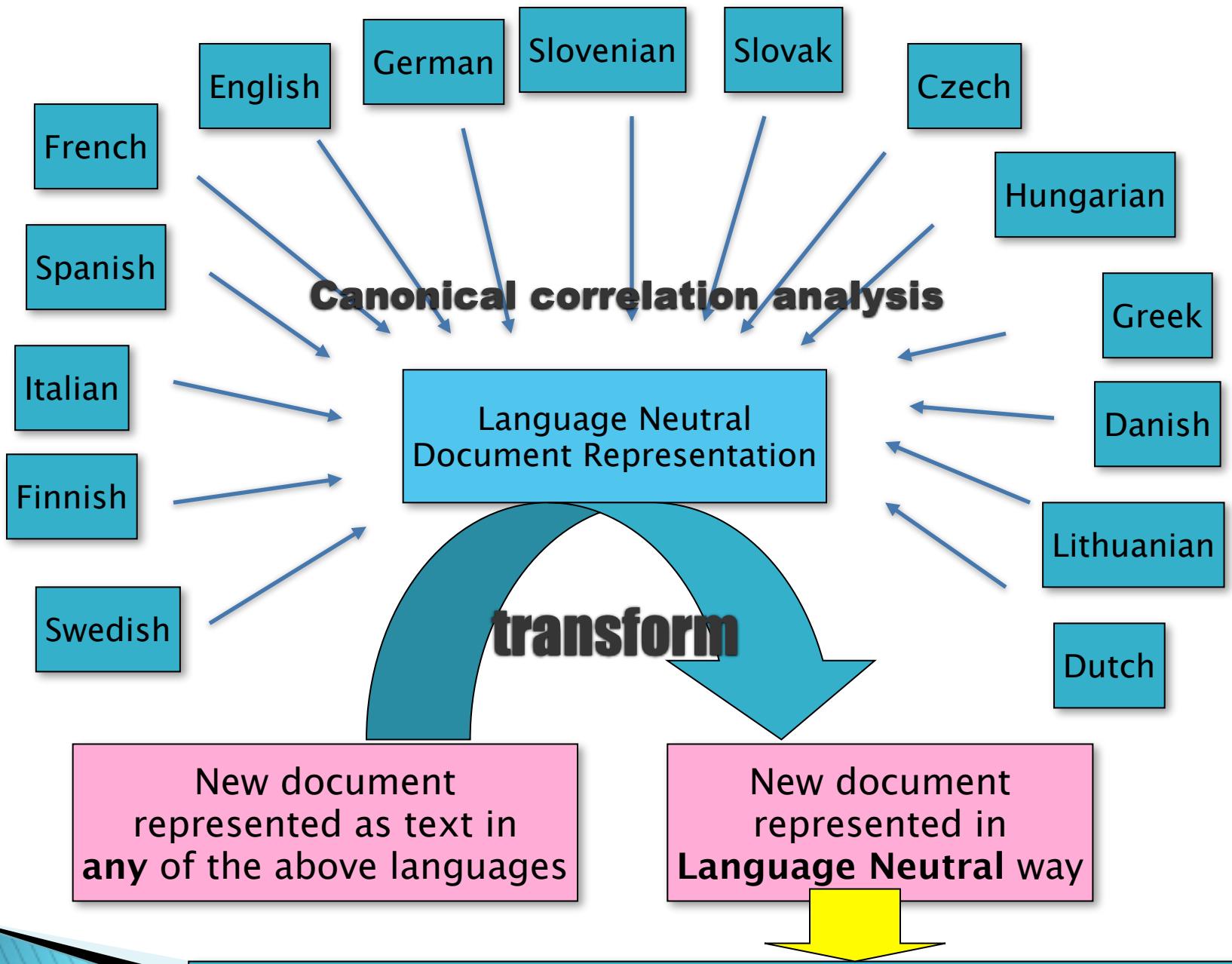
Cross-modality level

- ▶ It is 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

Multilingual search

<http://www.smart-project.eu/>





Example of cross-lingual information retrieval on Reuters news corpus (using KCCA)

Bili Bau Bau - Mozilla Firefox

File Edit View Go Bookmarks Tools Help | Slovenija Svet Gospodarstvo Delo CNet InfoWorld > ↗

http://127.0.0.1:8080/search?n=horse&l=de

Bili Bau Bau - Mozilla Firefox

File Edit View Go Bookmarks Tools Help | Slovenija Svet Gospodarstvo Delo CNet InfoWorld > ↗

http://127.0.0.1:8080/search?q=stock+exchange&l=en

stock exchange

en

de

Bili Search

Weight	Document Name
0.310658	
REPUBLIC OF IRELAND	
Countyglen Irish Stock Exchange	
The Irish Stock Exchange	
with effect from 1500 GMT	
-- Dublin Newsroom +353	
0.254181	
MACEDONIA: ONE COMPANY TRADES	
ONE COMPANY TRADES	
Stock turnover and traded	
MSE said.	
Total turnover rose to 296	
volume grew to 159 shares	
Only Gradski Tarnovski C	
Done	
0.311615	143935.en.txt
REPUBLIC OF IRELAND: Countyglen Irish Stock Exchange listing cancelled.	
Countyglen Irish Stock Exchange listing cancelled.	
The Irish Stock Exchange said in a statement on Friday that the listing for Countyglen Plc had been cancelled	
with effect from 1500 GMT (1600 local time).	
-- Dublin Newsroom +353 1 676 9775	
0.240322	328995.de.txt
BORSE ANDERT PLANE FÜR OPTIONSSCHEINHANDEL.	
Frankfurt (Reuter) - Die Deutsche Börse AG hat ihre Pläne für den elektronischen Handel mit Optionsscheinen geändert. Das ursprünglich angedachte Computersystem OHS werde nicht verwirklicht, weil die Anforderungen der Marktteilnehmer hier nicht zu sinnvollen Investitions- und Betriebskosten hätten verwirklicht werden können, erklärte die Börse am Freitag in Frankfurt. Stattdessen soll der elektronische Handel mit Optionsscheinen im Rahmen des voralberten allgemeinen Elektronischen Handelsystems (EHS) der Börse realisiert und im Laufe des	
Done	

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
- ▶ Linked Data
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Lexical

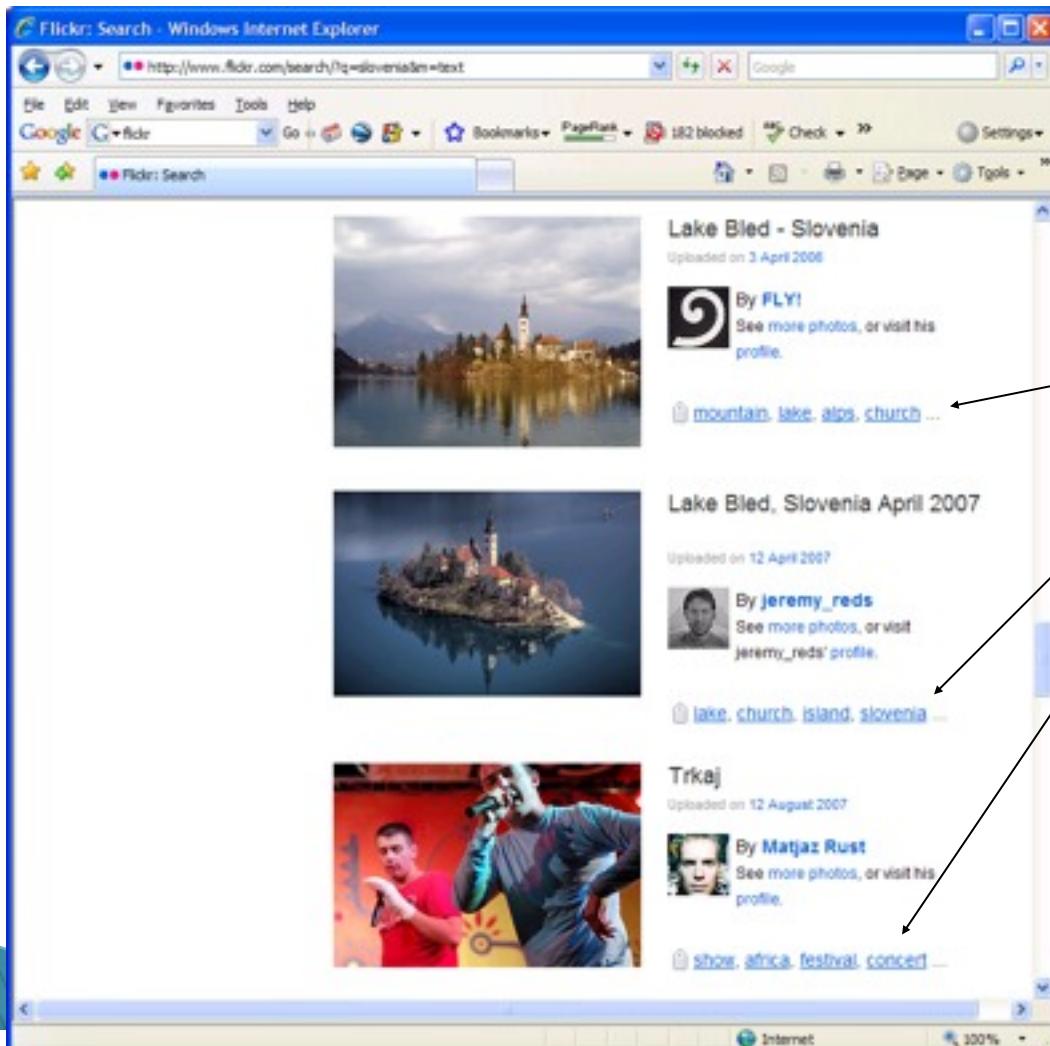
Syntactic

Semantic

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

del.icio.us search for "textmining" - Windows Internet Explorer
http://del.icio.us/search/?f=del_ocio_us&p=textmining&ttype=all

File Edit View Favorites Tools Help
Google C... Go Bookmarks PageRank 182 blocked Check Settings
del.icio.us search for "textmining"

del.icio.us / search popular | recent
login | register | help

Search results for textmining textmining del.icio.us search

Related tags: textmining datamining search nlp text software research java bioinformatics programming
showing 1 - 10 of 1378
« previous | next »

Text Analytics Solutions from ClearForest save this
to textmining datamining fed search semantic ... saved by 104 people

» Text mining the New York Times | Emerging Technology Trends | ZDNet.com save this
to textmining datamining research text ... saved by 113 people

GATE, A General Architecture for Text Engineering save this
to nlp java opensource information_extraction language ... saved by 104 people

Text mining - Wikipedia, the free encyclopedia save this
to textmining datamining wikipedia mining ai ... saved by 93 people

press release @ the bren school of information and computer sciences save this
to datamining textmining fed search mining ... saved by 80 people

Topic Modeling Toolbox save this
to matlab datamining textmining tools topic ... saved by 76 people

text-mining.org save this
to textmining research text_mining fed text-mining ... saved by 34 people

text-mining.org save this
to textmining datamining fed nlp analysis ... saved by 53 people

Text-Garden -- Text-Mining Software Tools save this
to textmining datamining clustering tools software ... saved by 57 people

KH Coder Index Page save this
to textmining software テキストマイニング datamining text ... saved by 54 people

« previous | next »

Internet 100%

Tags entered
by users
annotating
Web sites

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
- ▶ **Linked Data**
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Lexical

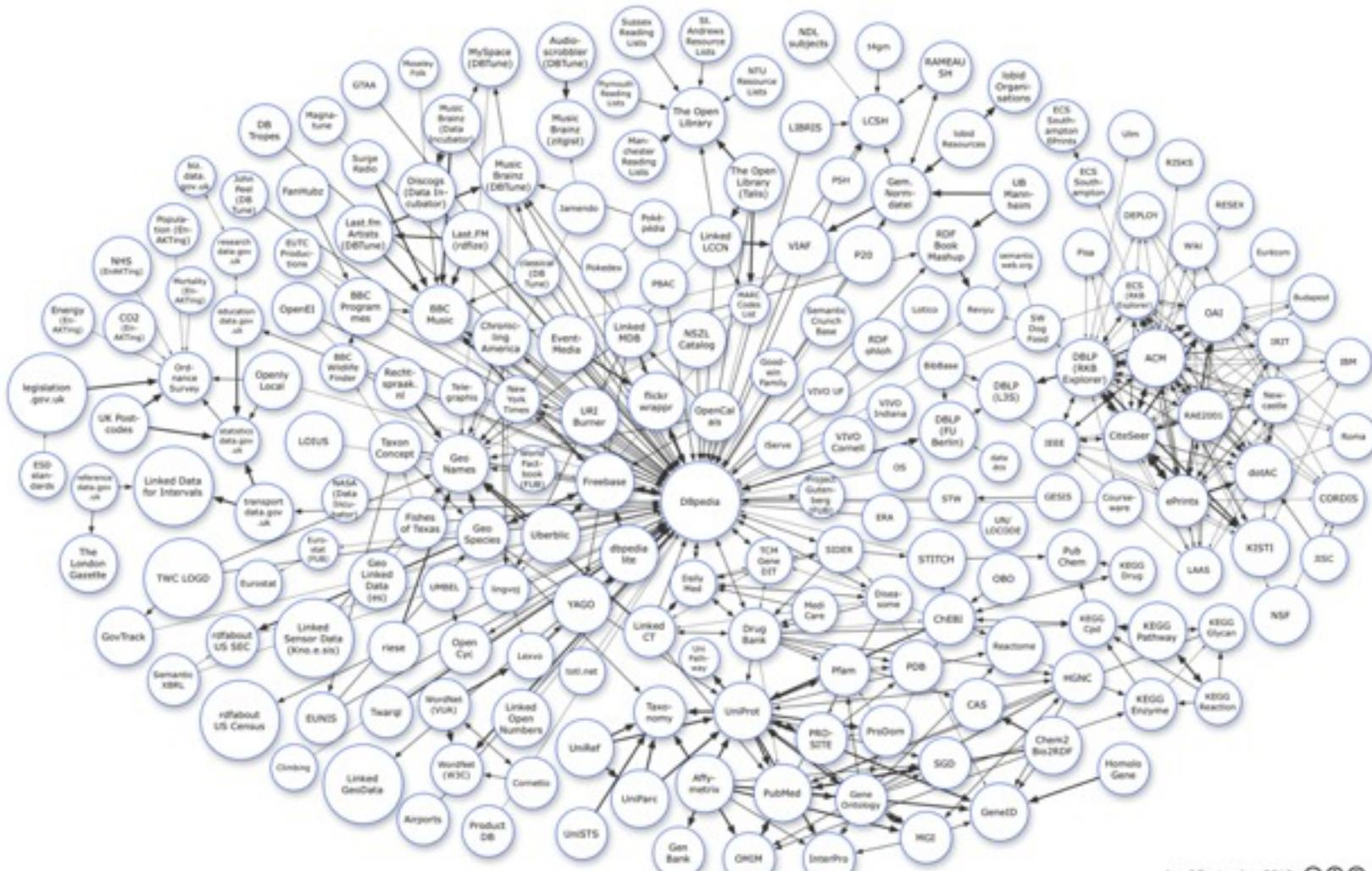
Syntactic

Semantic

The idea of Linked Data (Web-of-Data)

- ▶ The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data
- ▶ Four basic rules how to publish the data:
 - Use URIs as names for things
 - Use HTTP URIs so that people can look up those names
 - When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
 - Include links to other URIs, so that they can discover more things
- ▶ <http://linkeddata.org/>

Linked Open Graph



As of September 2010

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
- ▶ Linked Data
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Lexical

Syntactic

Semantic

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 template / frames approaches:
 - FrameNet: <http://framenet.icsi.berkeley.edu/>
 - TextRunner: <http://www.cs.washington.edu/research/textrunner/reverbdemo.html>

Examples of simple templates used by

- ▶ 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:
 - NP1 “such as” NPList2
 - NP1 “and other” NP2
 - NP1 “including” NPList2
 - NP1 “is a” NP2
 - NP1 “is the” NP2 “of” NP3
 - “the” NP1 “of” NP2 “is” NP3
- ▶ ...each template represents specific relationship between the words appearing in the variable slots
- ▶ From template patterns KnowItAll bootstraps new templates

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
 - ▶ Linked Data
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Lexical

Syntactic

Semantic

Ontologies level

- ▶ Ontologies are the most general formalism for describing data objects
 - ...in the recent years ontologies got popular through Semantic Web and OWL standard
 - Ontologies can be of various complexity:
 - ...from relatively simple ones (light weight described with simple)
 - ...to heavy weight (described with first order theories).
 - Ontologies could be understood also as very generic data-models where we can store extracted information from text

Cyc Knowledge Base and Reasoning

The Cyc Ontology

Cyc contains:

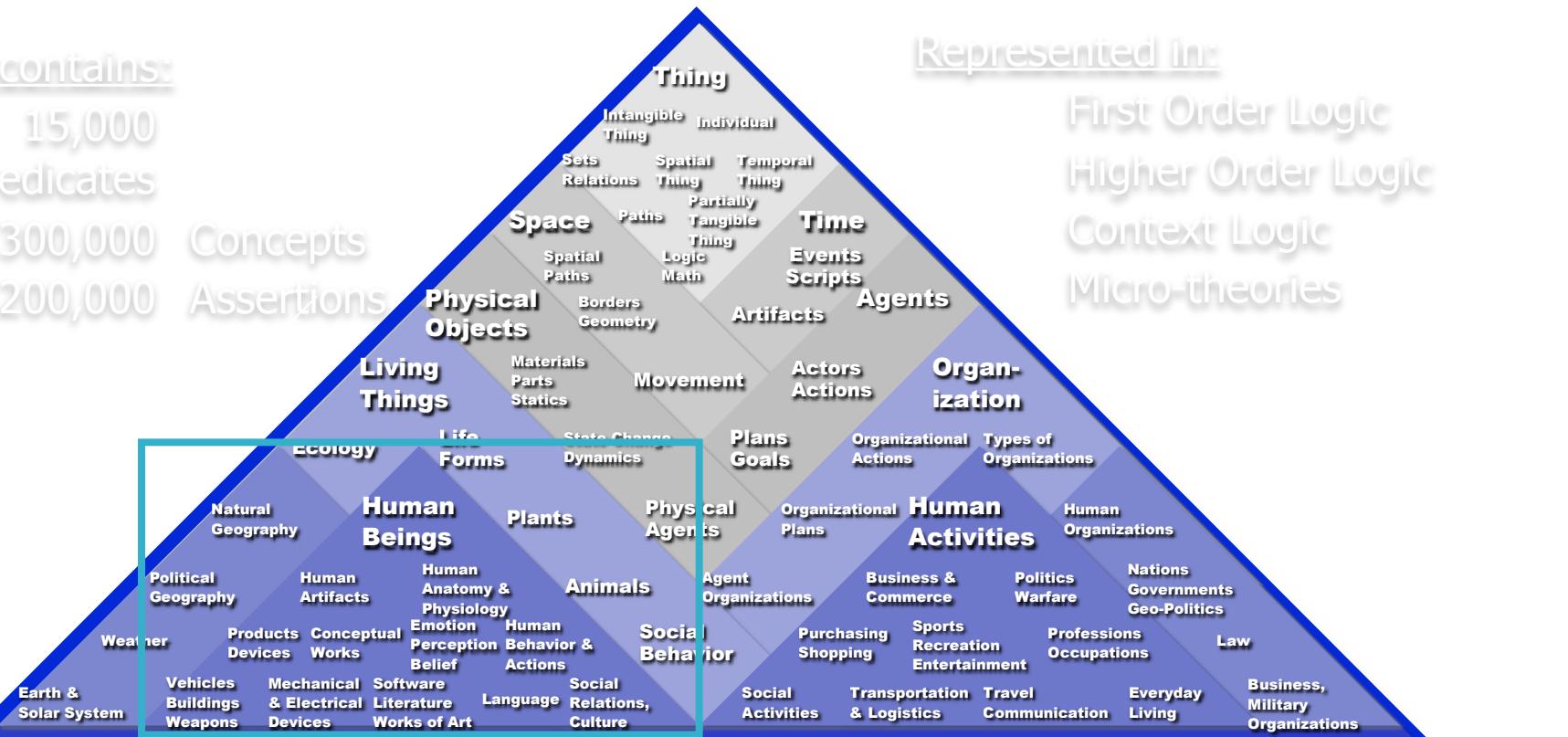
15,000
Predicates

300,000 Concepts

3,200,000 Assertions

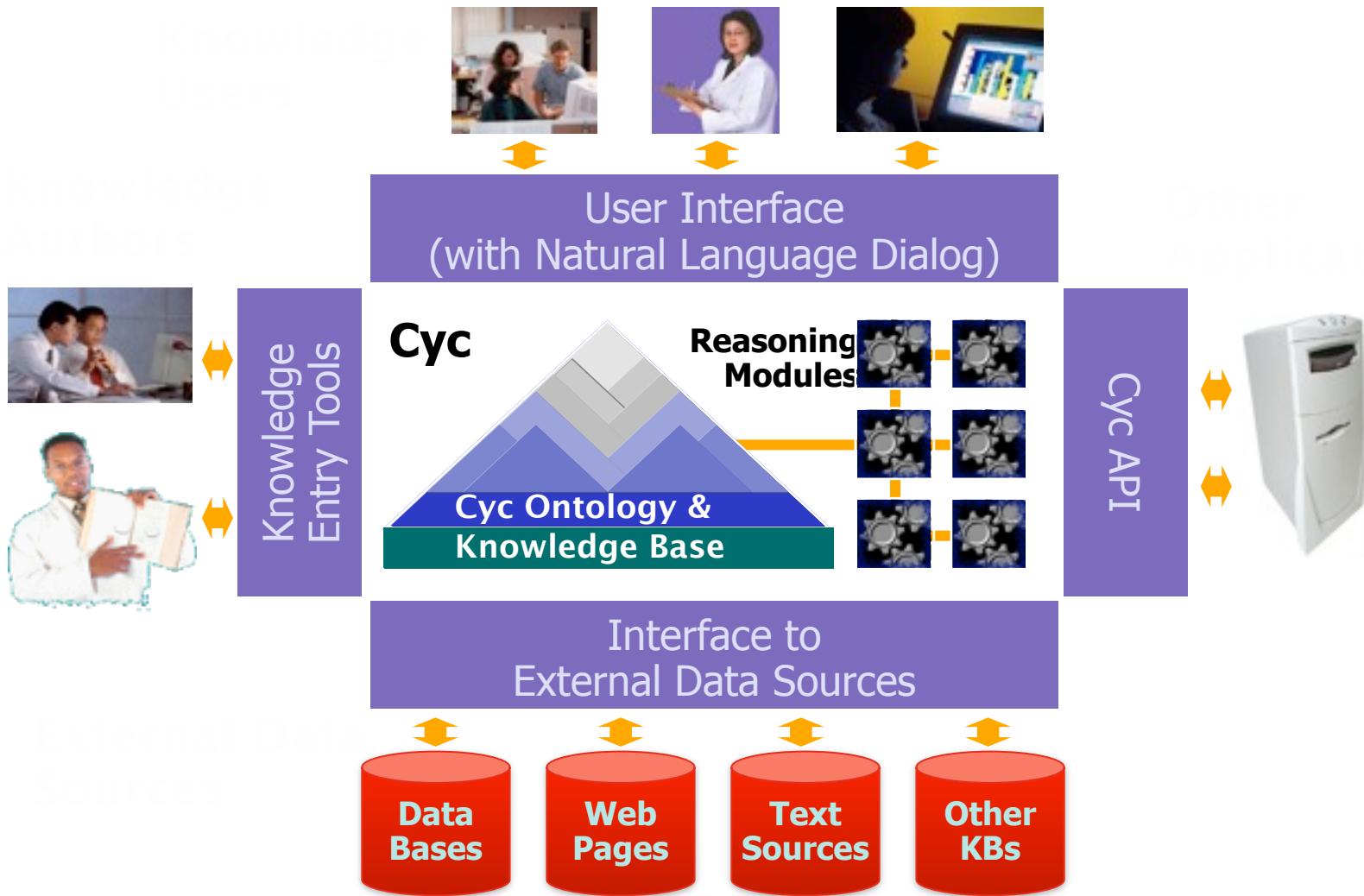
Represented in:

First Order Logic
Higher Order Logic
Context Logic
Micro-theories



General Knowledge about Various Domains

Specific data, facts, and observations



Cyc High-level Architecture



C. Matuszek, R.C. Kahlert
FACTory © Cycorp 2006

Cyc KB Extended w/ Domain



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))

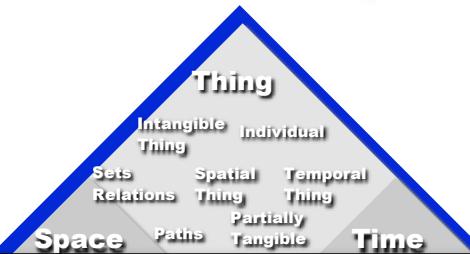
EARTH &
Solar System Buildings
Buildings & Weapons & Electrical
Devices Literature
Works of Art Language
Relations,
Culture SOCIAL
Activities TRANSPORTATION
& Logistics TRAVEL
Communication EVERYDAY
Living MILITARY
Organizations



General Knowledge about Terrorism

**Specific data, facts, and observations
about terrorist groups and activities**

Cyc KB Extended w/ Domain



Specific Facts about Al Qaida:

(basedInRegion AlQaida Afghanistan) Al-Qaida is based in Afghanistan.

(hasBeliefSystems AlQaida IslamicFundamentalistBeliefs) Al-Qaida has Islamic fundamentalist beliefs.

(hasLeaders AlQaida OsamaBinLaden) Al-Qaida is led by Osama bin Laden.

...

(affiliatedWith AlQaida AlQudsMosqueOrganization) Al-Qaida is affiliated with the Al Quds Mosque.

(affiliatedWith AlQaida SudaneseIntelligenceService) Al-Qaida is affiliated with the Sudanese Intell Service

...

(sponsors AlQaida HarakatUIAnsar) Al-Qaida sponsors Harakat ul-Ansar.

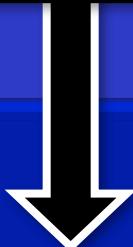
(sponsors AlQaida LaskarJihad) Al-Qaida sponsors Laskar Jihad.

...

(performedBy EmbassyBombingInNairobi AlQaida) Al-Qaida bombed the Embassy in Nairobi.

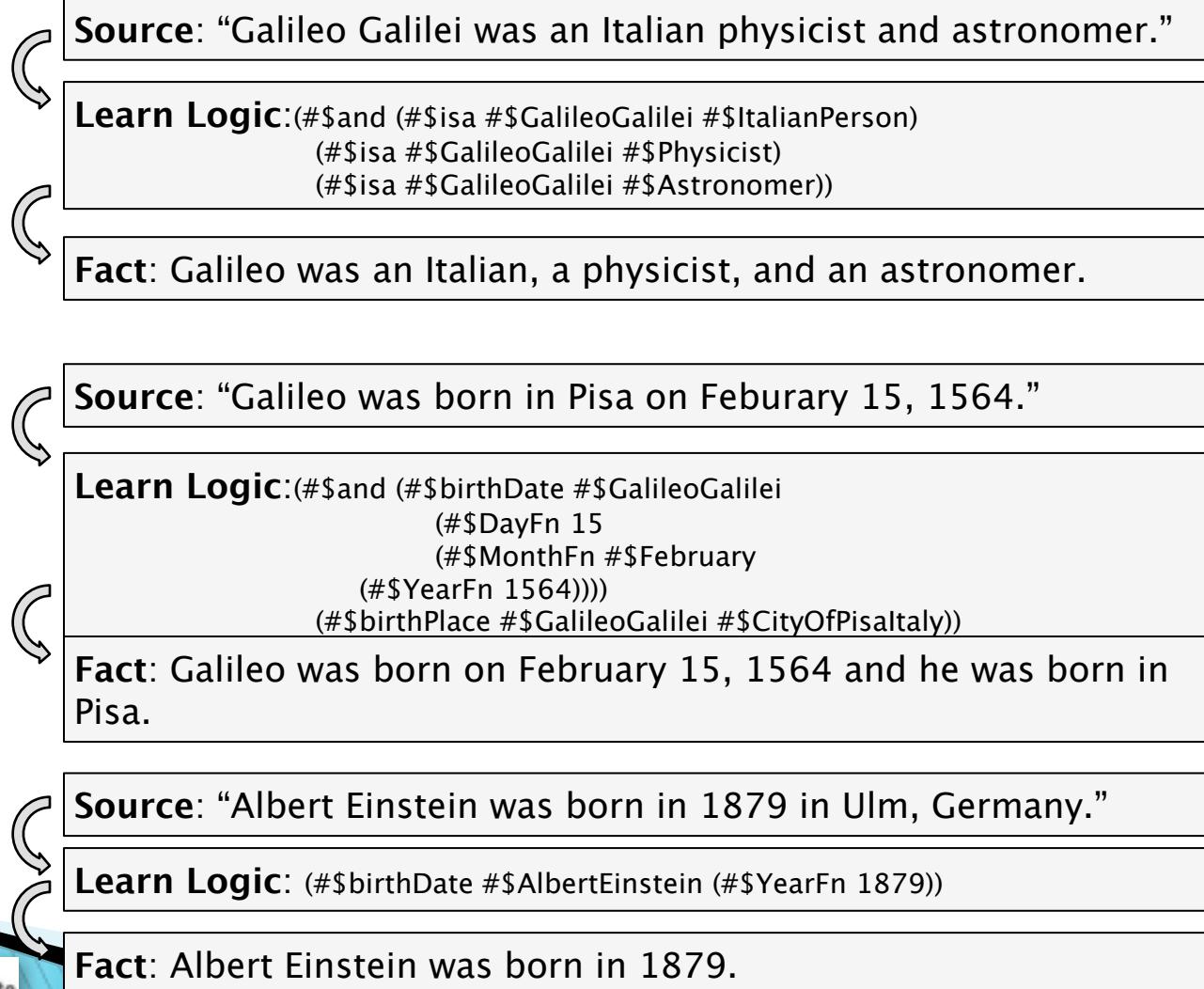
(performedBy EmbassyBombingInTanzania AlQaida) Al-Qaida bombed the Embassy in Tanzania.

General Knowledge about Terrorism



**Specific data, facts, and observations
about terrorist groups and activities**

Example of automatic translating text into Cyc Logic



Cyc Analytical Environment: General Intelligence Analysis using the CAE

File Edit Tools Window

Cyc's front-end: “Cyc Analytic Environment” – querying (1/2)

The screenshot shows the Cyc Analytic Environment interface. On the left, a sidebar lists various examples and queries. A specific query is highlighted: "Who has a motive for the assassination of Rafik Hariri?". This query is annotated with a callout "Text query" pointing to it. In the center, the query is displayed in a text box: "WHO had a motive for the assassination of Hariri.". This box is annotated with a callout "Query (semi) automatically translated in the First Order Logic". Below this, a table titled "Answers (5)" lists five potential motives with their speculation levels and sources. The table is annotated with a callout "Answers to the query" pointing to it.

Text query

WHO had a motive for the assassination of Hariri.

Query (semi) automatically translated in the First Order Logic

Answers (5)

Answer	Speculation Level	Sources
Bashar al-Assad	No Speculation	W CNN
Syria	Mildly Speculative	CNN M
al Qaeda	Moderately Speculative	SAIC CNN M
United States, the	No Speculation	2
Israel	No Speculation	2

Answers to the query

Status: Finished Message: No appropriate visualizations found

Cyc Analytical Environment: General Intelligence Analysis using the CAE

Cyc's front-end: “Cyc Analytic Environment” – justification (2/2)

File Edit Tools View Task Info Document Search Concepts Related-to Query Creator Queries Justification Justification Justification Proof 1 Save... Copy ▾ Options ▾ Options

Query: Who or what had a motive for the assassination of Hariri?
Answer: al Qaeda
Because:

Since 2000, Lebanon has been responsible for according with Lebanese economic reform. [1]

February 14, 2005 was the date of the assassination of Hariri. [2]

Rafik Hariri was killed during the assassination of Hariri.
Rafik Hariri is an advocate of Lebanese economic reform.

Al Qaeda opposes Lebanese economic reform.

Detailed Justification:
► Al Qaeda had a motive for the assassination of Hariri

External Sources:
Gary C. Gambill, "Dossier: Rafiq Hariri", United States Committee for a Free Lebanon, July 2001, http://www.meb.org/articles/0107_id1.htm.
"Huge blast kills Lebanese ex-PM", the Cable News Network, February 14, 2005, http://www.cnn.com/2005/WORLD/meast/02/14/beirut.explosion_1910/.

If

- some intelligent agent opposes some policy,
- and some other intelligent agent *VICTIM* is an advocate of that policy,
- and some other intelligent agent *ADOPTER* is responsible for according with the policy,
- and it is adopted by *ADOPTER* in any some *ADOPT-TYPE*,
- and some *ACT* prevents *VICTIM* from playing the role "key participant" in any *ADOPT-TYPE*,

then that intelligent agent had a motive for *ACT*.

Query & Answer

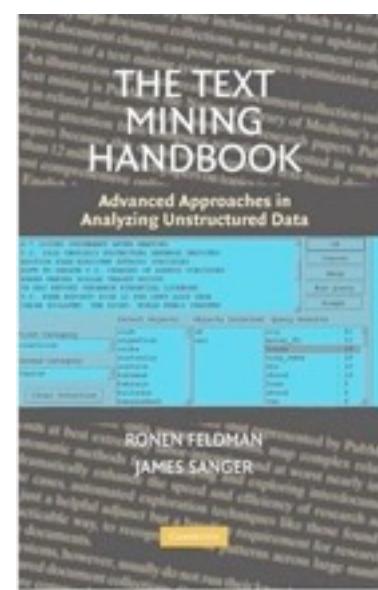
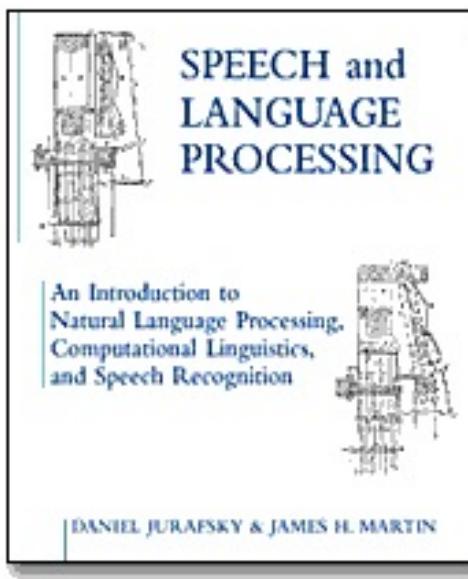
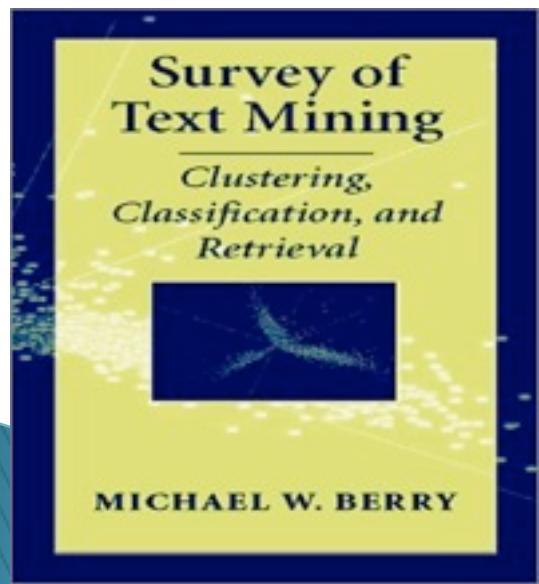
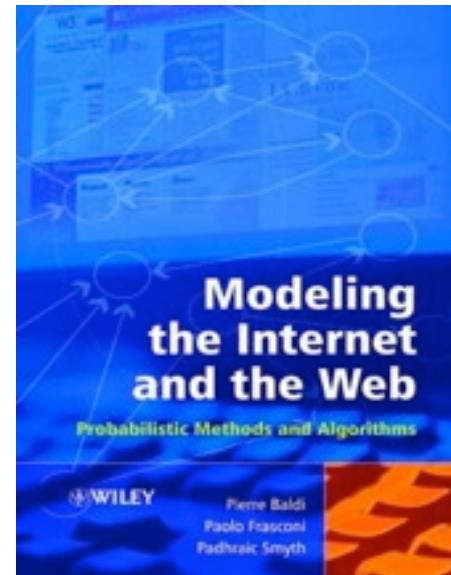
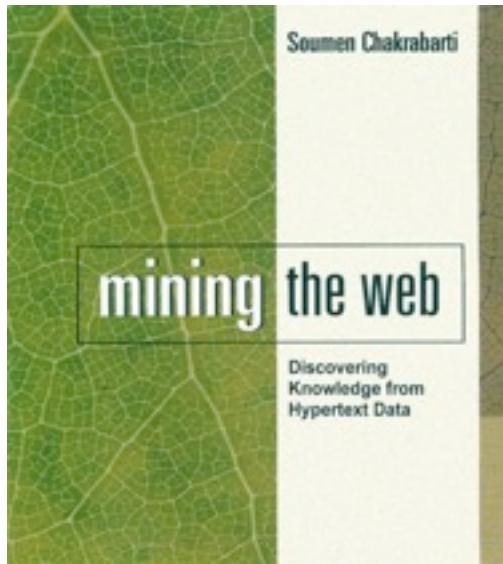
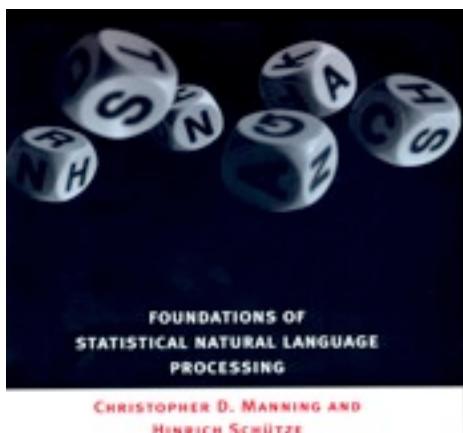
Justification

Sources for Reasoning and Justification

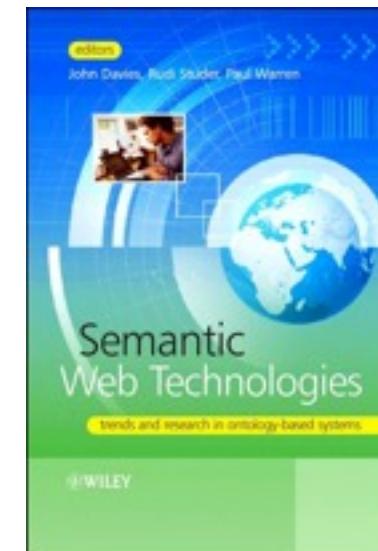
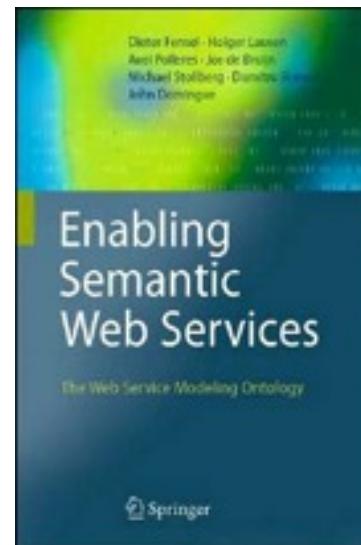
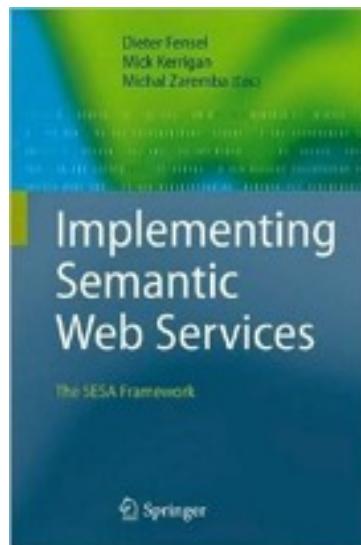
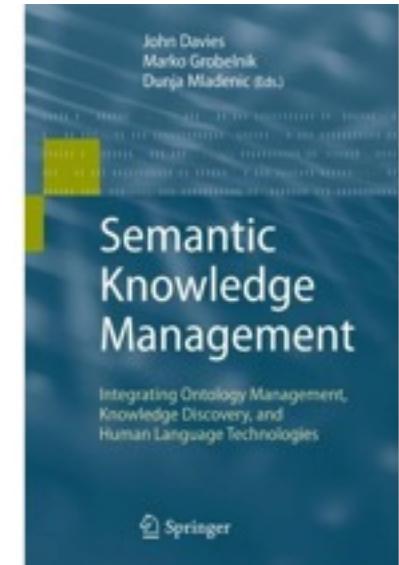
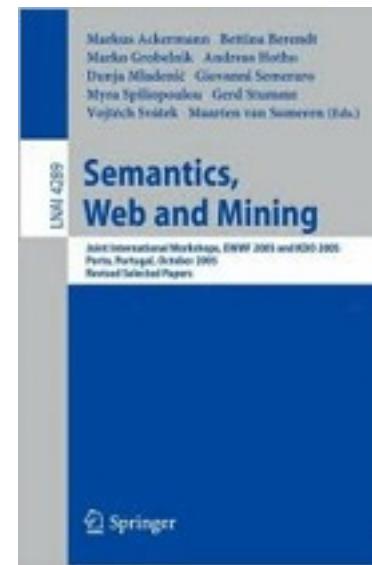
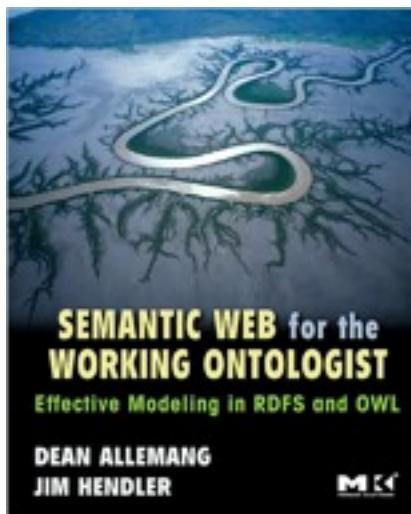
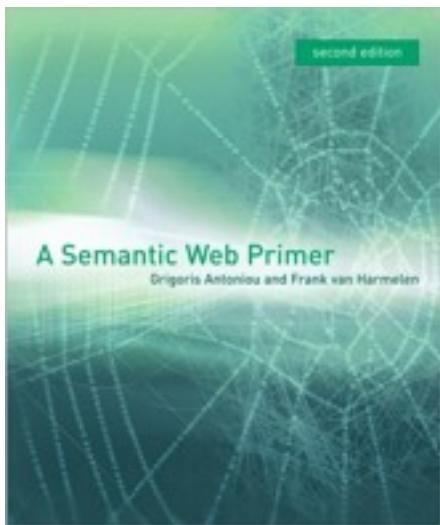


Further references...

References to some Text-Mining books



Books on Semantic Technologies



References to the main conferences

- ▶ **Information Retrieval:**
 - SIGIR, ECIR
- ▶ **Machine Learning/Data Mining:**
 - ICML, ECML/PKDD, KDD, ICDM, SDM
- ▶ **Computational Linguistics:**
 - ACL, EACL, NAACL
- ▶ **Semantic Web:**
 - ISWC, ESWC, ASWC

Videos on Text and Semantic Technologies

http://videolectures.net/Top/Computer_Science/

- ▶ Recorded tutorials, workshops, conferences, summer schools available from

<http://videolectures.net>

The screenshot shows the homepage of VideoLectures.net. At the top, there's a navigation bar with links for Home, Most Popular, Latest Lectures, Categories, Events, People, Interviews, Tutorials, and About Us. Below the navigation is a section titled "FEATURED LECTURES" with five thumbnail images. To the right of these thumbnails, there are brief descriptions of the lectures. Below this is a "RECENT EVENTS" section featuring "ISWC '08 - Karlsruhe" and "MIT OpenCourseWare Collection". Further down are sections for "NEWS", "CATEGORIES", and "FEATURED". The "CATEGORIES" section lists various academic fields with their respective counts. The "FEATURED" section includes links to "ESTC '08 - Vienna" and "Carnegie Mellon Machine Learning Lunch seminar". The footer contains a link to "Transferring data from maps.amung.us...".