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

"Jožef Stefan" Institute

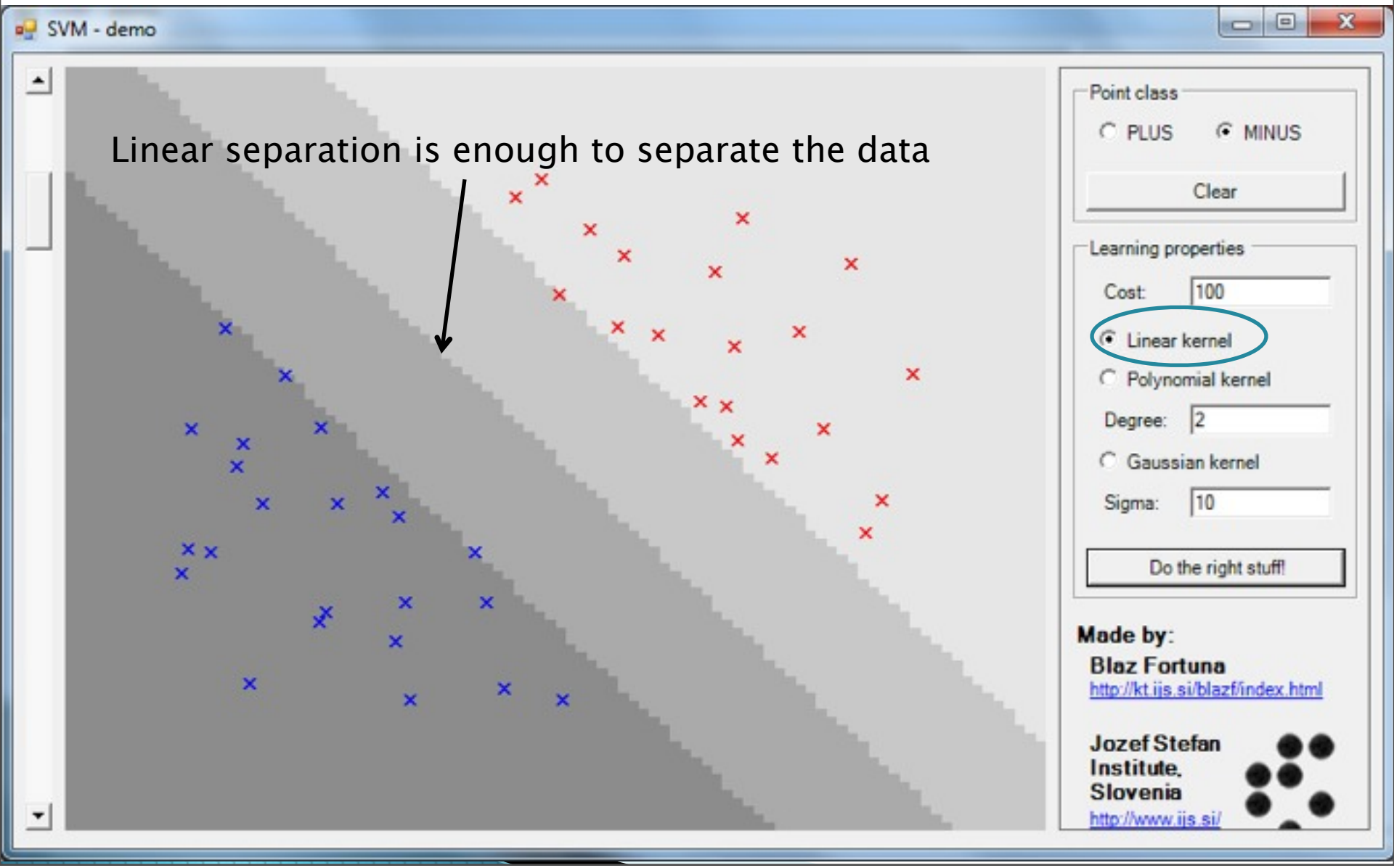
Some Initial thoughts

- 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 then **good algorithm with bad representation**

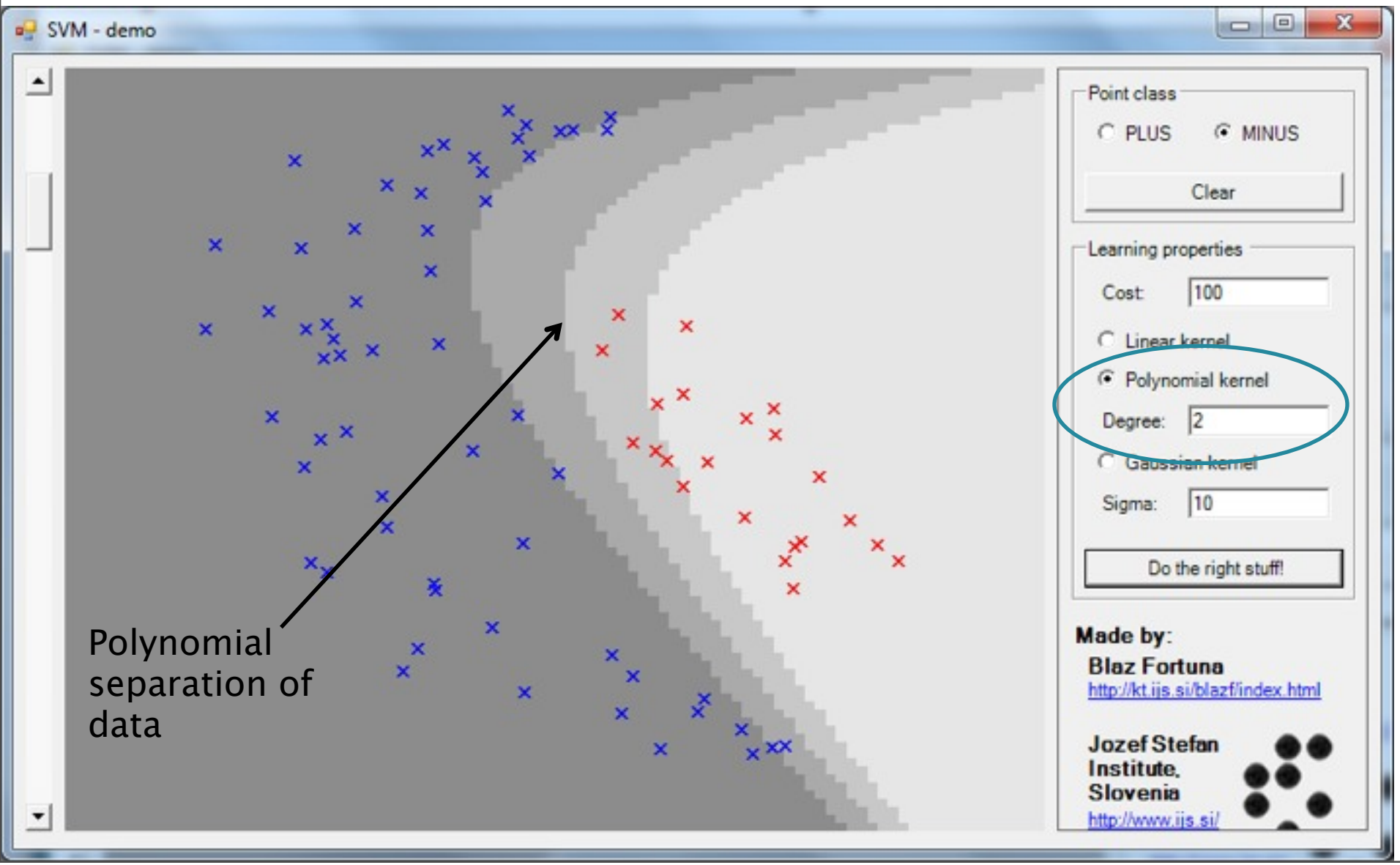
Quick example:
Why representation matters?

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

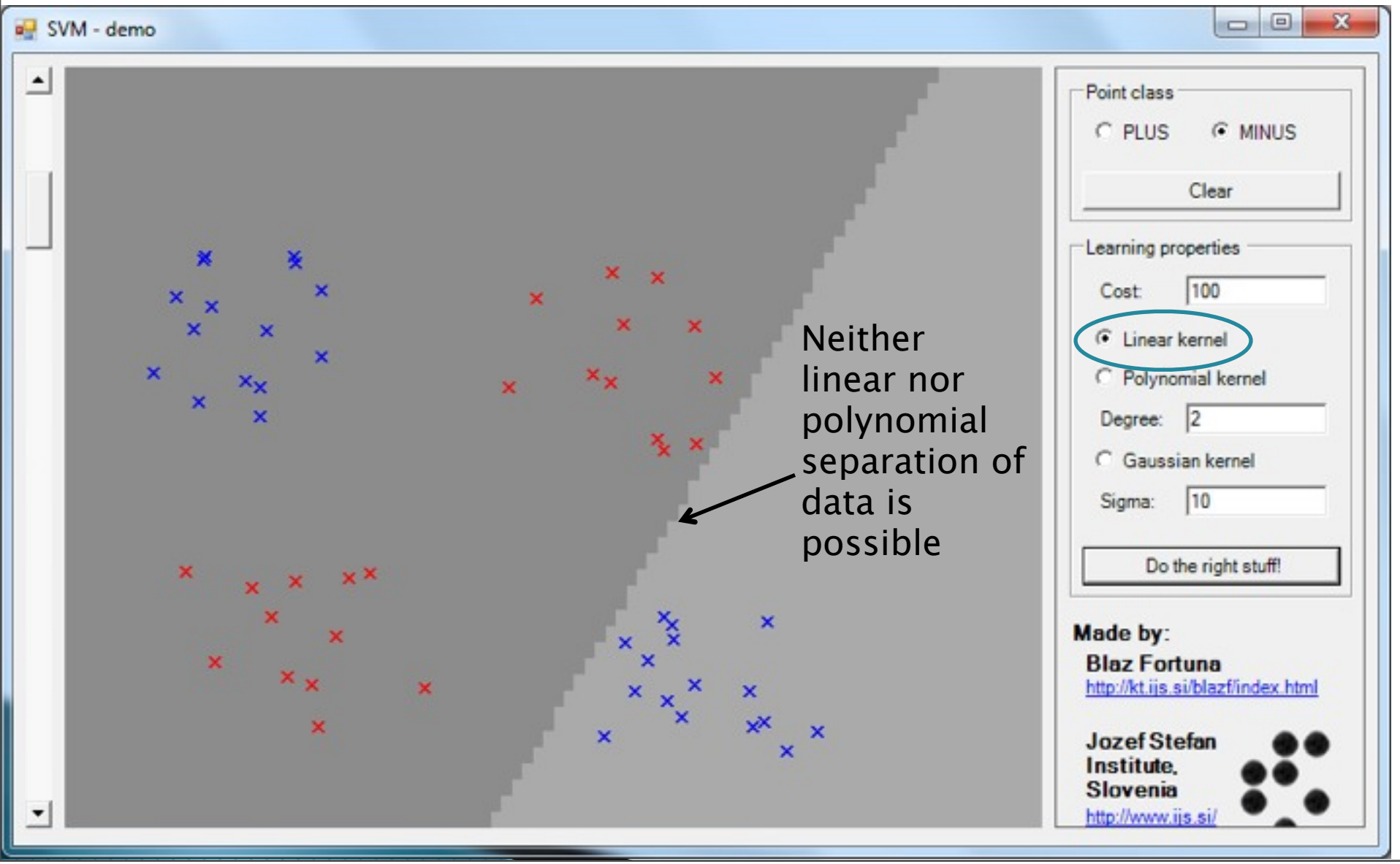
Easy decision problems require simple data representation



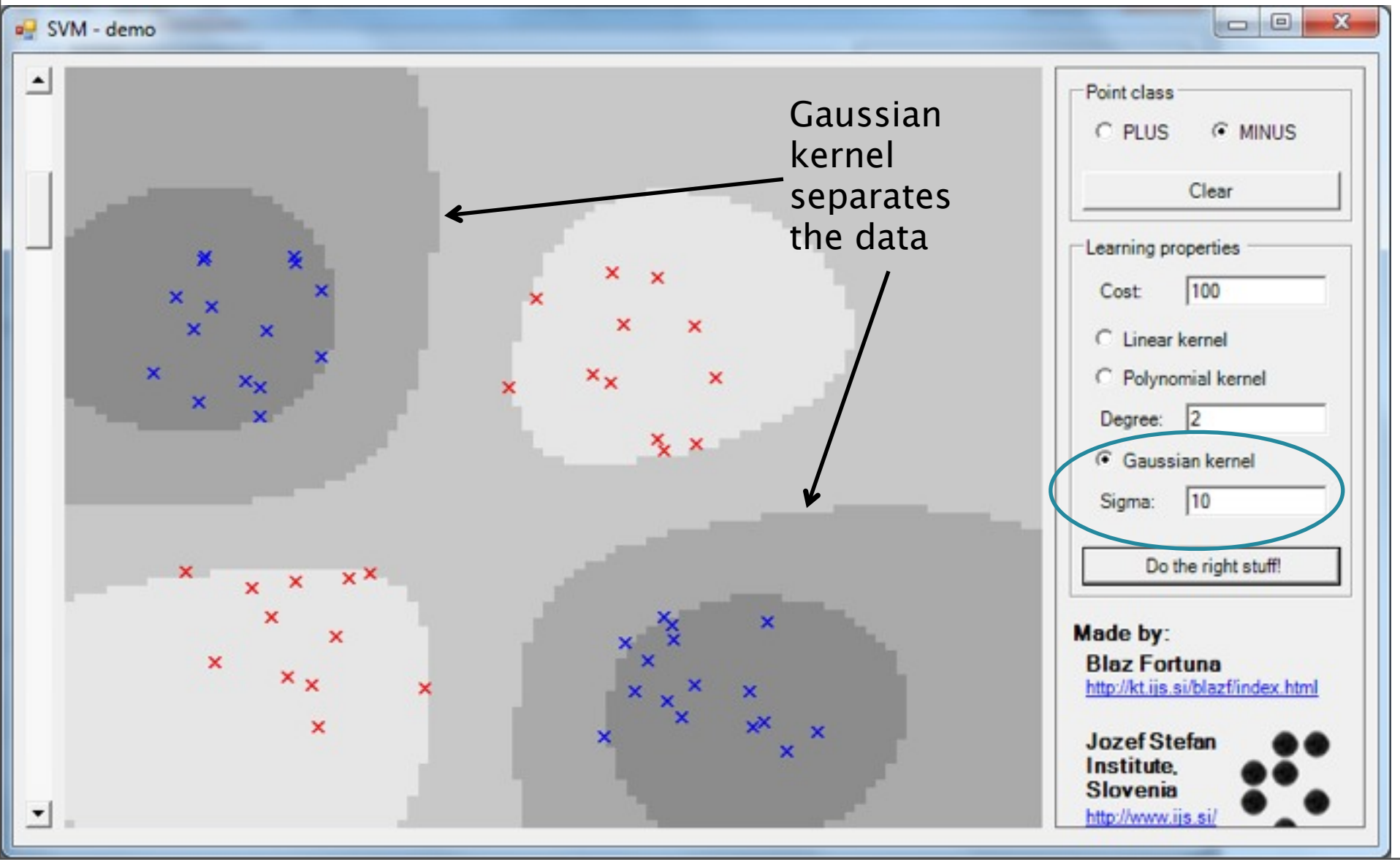
Harder decision problems require better data representation



Hard decision problems require sophisticated data representation



Hard decision problems require sophisticated data representation



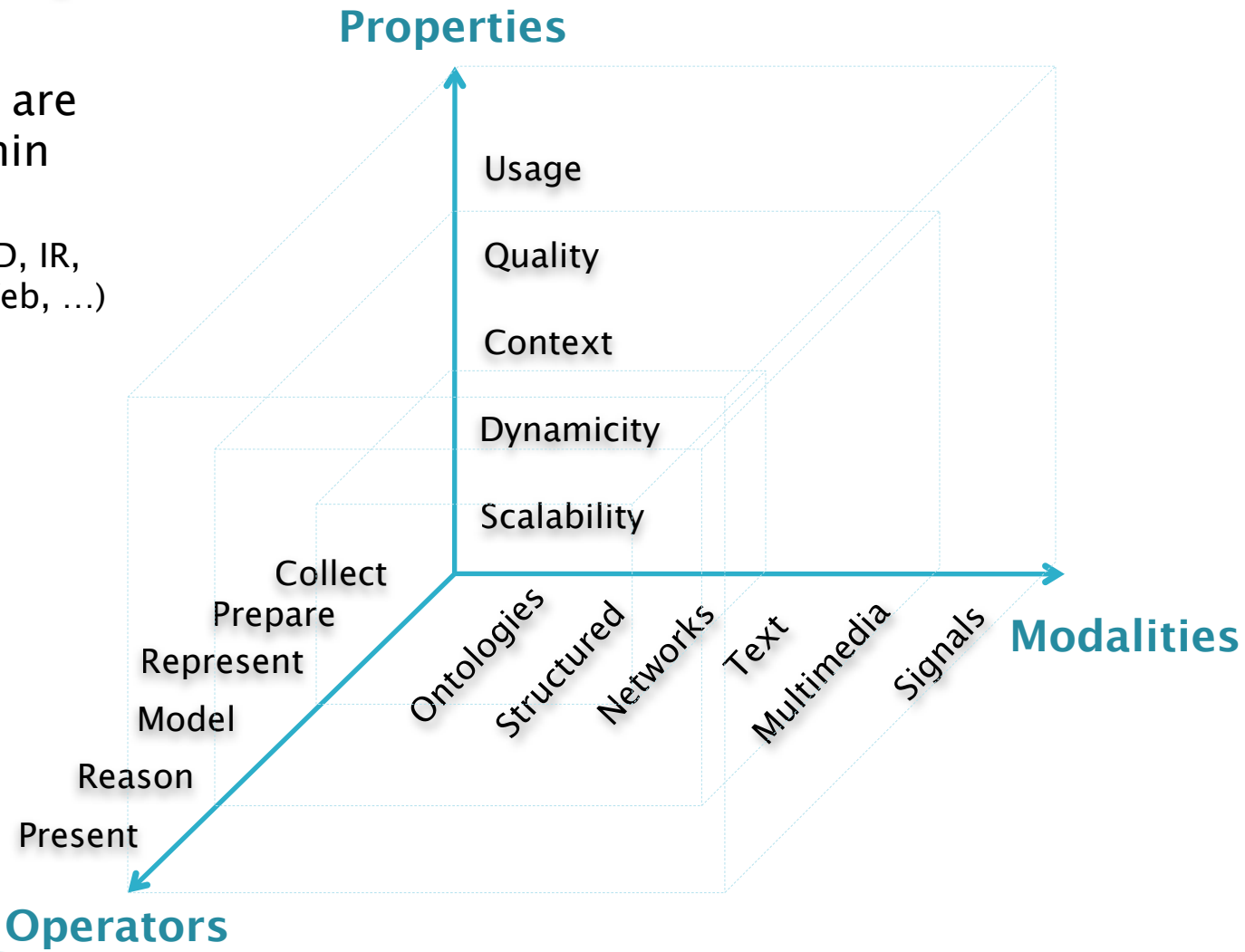
How we represent Text?

The Big picture

"Jožef Stefan" Institute

How we process data?

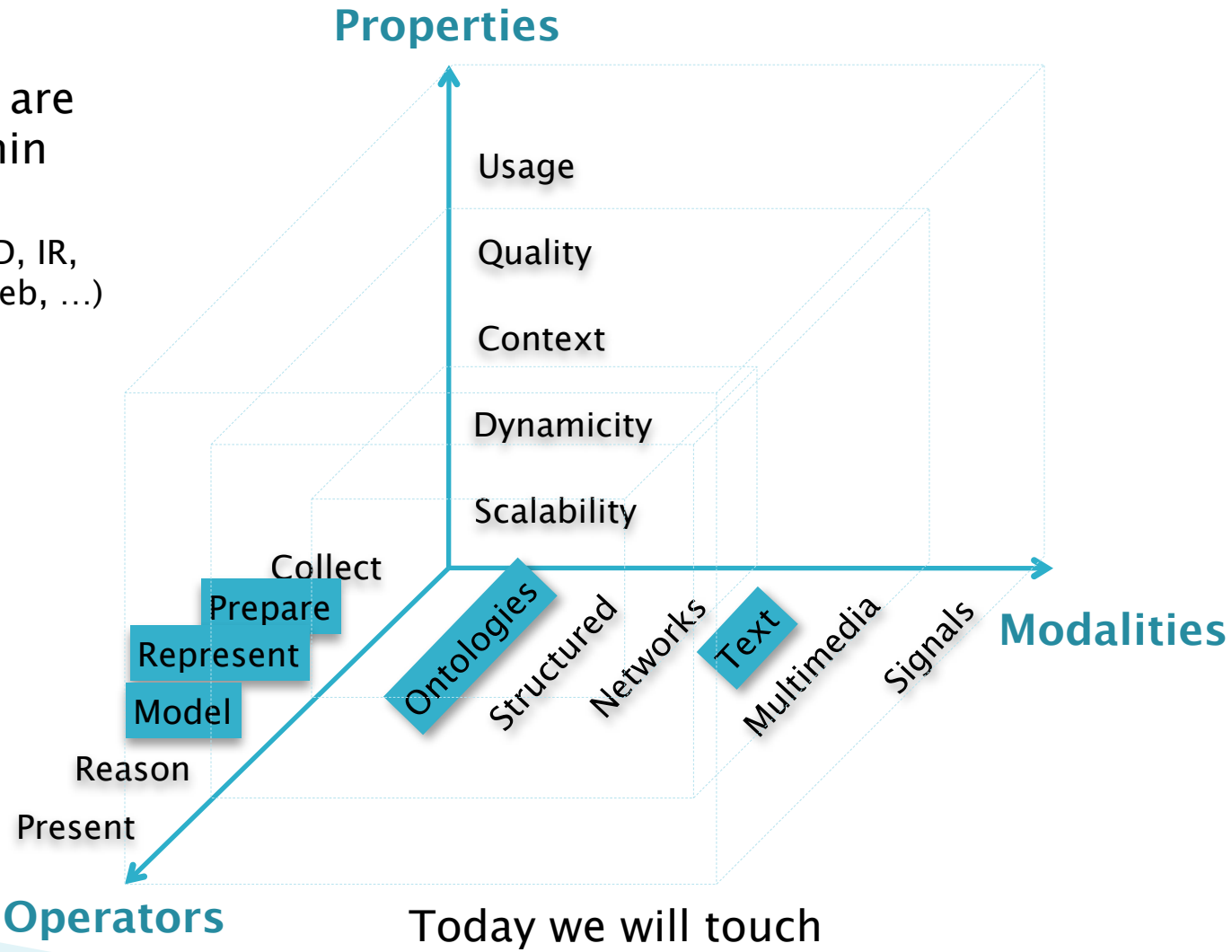
- Research areas are sub-cubes within the data cube
  - (such as ML, KDD, IR, SNA, NLP, SemWeb, …)



"Jožef Stefan" Institute

What we do with data?

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



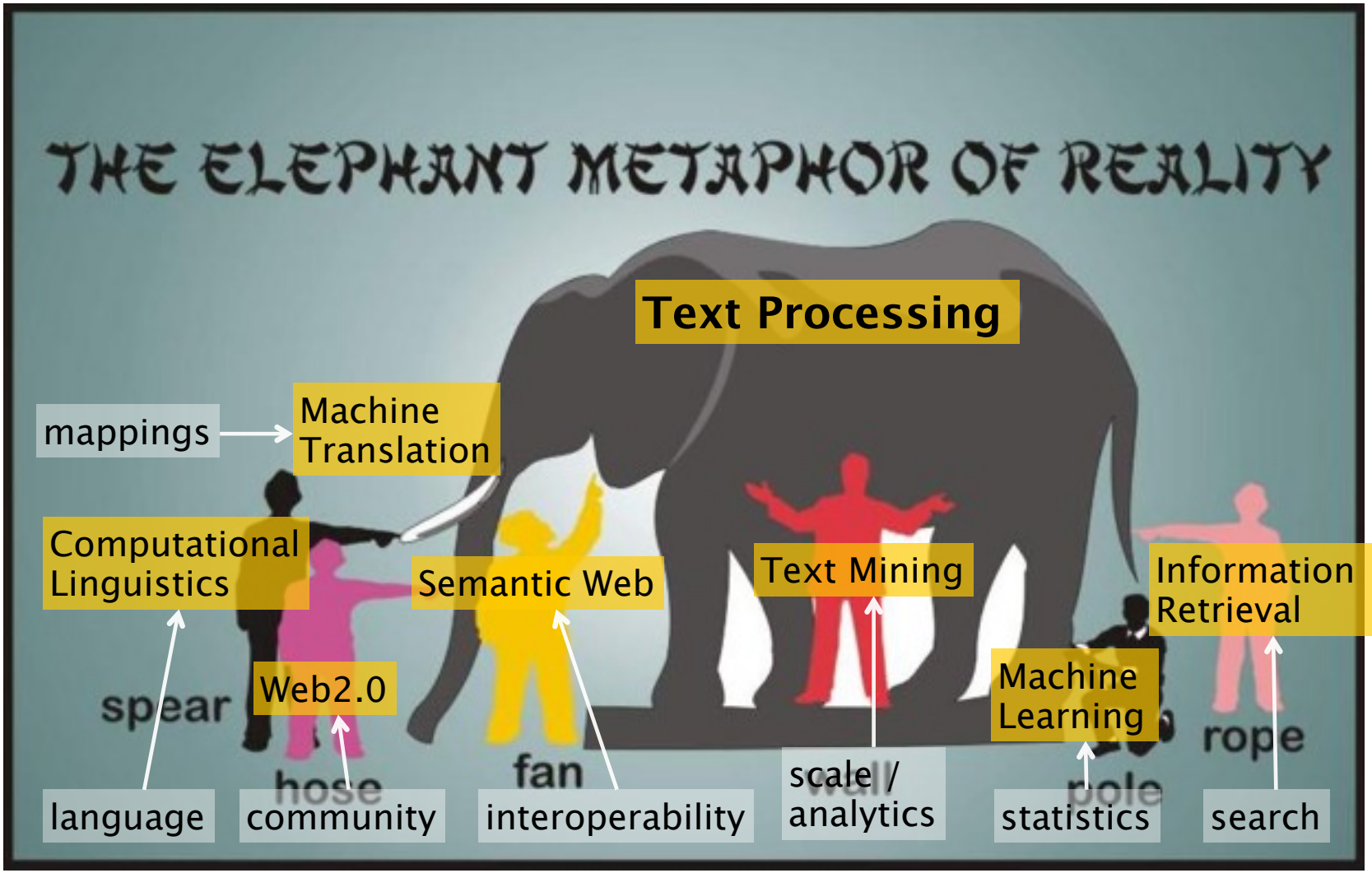
Today we will touch the highlighted parts of the cube

"Jožef Stefan" Institute

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

"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

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

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

Language identification, Copy detection

Named entity extraction (names of people, places, organizations)

Lexical

Text categorization, Clustering, Search , Summarization, …

Spam filtering, Machine translation

Syntactic

Multilingual search, Associating text with images, …

Data integration

Unifying semantics of data

Semantic

Reasoning, Semantic Search

"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

Syntactic

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

Semantic

"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

Syntactic

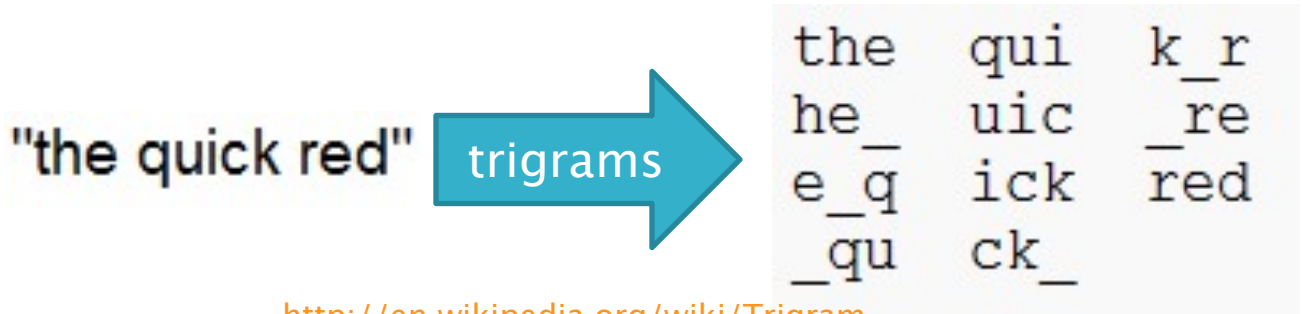
- Collaborative tagging / Web2.0
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Semantic

"Jožef Stefan" Institute

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



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

"Jožef Stefan" Institute

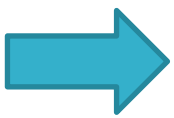
Good and bad sides of character level representation (n-grams)

- 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

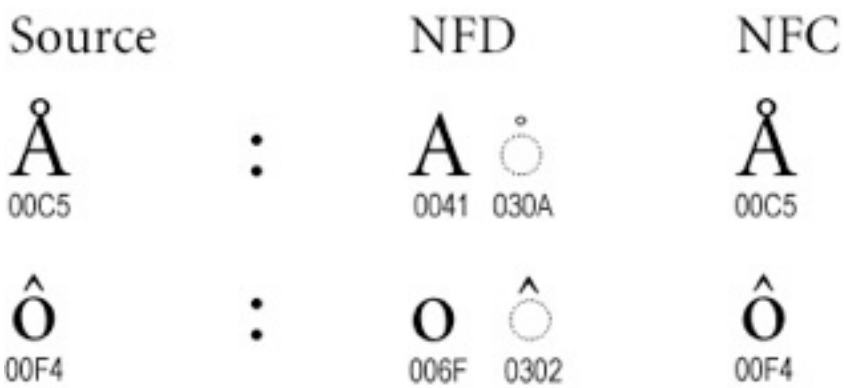


th 1.52%	en 0.55%	ng 0.18%
he 1.28%	ed 0.53%	of 0.16%
in 0.94%	to 0.52%	al 0.09%
er 0.94%	it 0.50%	de 0.09%
an 0.82%	ou 0.50%	se 0.08%
re 0.68%	ea 0.47%	le 0.08%
nd 0.63%	hi 0.46%	sa 0.06%
at 0.59%	is 0.46%	si 0.05%
on 0.57%	or 0.43%	ar 0.04%
nt 0.56%	ti 0.34%	ve 0.04%
ha 0.56%	as 0.33%	ra 0.04%
es 0.56%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



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

"Jožef Stefan" Institute

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:
  - **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 an 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

"Jožef Stefan" Institute

Stemming

- For English is mostly used Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- Example cascade rules used in English Porter stemmer
  - ATIONAL -> ATE relational -> relate
  - TIONAL -> TION conditional -> condition
  - ENCI -> ENCE valenci -> valence
  - ANCI -> ANCE hesitanci -> hesitance
  - IZER -> IZE digitizer -> digitize
  - ABLI -> ABLE conformabli -> conformable
  - ALLI -> AL radicalli -> radical
  - ENTLI -> ENT differentli -> different
  - ELI -> E vileli -> vile
  - OUSLI -> OUS analogousli -> analogous

Levels of text representations

- Character (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

"Jožef Stefan" Institute

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 indispensible 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

"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

Syntactic

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

Semantic

"Jožef Stefan" Institute

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
Verb	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com **is** a web site. I **like** EnglishClub.com.
Noun	thing or person	pen, dog, work, music, town, London, teacher, John	This is my **dog**. He lives in my **house**. We live in **London**.
Adjective	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is **big**. I like **big** dogs.
Adverb	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.
Pronoun	replaces a noun	I, you, he, she, some	Tara is Indian. **She** is beautiful.
Preposition	links a noun to another word	to, at, after, on, but	We went **to** school **on** Monday.
Conjunction	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.
Interjection	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	**Ouch**! That hurts! **Hi**! How are you? **Well**, I don't know.

"Jožef Stefan" Institute

http://www.englishclub.com/grammar/parts-of-speech_1.htm

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.

http://www.englishclub.com/grammar/parts-of-speech_2.htm

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

"Jožef Stefan" Institute

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
Hypernym	From lower to higher concepts	breakfast → meal
Hyponym	From concepts to subordinates	meal → lunch
Has-Member	From groups to their members	faculty → professor
Member-Of	From members to their groups	copilot → crew
Has-Part	From wholes to parts	table → leg
Part-Of	From parts to wholes	course → meal
Antonym	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

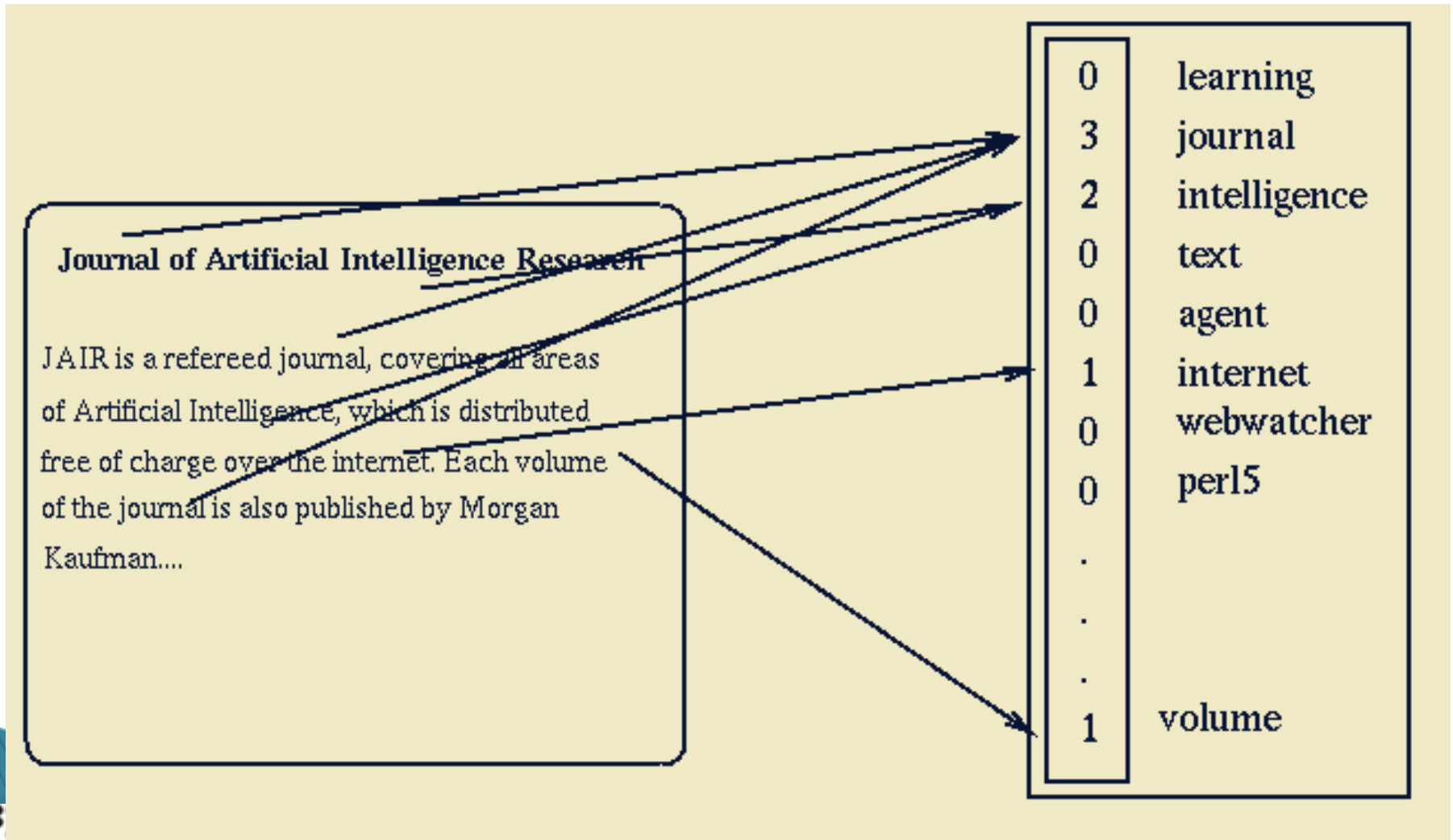
"Jožef Stefan" Institute

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

"Jožef Stefan" Institute

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.

Original text

- [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]

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

"Jožef Stefan" Institute

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 = <x>
- 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)

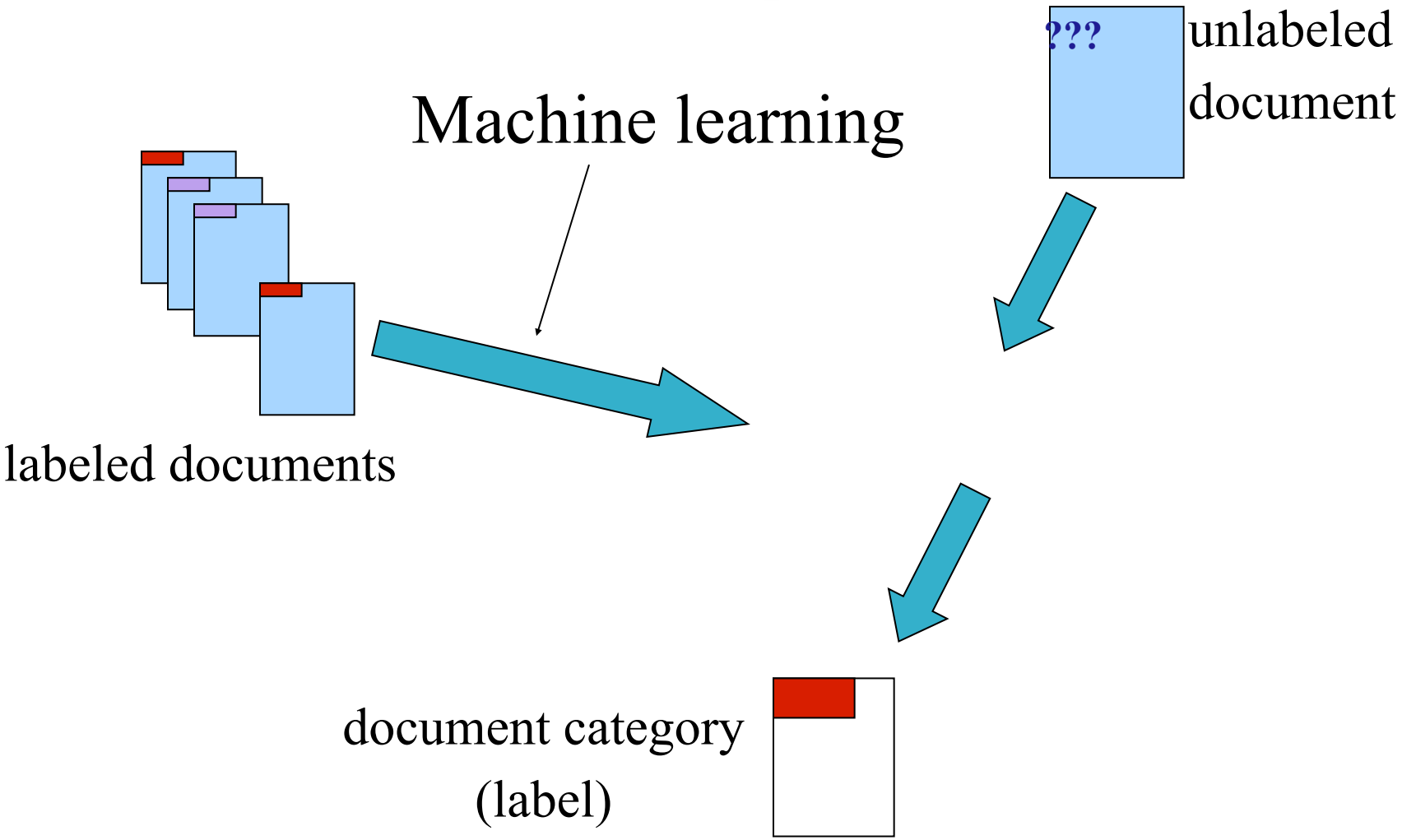
$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2} \sqrt{\sum_k x_k^2}}$$

"Jožef Stefan" Institute

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 $\mathbf{w_i}$ (one weight per word from the vocabulary)

Algorithm:

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

Measuring success – Model quality estimation

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

The truth, and

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

..the whole truth

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

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

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

"Jožef Stefan" Institute

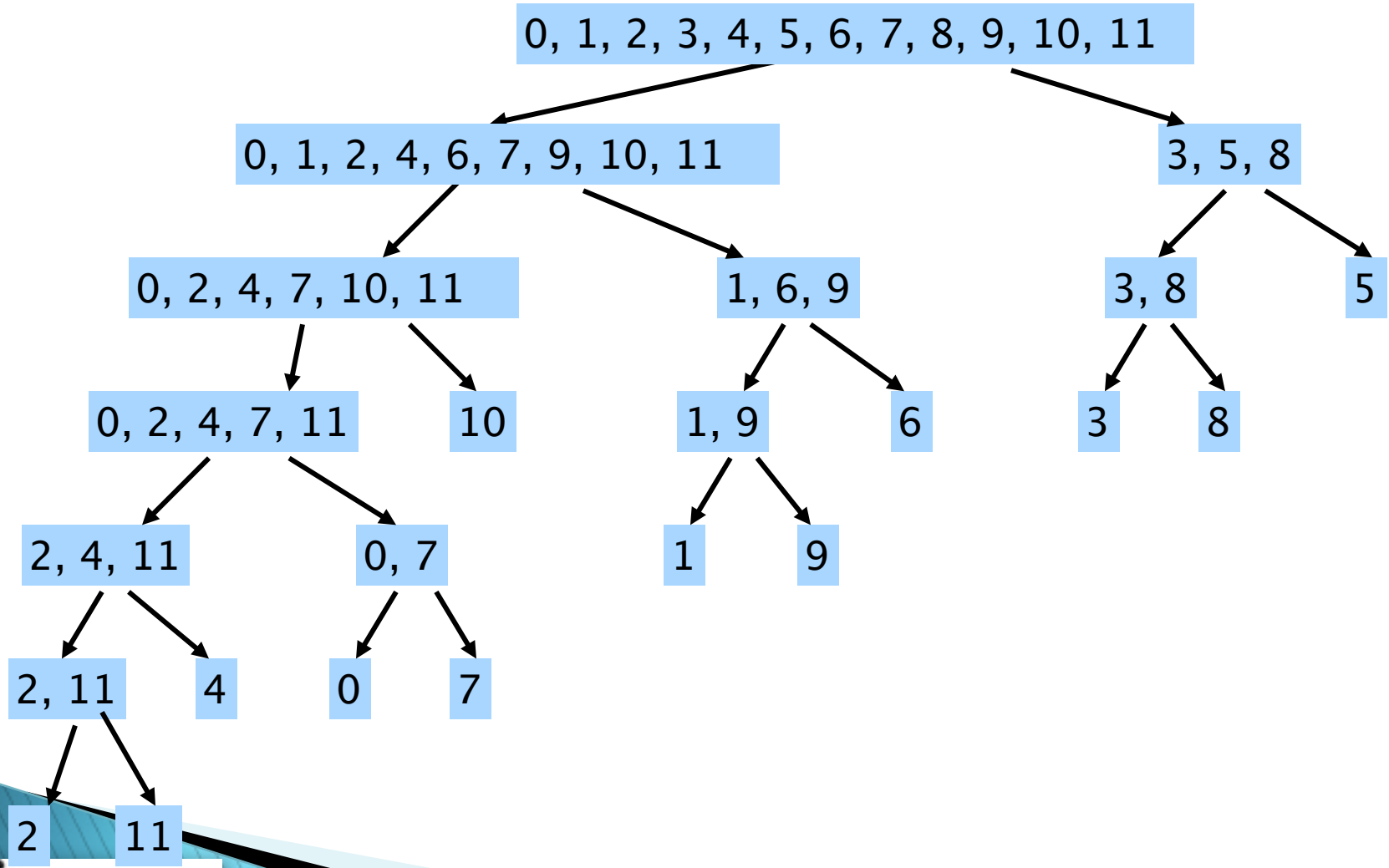
Document Clustering Task

- 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)



"Jožef Stefan" Institute

Latent Semantic Indexing

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

LSI Example

Original document-term mantrix

	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

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 →

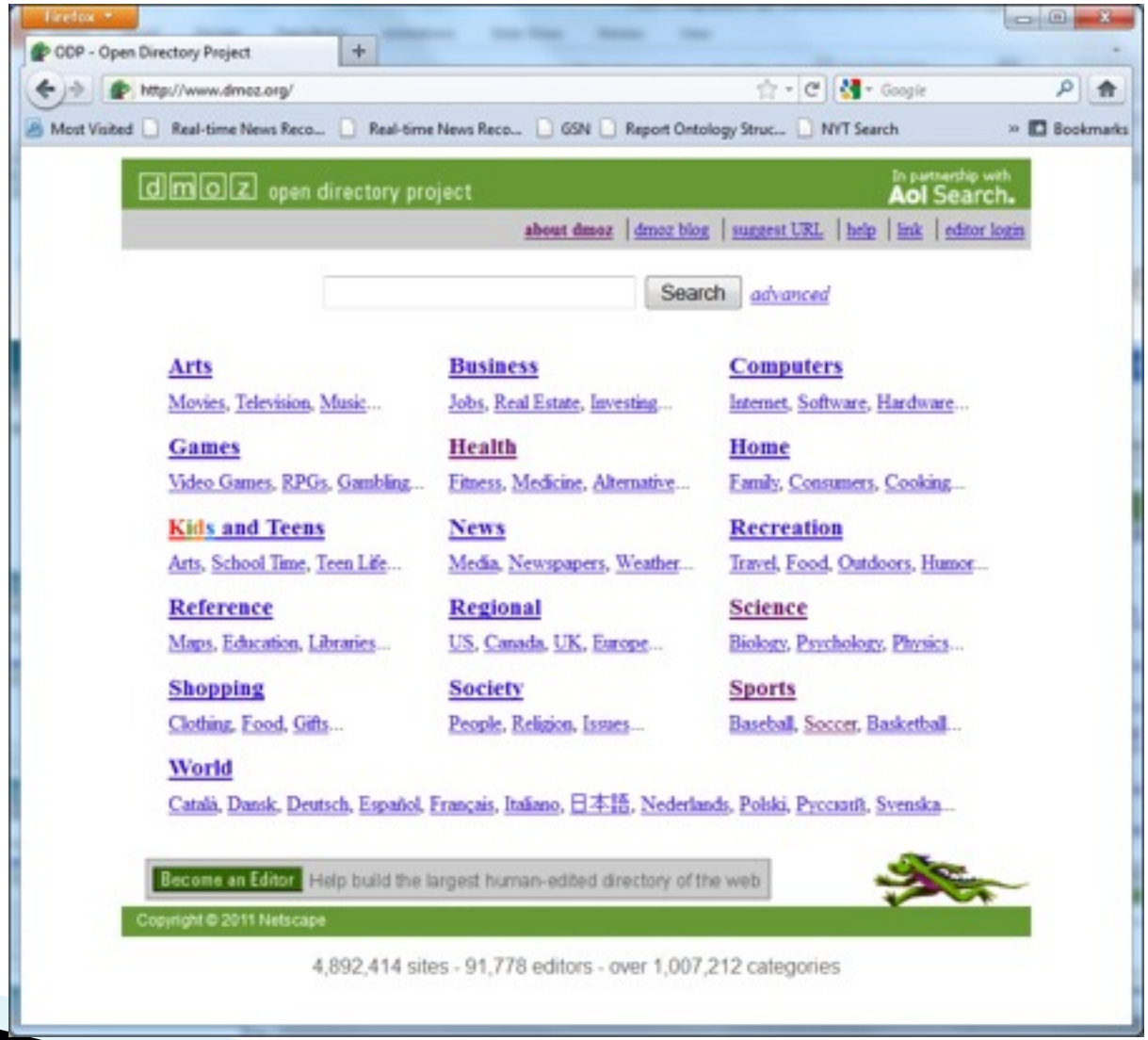
	d1	d2	d3	d4	d5	d6
d1	1.00					
d2	0.8	1.00				
d3	0.4	0.9	1.00			
d4	0.5	-0.2	-0.6	1.00		
d5	0.7	0.2	-0.3	0.9	1.00	
d6	0.1	-0.5	-0.9	0.9	0.7	1.00

Document Classification into large taxonomies

"Jožef Stefan" Institute

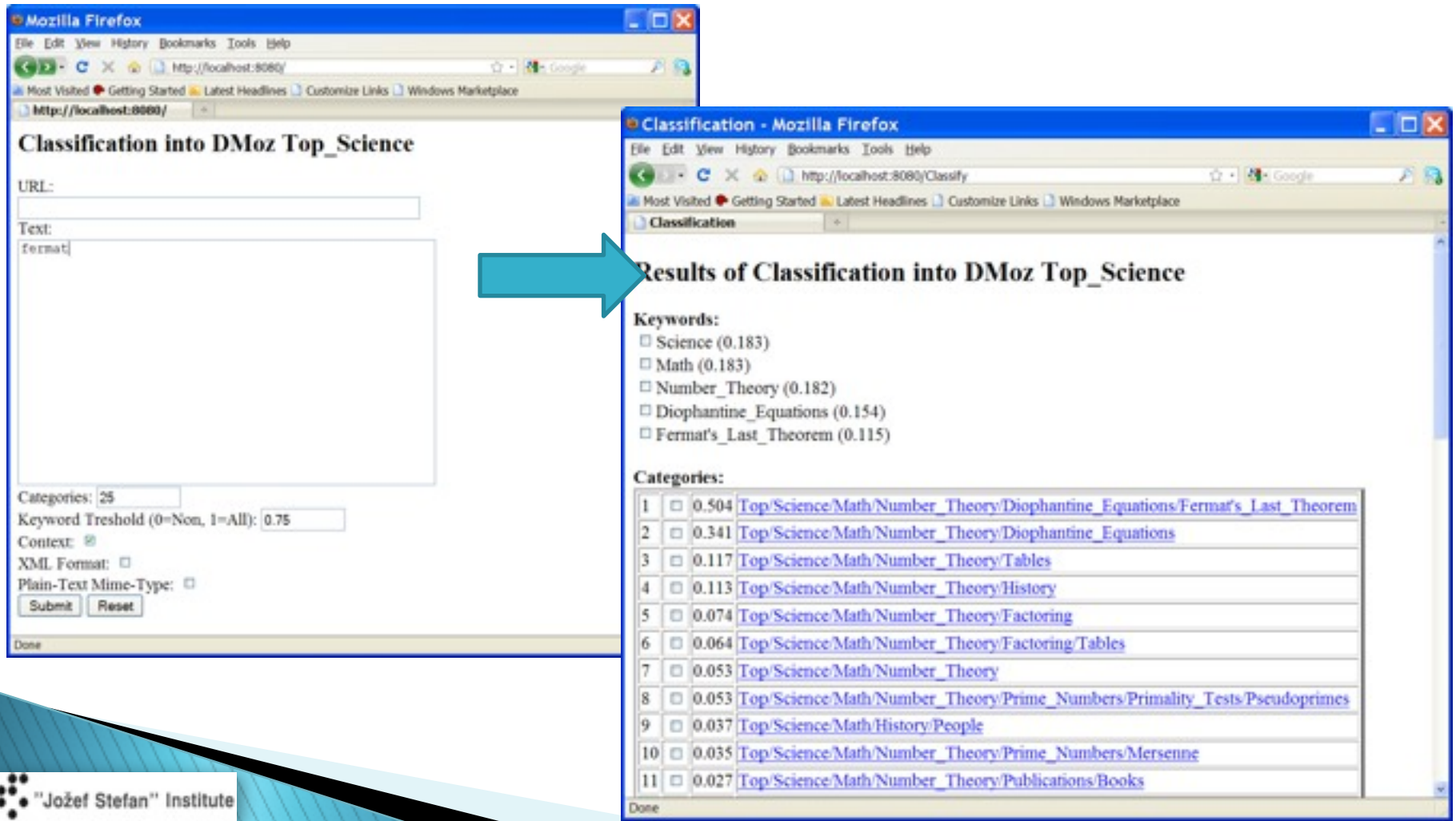
DMoz (Open Directory Project) http://dmoz.org

- 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



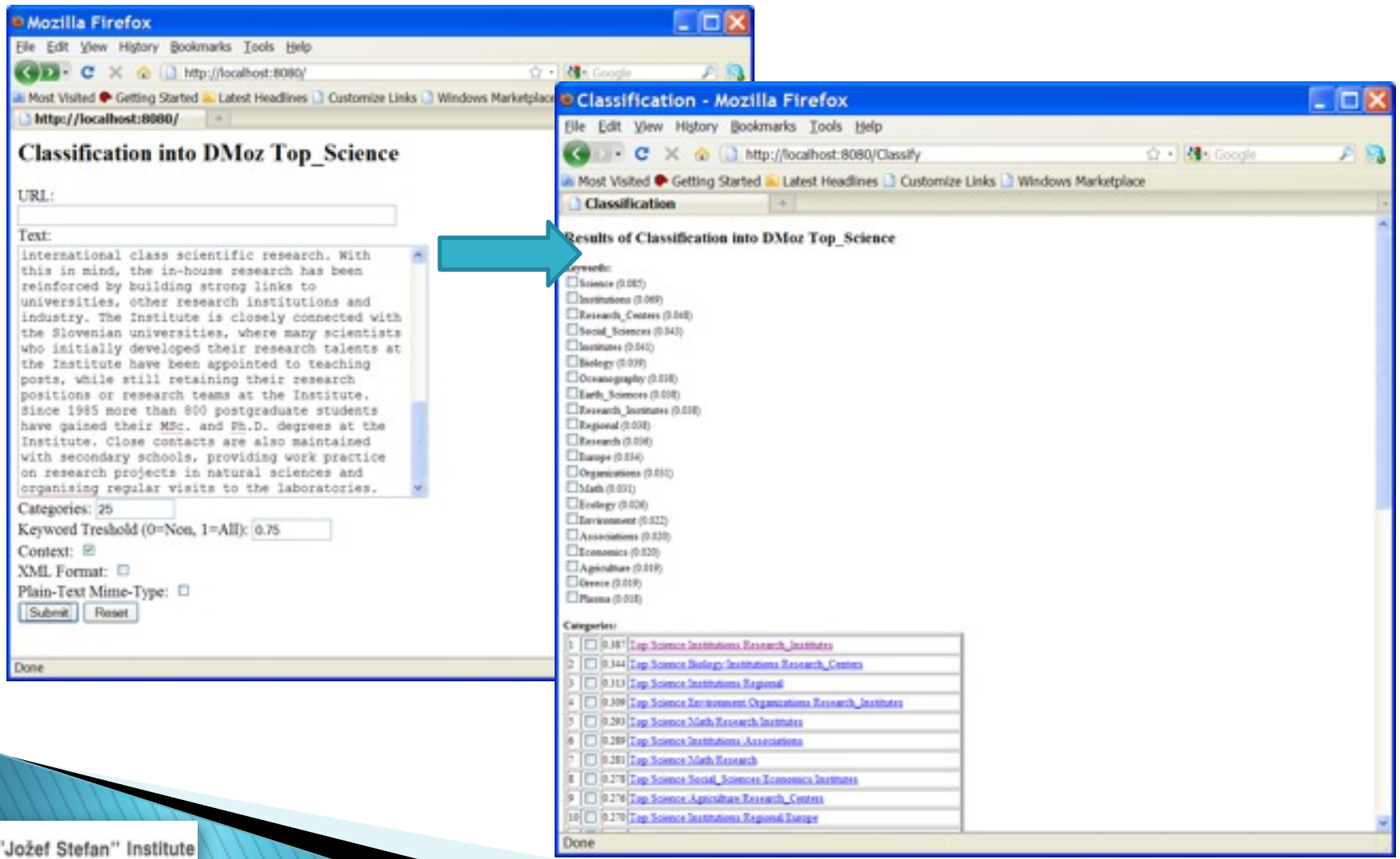
"Jožef Stefan" Institute

Classification of a query into DMoz



"Jožef Stefan" Institute

Classification of a document into DMoz



"Jožef Stefan" Institute

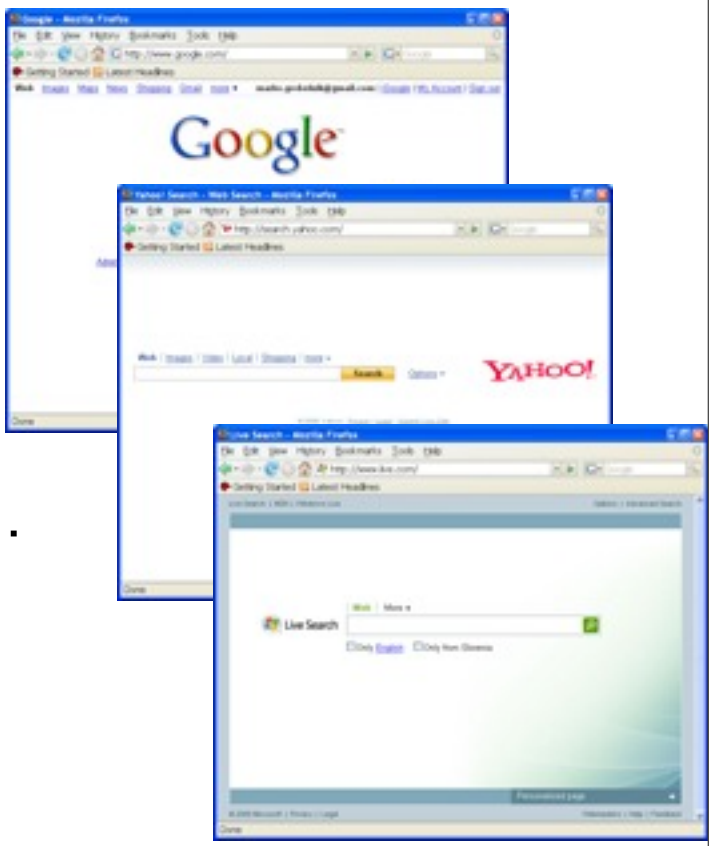
Visual & Contextual Search

(WWW2008)

"Jožef Stefan" Institute

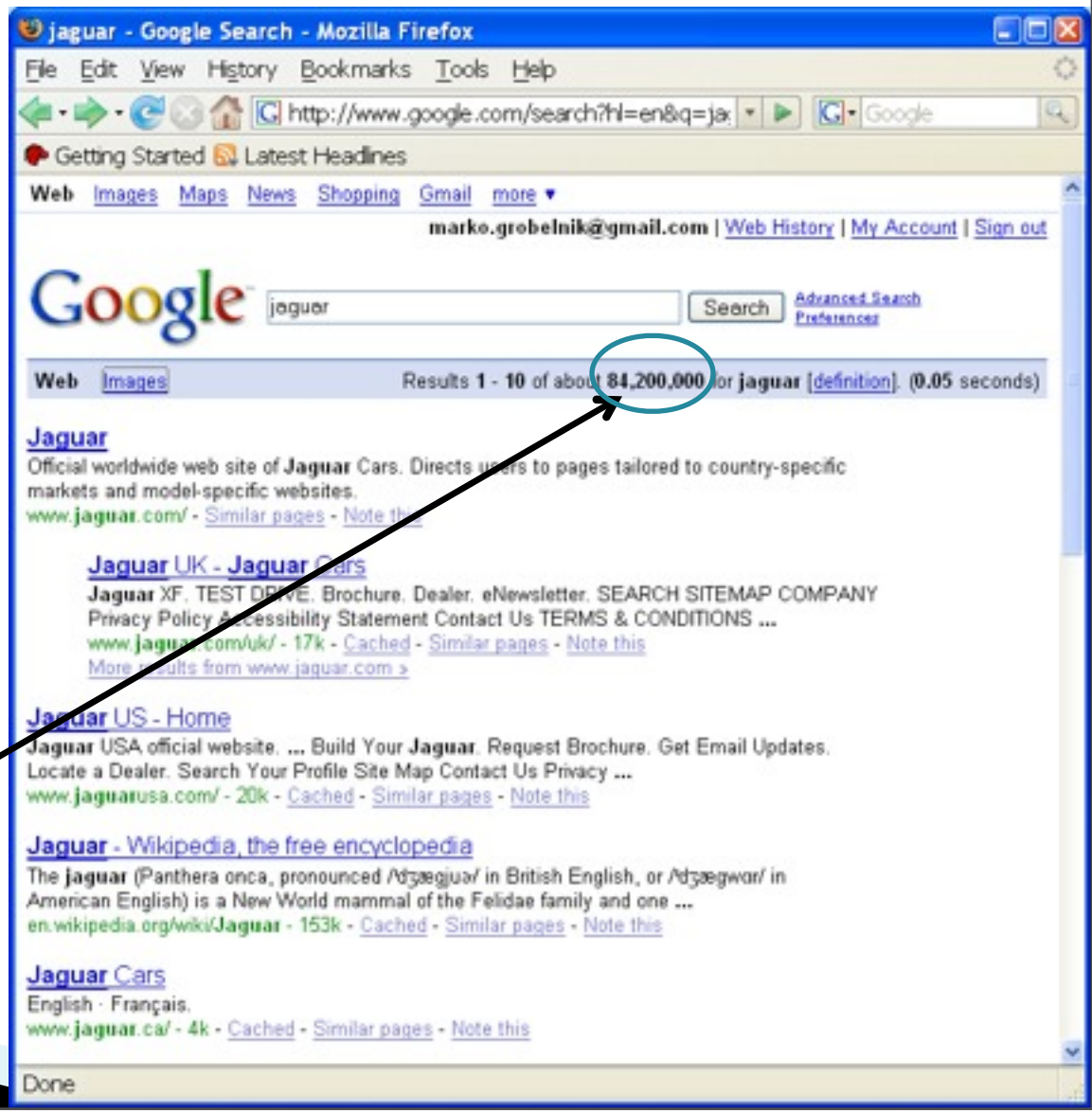
Contextualized search

- 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…

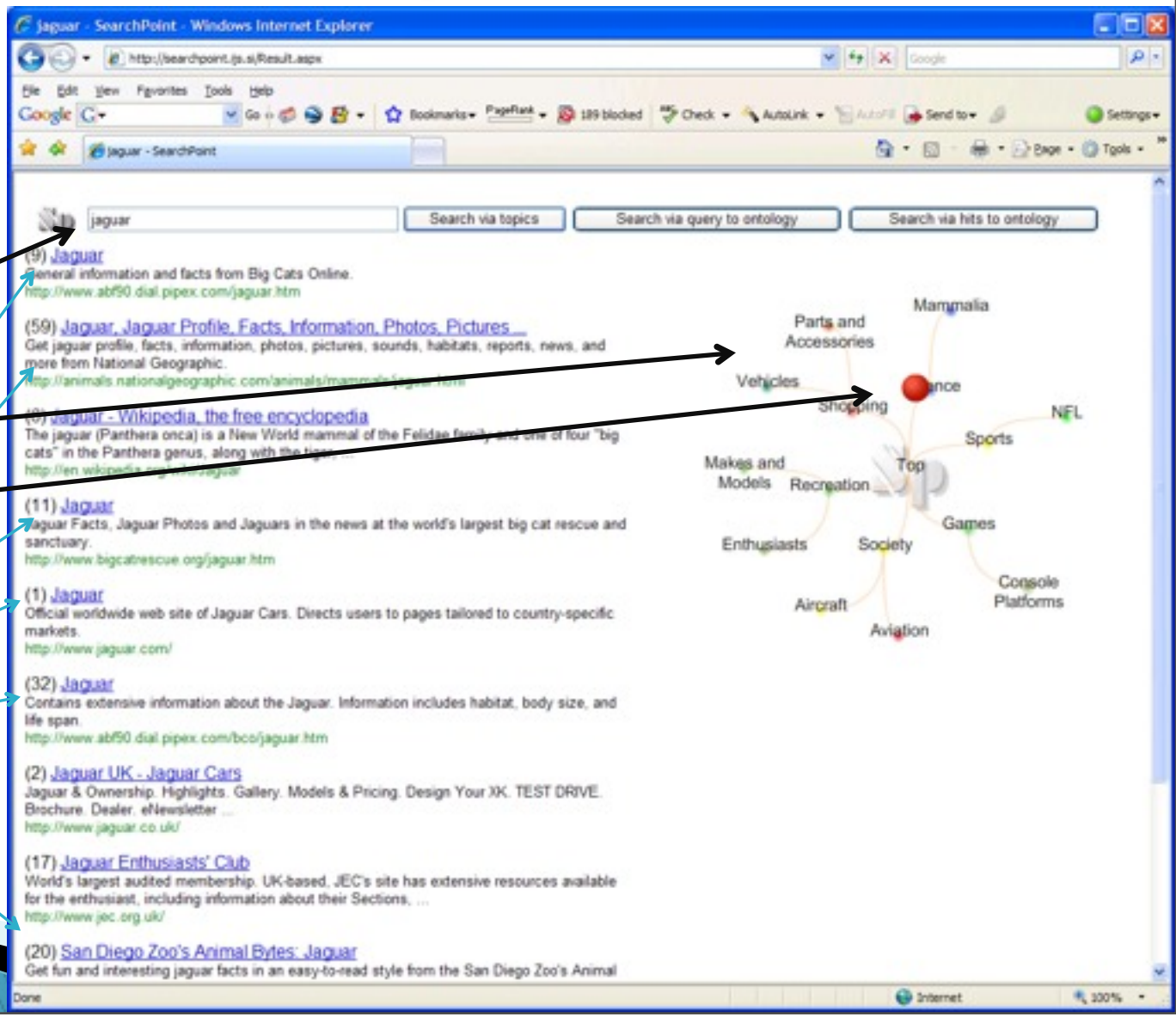


Example: searching for

- Query "**jaguar**" has many meanings…
- …but the first page of search engines doesn't provide us with many answers
- …there are 84M more results



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



Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point

"Jožef Stefan" Institute

News reporting bias

(Fortuna, Galleguillos, Cristianini 2008)

"Jožef Stefan" Institute

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	–

"Jožef Stefan" Institute

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

"Jožef Stefan" Institute

Detecting News Reporting

- The task:
  - Given a news story, are we able to say from which news source it came?

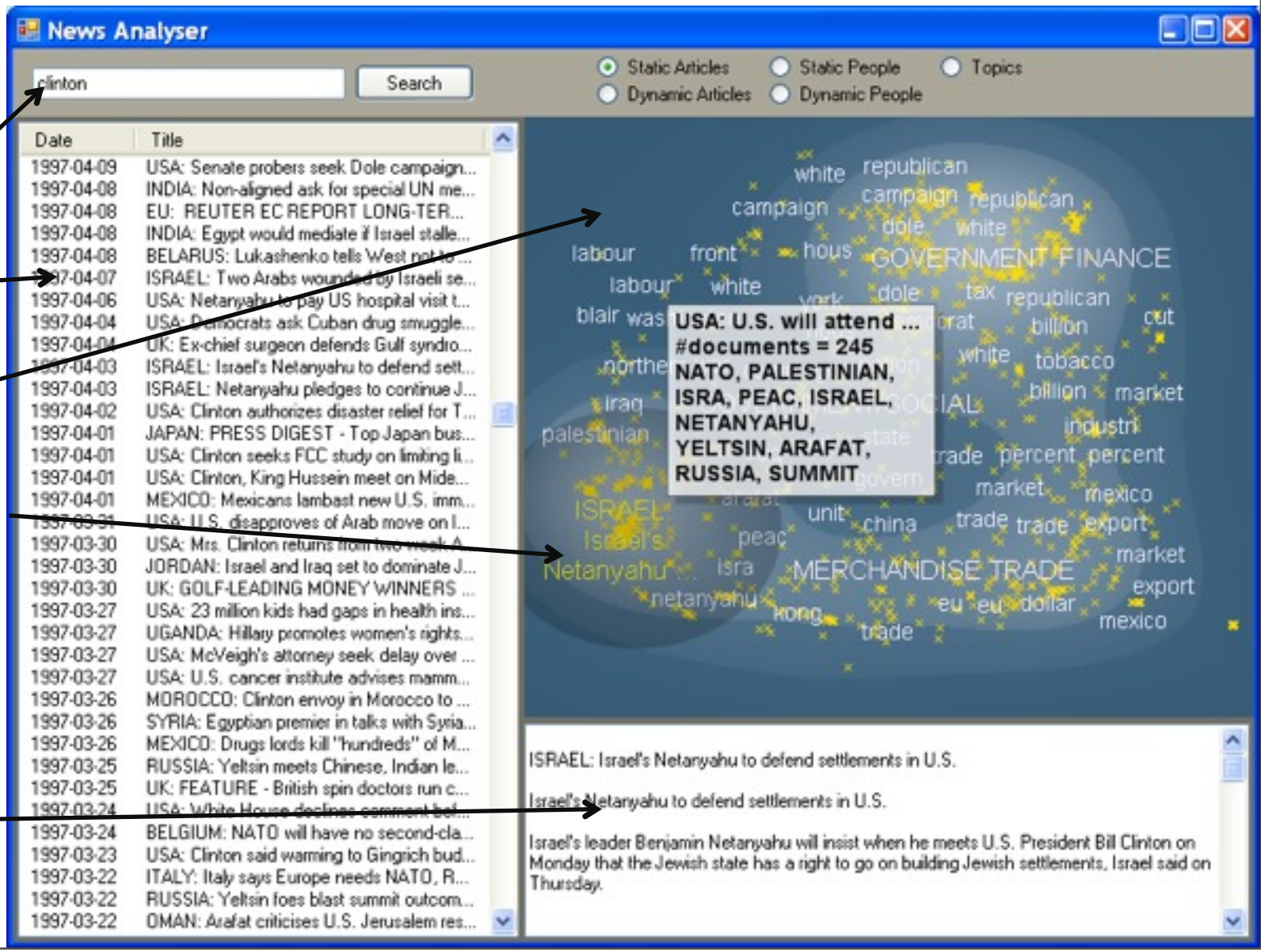
- We compared **CNN** and **Aljazeera** reports about the same events from the war in Iraq
  - …300 aligned articles describing the same story from both sources
- The same topics are expressed in both sources with the following keywords:
  - CNN with:
    - **Insurgents**, Troops, Baghdad, Iran, **Militant**, 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, …

"Jožef Stefan" Institute

News Visualization

"Jožef Stefan" Institute

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



Query

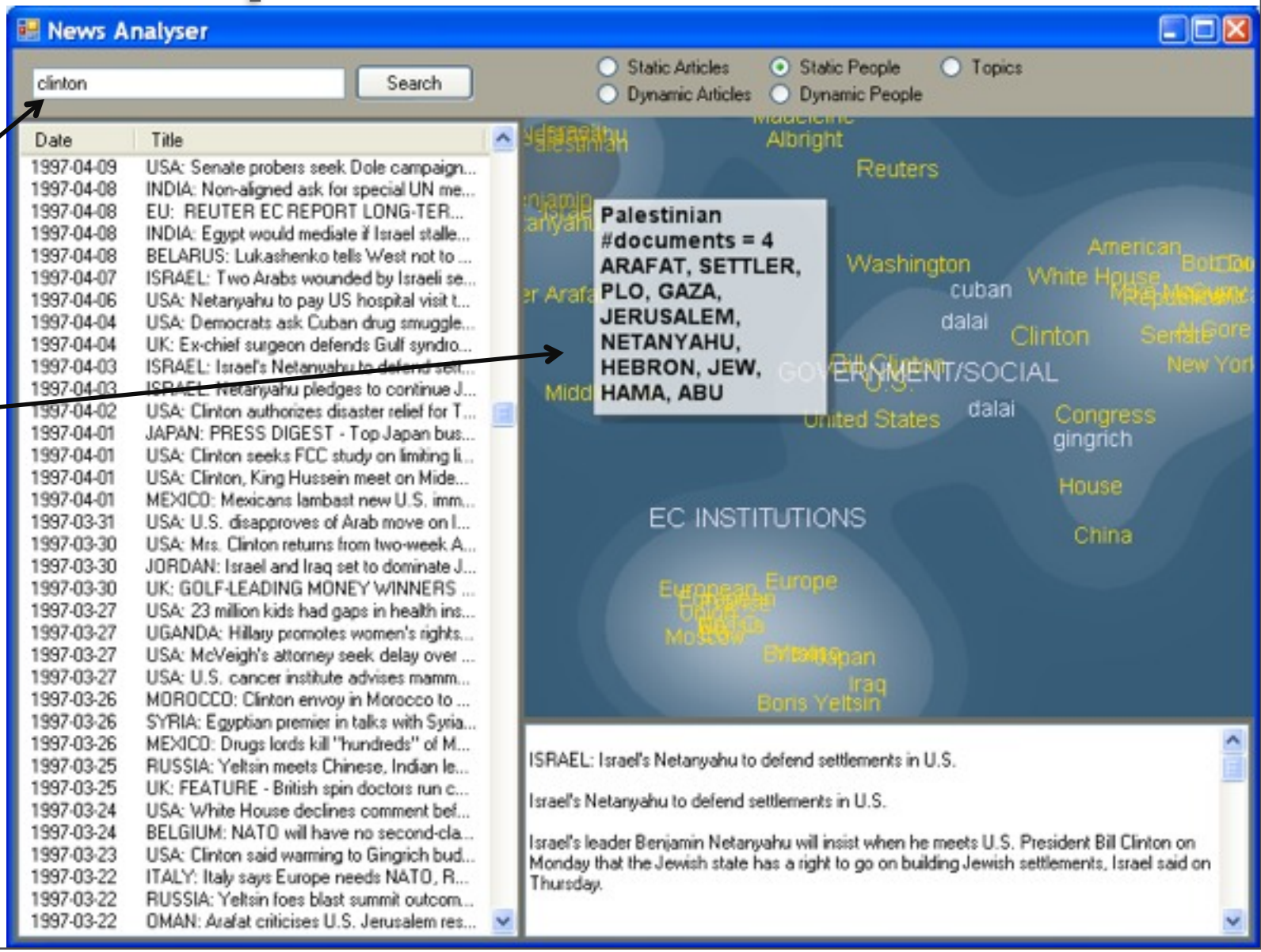
Search Results

Topic Map

Selected group of news

Selected story

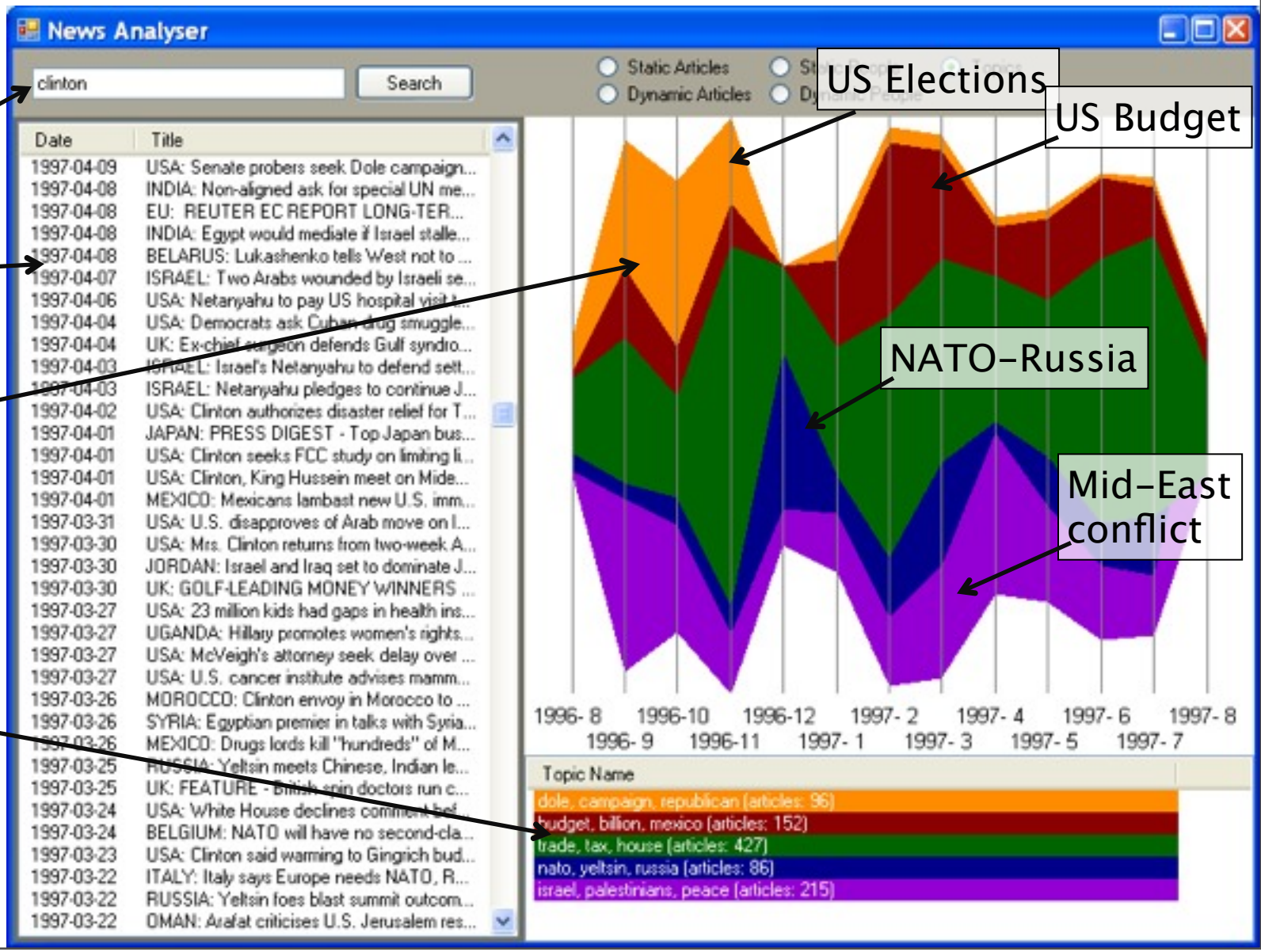
Visualization of social relationships between “Clinton”



Query

Named entities in relation

Topic Trends Tracking of the documents including “Clinton”



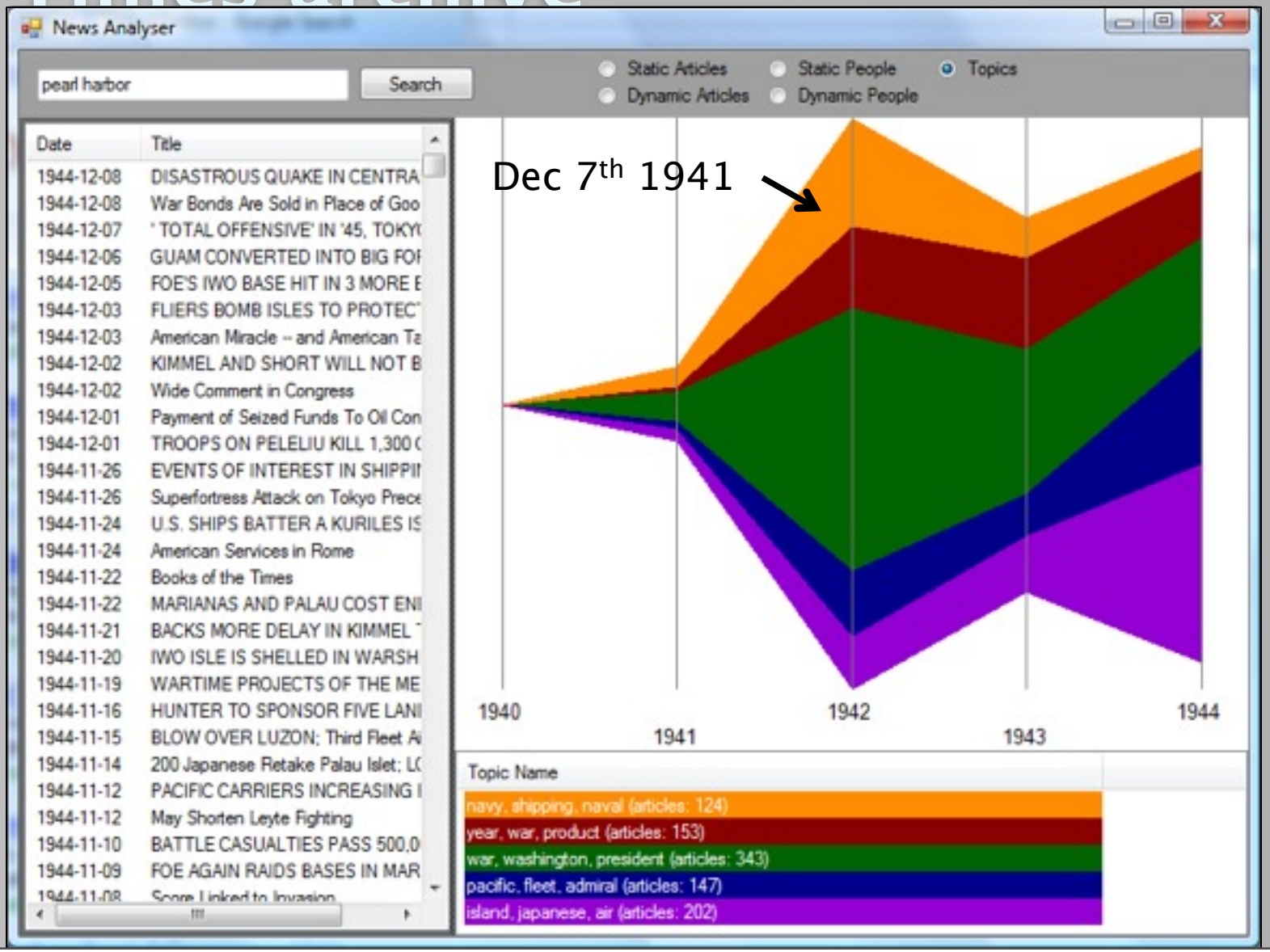
Query

Result set

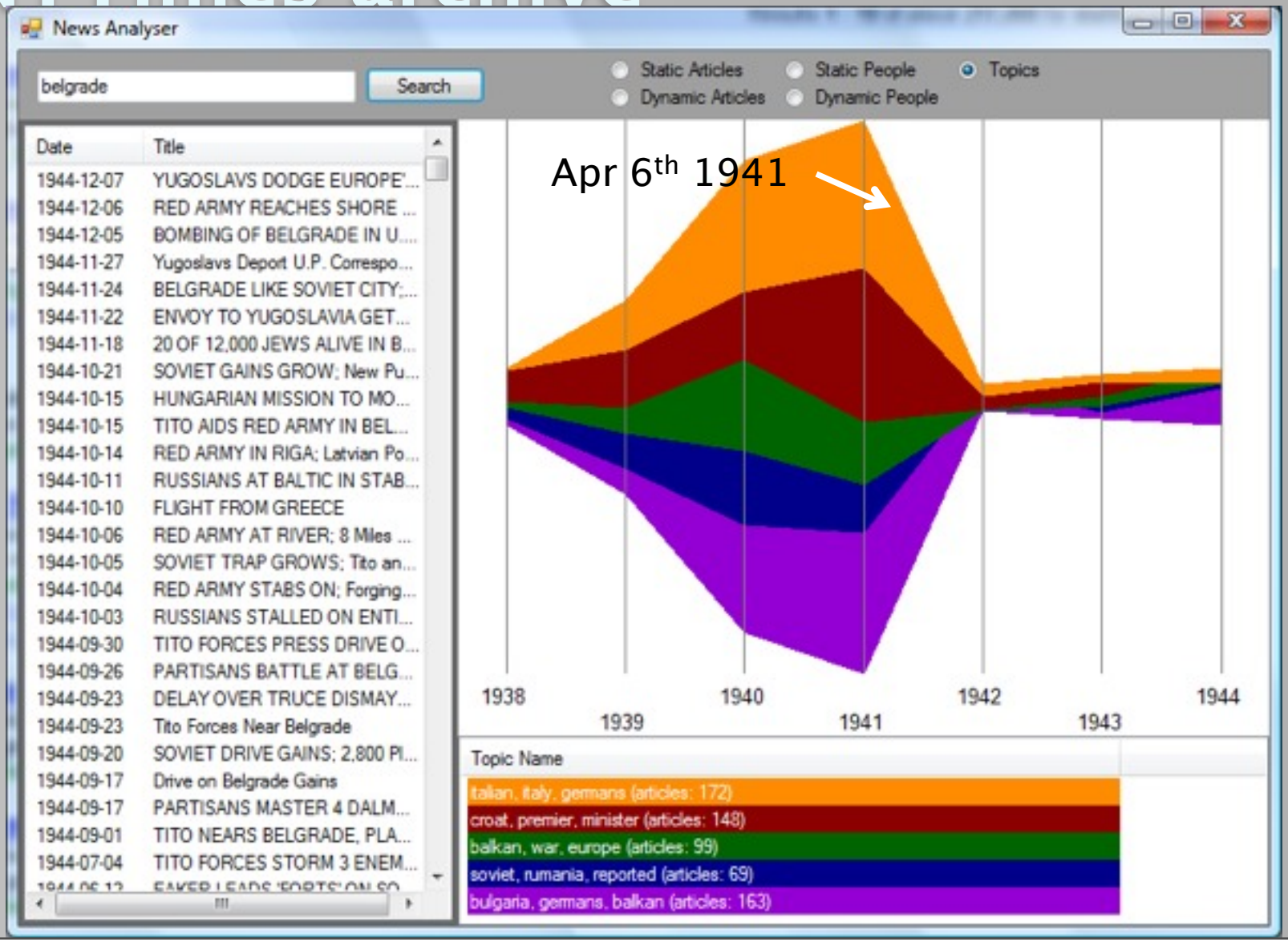
Topic Trends Visualization

Topics description

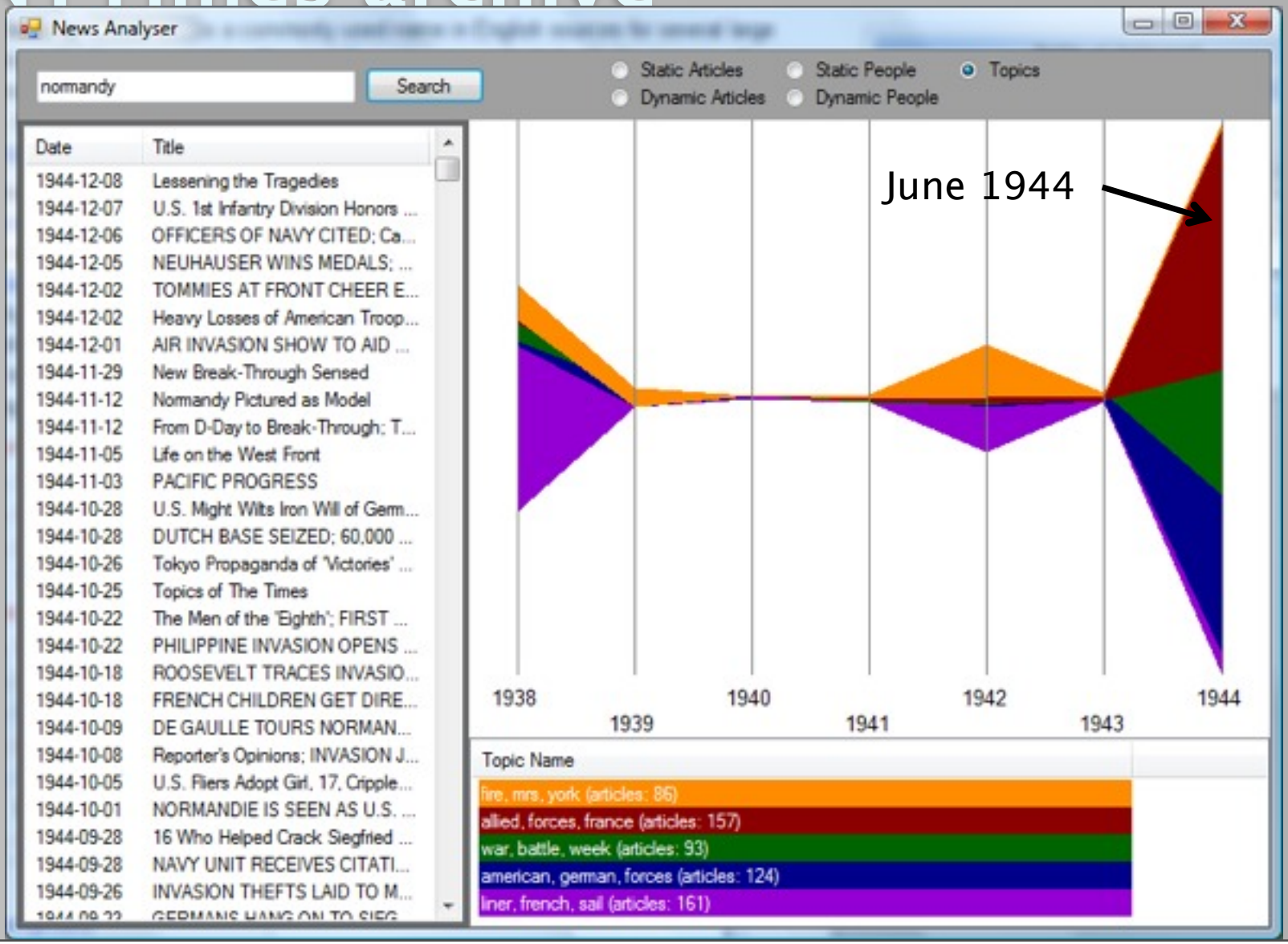
WW2 query "Pearl Harbor" into NYTimes archive



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

Syntactic

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

Semantic

"Jožef Stefan" Institute

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)

"Jožef Stefan" Institute

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

Lexical

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

Syntactic

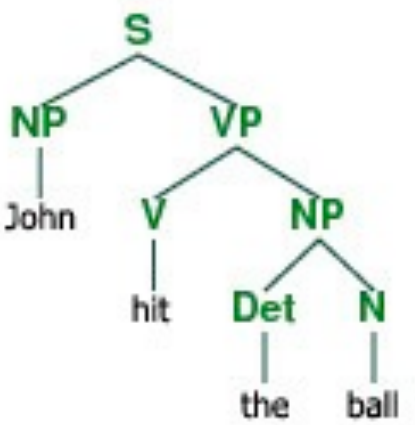
- Collaborative tagging / Web2.0
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Semantic

"Jožef Stefan" Institute

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

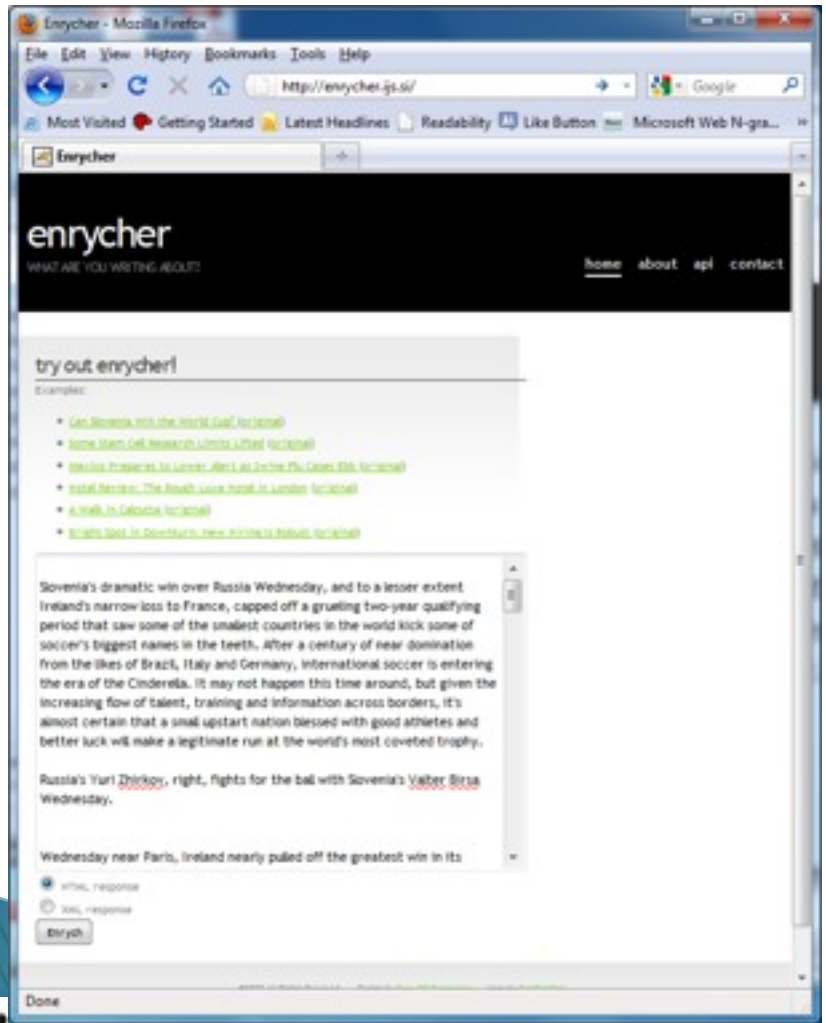


Text Enrichment

WWW2009-SemSearch

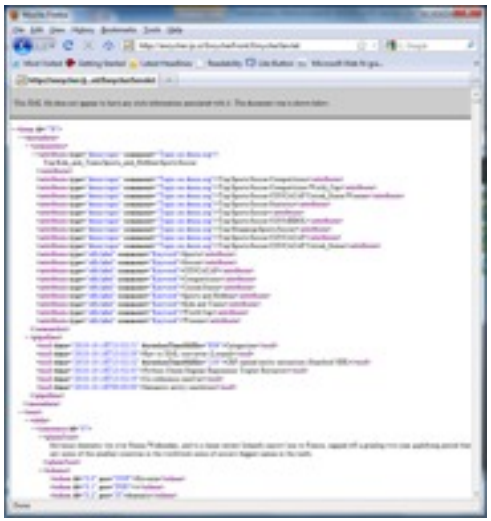
"Jožef Stefan" Institute

Text enrichment with http://
Enrycher.ijs.si

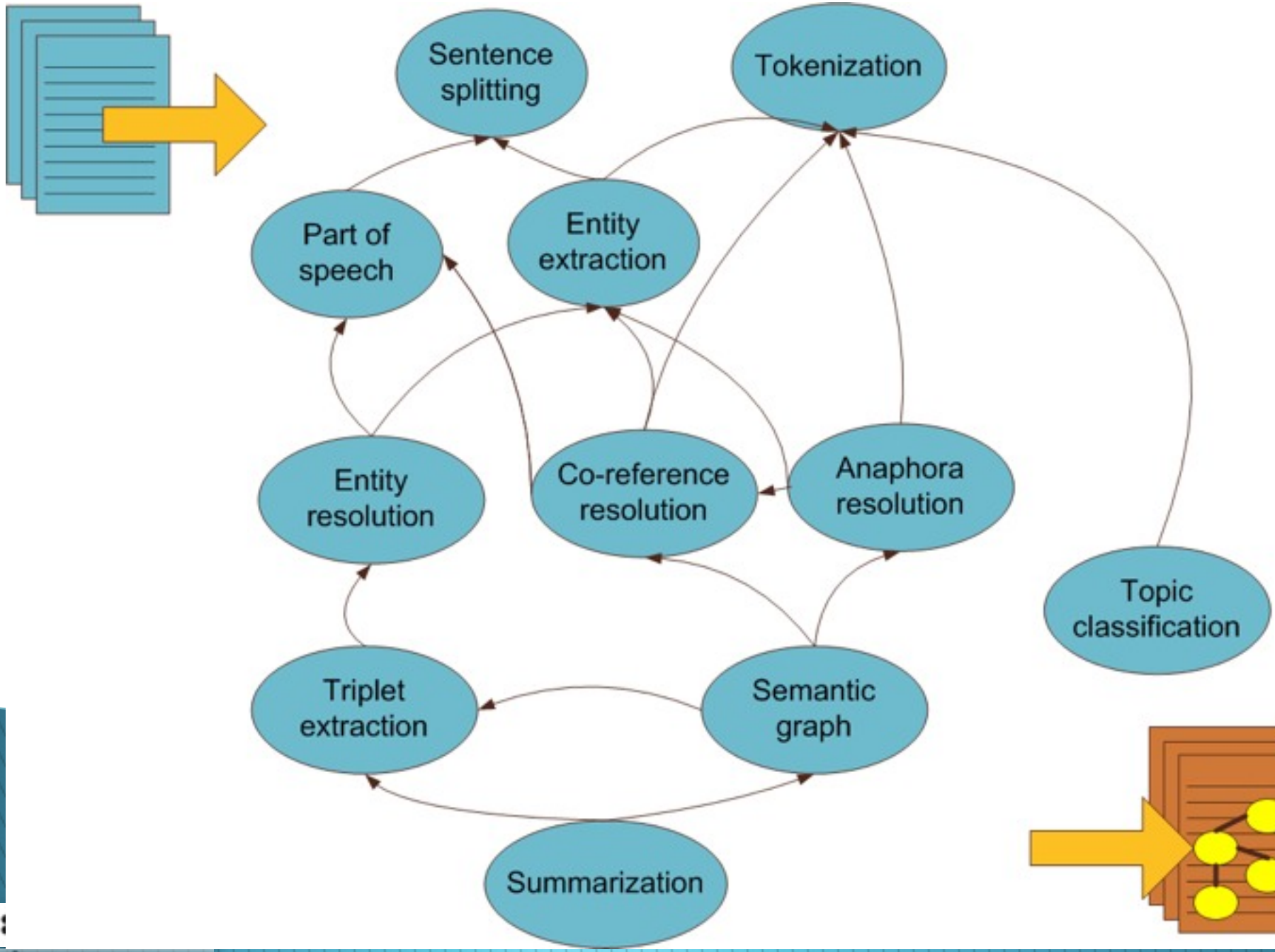


HTML

Web service



"Jožef Stefan" Institute


Sentence splitting
Tokenization
Part of speech
Entity extraction
Entity resolution
Co-reference resolution
Anaphora resolution
Topic classification
Triplet extraction
Semantic graph
Summarization

Knowledge based summarization

AAAI 2005

"Jožef Stefan" Institute

Summarization via semantic graphs

Original document

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

Hussein sought to circumvent the economic noose around his country. Japan, forcing the sanctions. Hoping to defuse criticism that it is not doing its share to affected by the U.N. embargo on Iraq, President Bush on Tuesday night promised a `Saddam Hussein will fail" to make his conquest of Kuwait permanent. ``America military may remain in the Saudi Arabian desert indefinitely. ``I cannot predict just re than 150,000 U.S. troops have been sent to the Persian Gulf region to deter a d follow his address to Congress with a televised message for the Iraqi people, dam had offered Bush time on Iraqi TV. The Philippines and Namibia, the first of the hange for sending their own tankers to get it _ said no to the Iraqi leader. Saddam's once was seen as a more 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 maintain their military sales to Iraq. A well-informed source in Tehran told The Associated Press that Iran has agreed to an Iraqi request to exchange food and medicine for up to 200,000 barrels of refined oil a day and cash payments. There was no official comment from Tehran or Baghdad on the reported food-for-oil deal. But the source, who requested anonymity, said the deal was struck during 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 The AP, said that although Iran 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 to refineries in the gulf war. Along similar lines, ABC News reported that following Aziz's visit, Iraq is apparently prepared to give Iran all the oil it wants to make up for the damage Iraq inflicted on Iran during their conflict. Secretary of State 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 parliament Tuesday the specialists had not reneged on those contracts for fear it would jeopardize the 5,800 Soviet citizens in Iraq. In his speech, Bush said his heart went out to the families of the hundreds of Americans held hostage by Iraq, but he declared, ``Our policy cannot 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 Iraqi-occupied Kuwait this week, most of them women and children leaving their husbands behind. Saddam has said he is keeping foreign men as human shields against attack. On Monday, a planeload of 164 Westerners arrived in Baltimore from Iraq. Evacuees spoke of food shortages in Kuwait, nighttime gunfire and Iraqi roundups of young people suspected of involvement in the resistance. ``There is no law and order," said Thuraya, 19, who would not give her last name. ``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 males residing in Iraq and Kuwait who were born in Arab countries will be allowed to leave. Iraq generally has not let American males leave. It was not known how many men the Iraqi move could affect. _A Pentagon spokesman said ``some increase in military activity" had been detected inside Iraq near its borders with Turkey and Syria. He said there was 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 $15 billion in the next fiscal year beginning Oct. 1. Cheney promised disgruntled lawmakers ``a significant increase" in help from Arab nations and other U.S. allies for Operation Desert Shield. Japan, which has been accused of responding too slowly to the crisis in the gulf, said Tuesday it may give $2 billion to Egypt, Jordan and Turkey, hit hardest by the U.N. prohibition on trade with Iraq. ``The pressure from abroad is getting so strong," said Hiroyasu 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 sent as early as mid-September. On Friday, Treasury Secretary Nicholas Brady visited Tokyo on a world tour seeking $10.5 billion to help Egypt, Jordan and Turkey. Japan has already promised a $1 billion aid package for multinational peacekeeping forces in Saudi Arabia, including food, water, vehicles and prefabricated housing for non-military uses. But critics in the United States have said Japan should do more because its economy depends heavily on oil from the Middle East. Japan imports 99 percent of its oil. Japan's constitution bans the use of force in settling international disputes and Japanese law restricts the military to Japanese territory, except for ceremonial occasions. On Monday, Saddam offered developing nations free oil if they would send their tankers to pick it up. The first two countries to respond Tuesday _ the Philippines and Namibia _ said no. Manila said it had already fulfilled its oil requirements, and Namibia said it would not ``sell its sovereignty" for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of free oil as a ``propaganda ploy." Venezuela, an OPEC member, has led a drive among oil-producing nations to boost production to make up for the shortfall caused [illegible] of Iraqi and Kuwaiti oil from the world market. Their oil makes up 20 percent of the world's oil reserves. Only Saudi Arabia has higher reserves. But acc[illegible] State Department, [illegible]nce U.N. sanctio[illegible]sed five weeks ago. [illegible]il Constantines[illegible]t claim Tuesday,

Linguistic processing and Creation of semantic graph

Select a sub-graph that would characterize extracted summaries

Learn the sub-graph selection model

Sentence extraction rule

Automatically generated document summary

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.

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

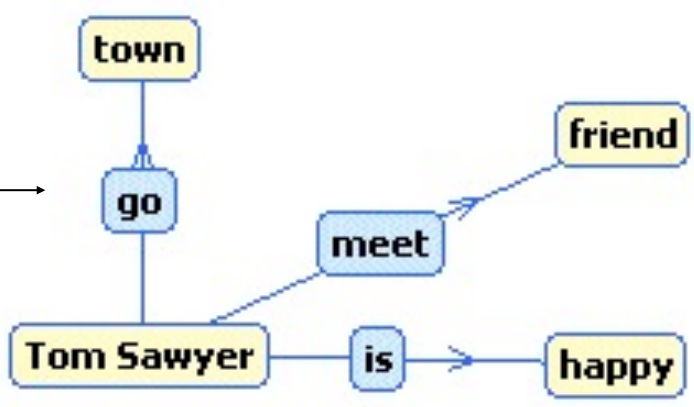
Generate a **summary graph**

Use summary graph to generate textual document summary

Tom Sawyer went to town. He met a friend. Tom was happy. …

Tom Sawyer went to town. He [Tom Sawyer] met a friend. Tom [Tom Sawyer] was happy. …

Tom ← go → town
Tom ← meet → friend
Tom ← is → happy



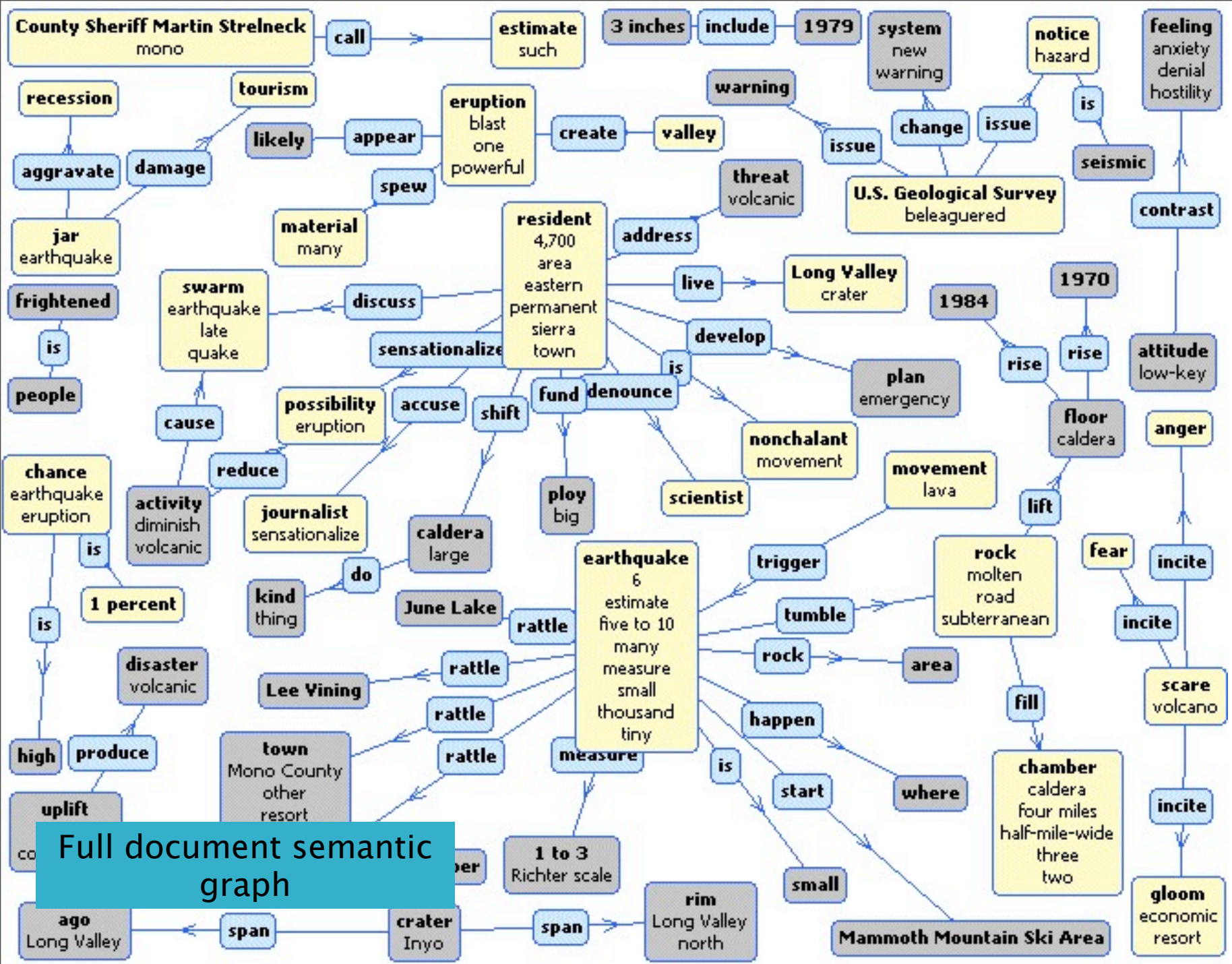
Example of automatic summary

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

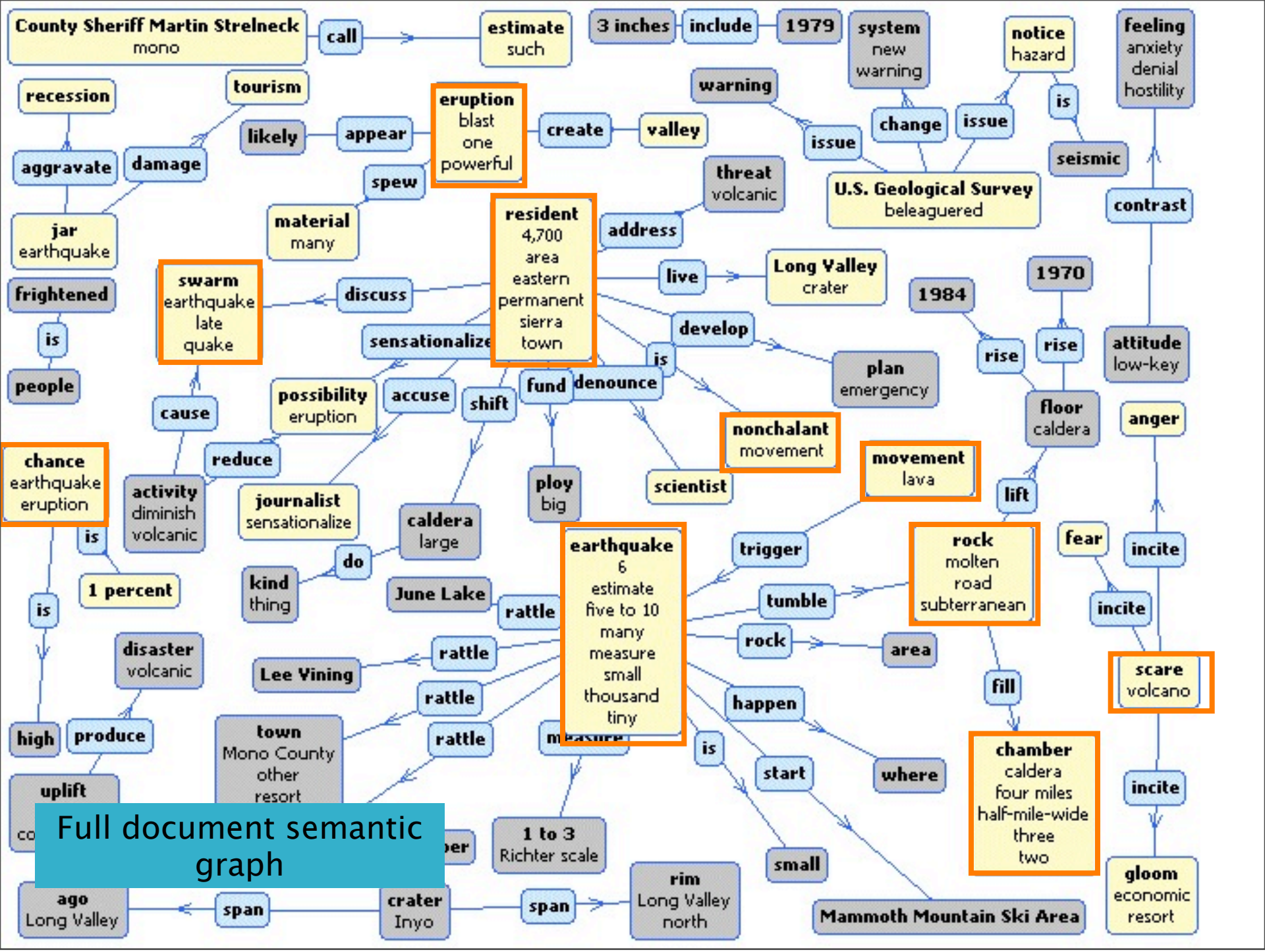
Human extracted summary

- 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.

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 to oppose Baghdad, Japan said up to $2 billion in aid may be sent to nations most affected by the U.N. embargo on Iraq. President Bush on Tuesday night promised a joint session of 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 maintain [illegible] to exchange food and medicine for [illegible] reported food-for-oil deal. But the source [illegible] by a senior Iraqi official since the [illegible] sources in the region, contacted by [illegible] for domestic use because of damages [illegible] Iran all the oil it wants to make up for [illegible] Soviet Foreign Minister Eduard Shevardnadze [illegible] summit, Bush encouraged Mikhail Gorbachev [illegible] parliament Tuesday the specialists [illegible] art went out to the families of the hundreds [illegible] world will not be blackmailed [illegible] of the Earth.'' In other developments [illegible] most of them women and children [illegible] ad of 164 Westerners arrived in Baghdad [illegible] ed of involvement in the resistance. [illegible] front of him and he can't do anything about [illegible] in Arab countries will be allowed to leave [illegible] gon spokesman said ``some increase in [illegible] ies are imminent. Defense Secretary Dick Cheney [illegible] lly used by government officials. Here [illegible] mised disgruntled lawmakers ``a significant [illegible] onding too slowly to the crisis in the Persian [illegible] The pressure from abroad is getting so strong [illegible] uld be extended through the World Bank [illegible] cholas Brady visited Tokyo on a world tour [illegible] al peacekeeping forces in Saudi Arabia [illegible] n should do more because its economy [illegible] ettling international disputes and Japanese [illegible] ons free oil if they would send their tankers [illegible] ady fulfilled its oil requirements, and Namibia said it would not ``sell its sovereignty'' for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of free oil as a ``propaganda ploy.'' Venezuela, an OPEC member, has led a drive among oil-producing nations to boost production to make up for the shortfall caused by the loss of Iraqi and Kuwaiti 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, Cuba, which faces an oil deficit because of reduced Soviet deliveries, has received a shipment of Iraqi petroleum since U.N. sanctions were imposed five weeks ago. And Romania, it said, expects to receive oil indirectly from Iraq. Romania's ambassador to the United States, Virgil Constantinescu, denied that claim Tuesday, calling it ``absolutely false and without foundation.''.

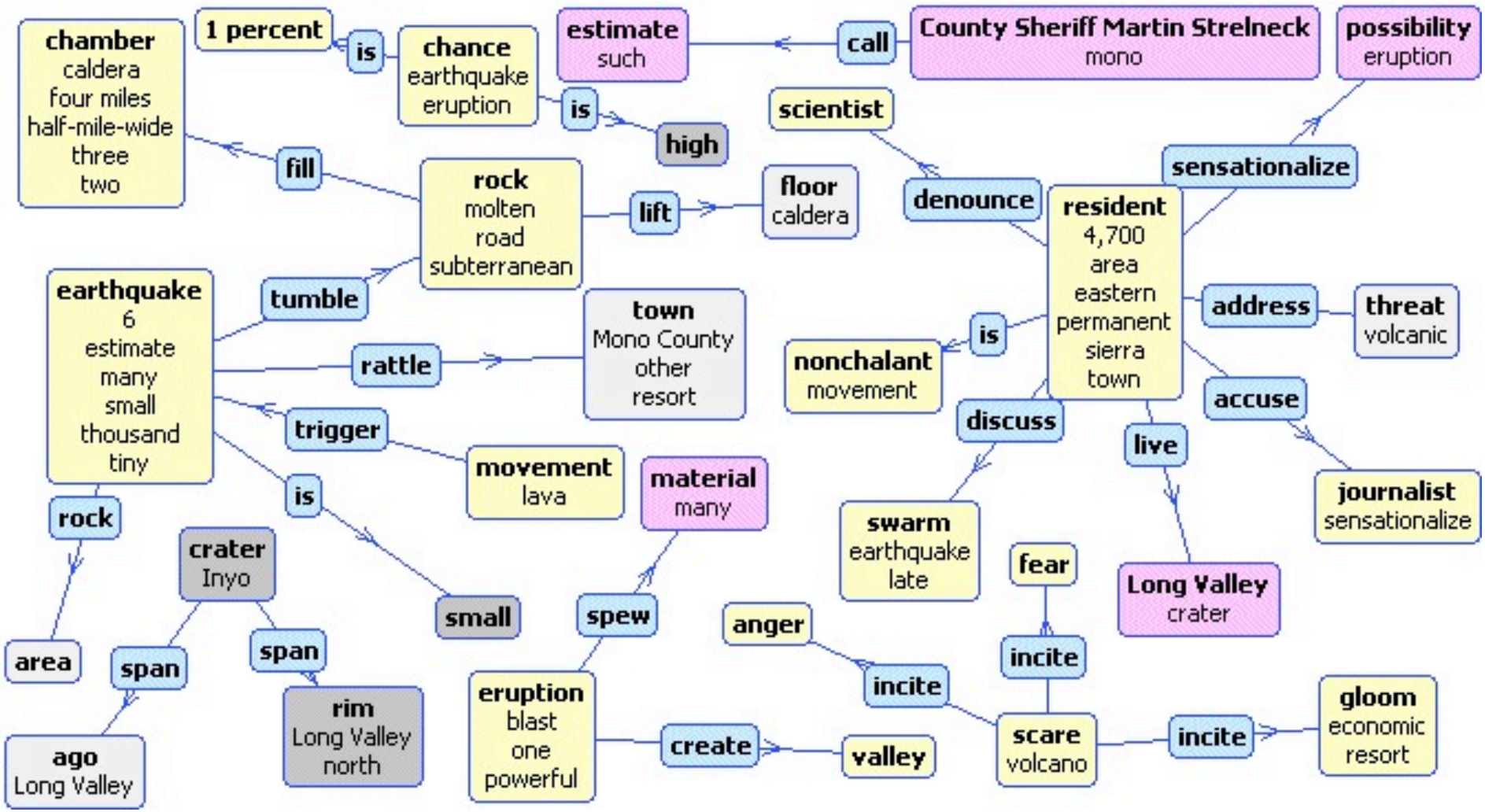


Full document semantic graph

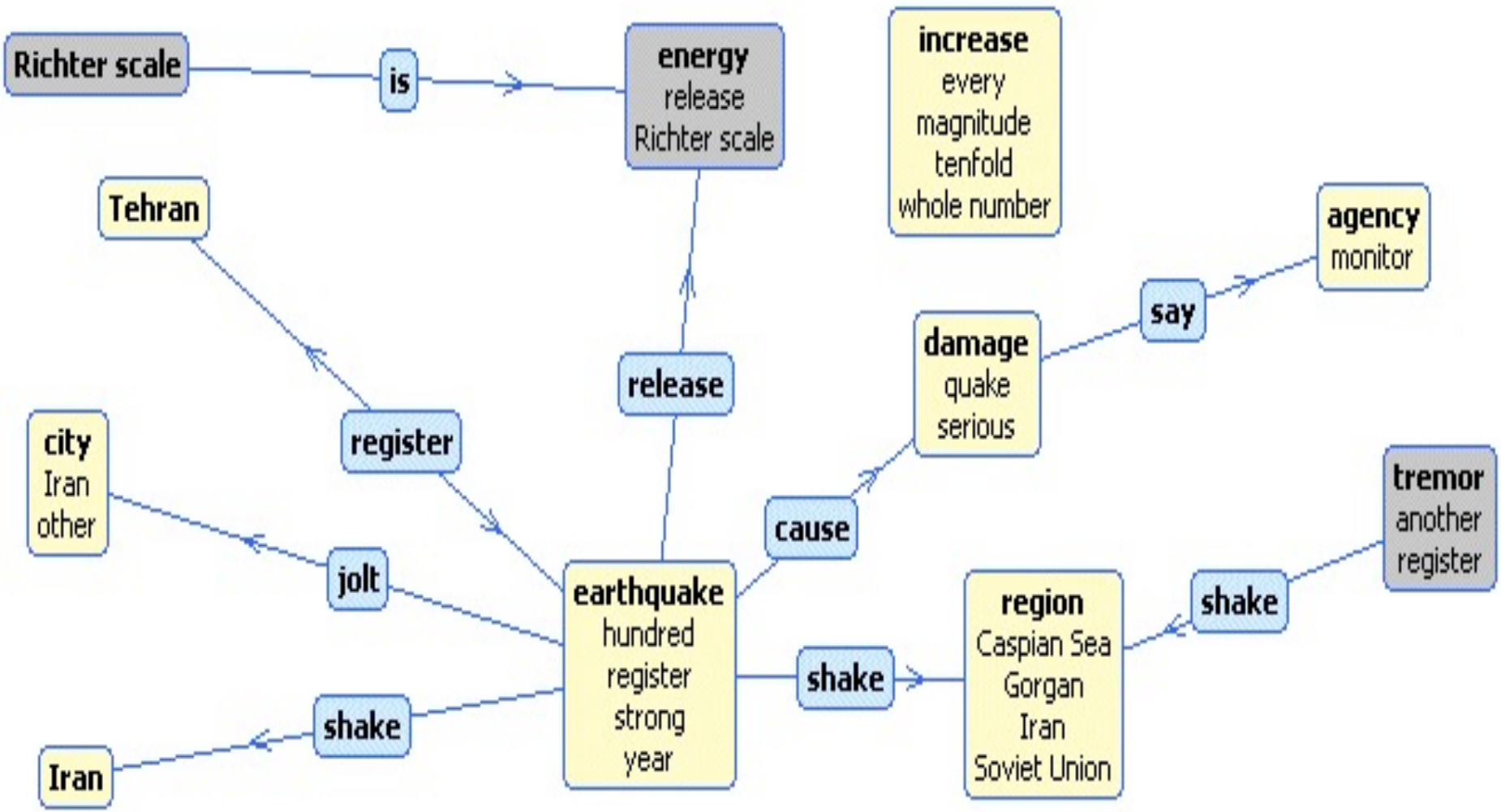


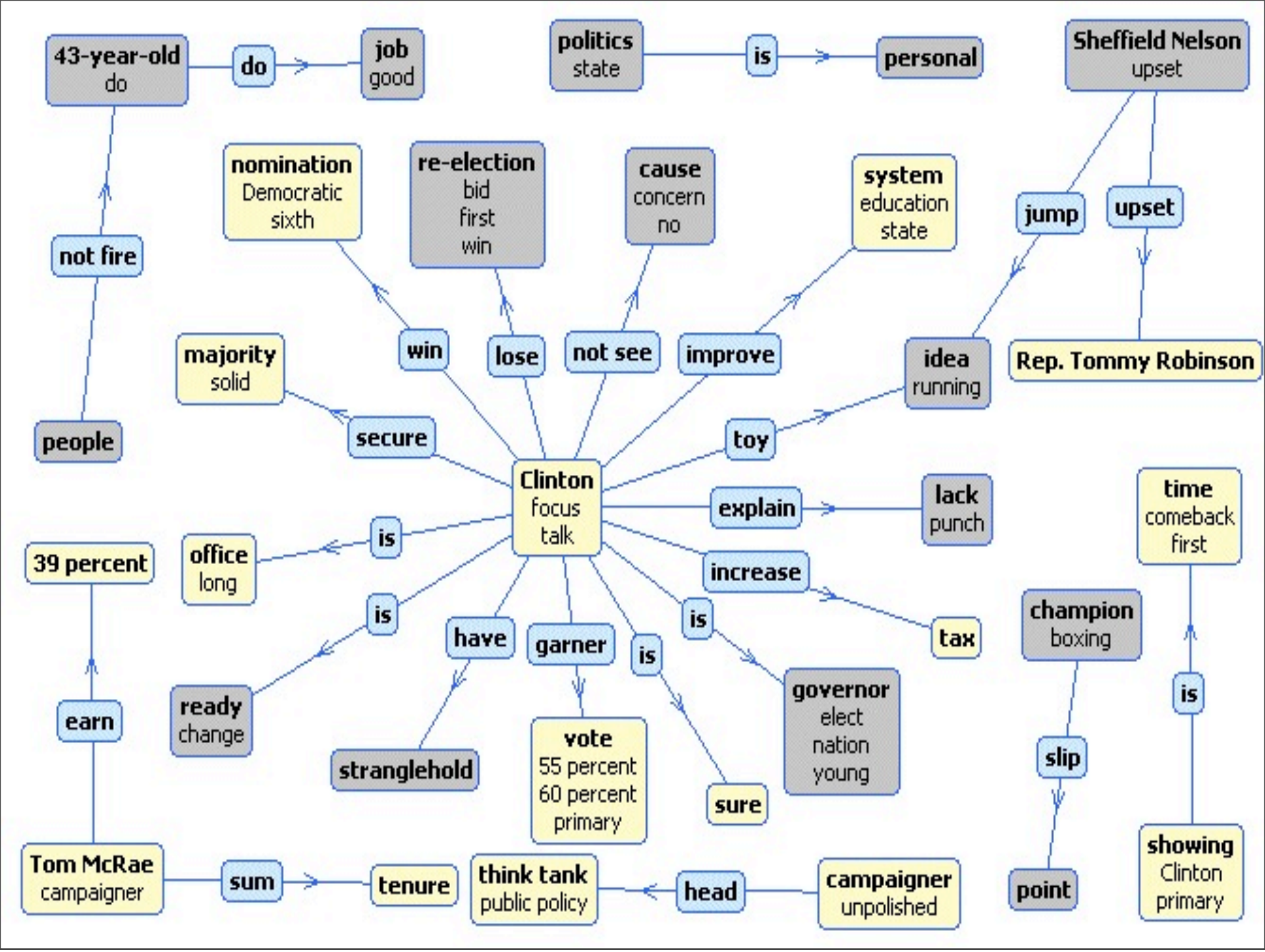
Full document semantic graph

Automatically generated summary graph



More examples



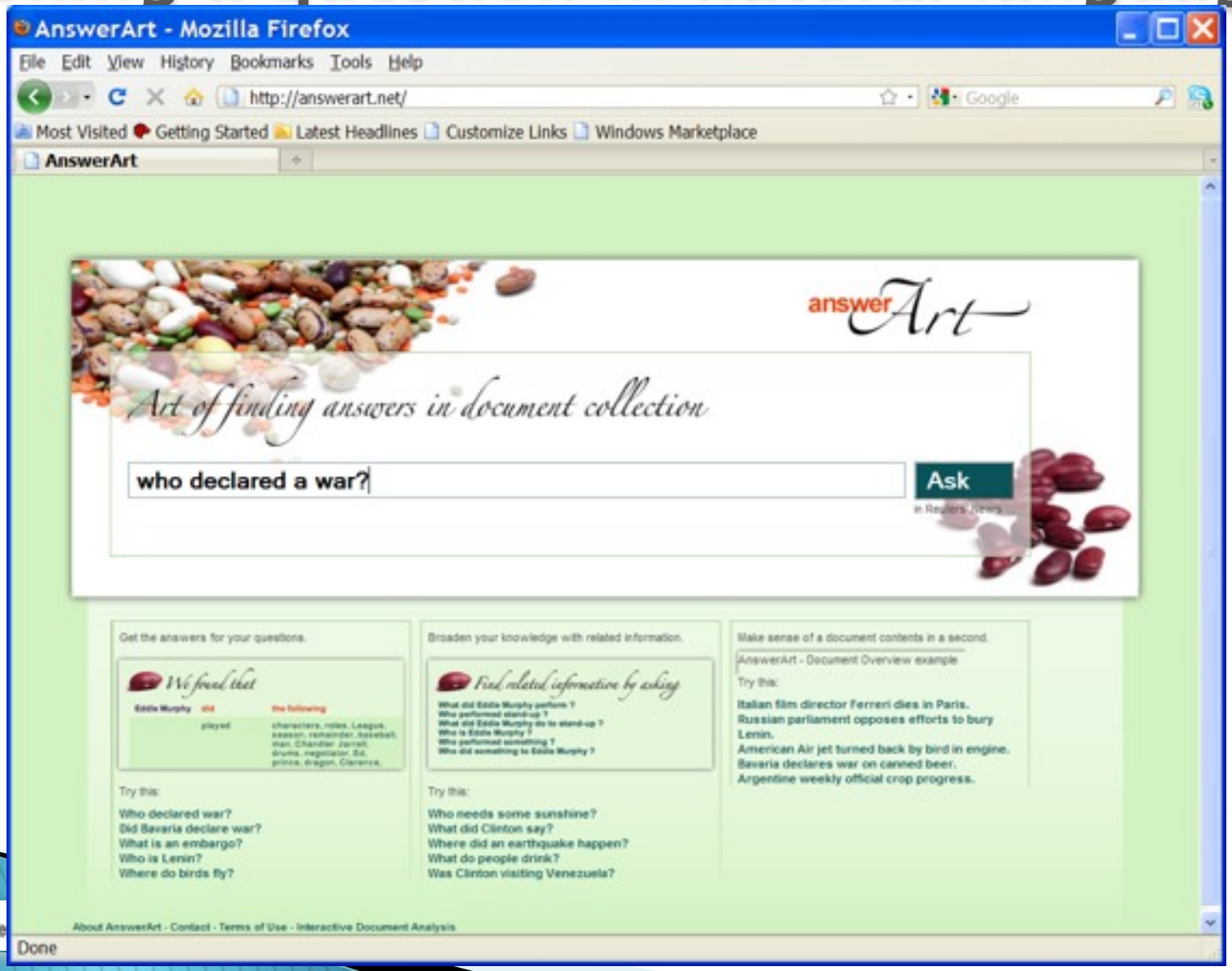

43-year-old
do
do
job
good
politics
state
is
personal
Sheffield Nelson
upset
nomination
Democratic
sixth
re-election
bid
first
win
cause
concern
no
system
education
state
jump
upset
not fire
win
lose
not see
improve
idea
running
Rep. Tommy Robinson
majority
solid
people
secure
toy
Clinton
focus
talk
explain
lack
punch
time
comeback
first
is
office
long
39 percent
increase
champion
boxing
is
is
have
garner
is
is
tax
earn
ready
change
governor
elect
nation
young
is
slip
stranglehold
vote
55 percent
60 percent
primary
sure
showing
Clinton
primary
Tom McRae
campaigner
sum
tenure
think tank
public policy
head
campaigner
unpolished
point

Question Answering

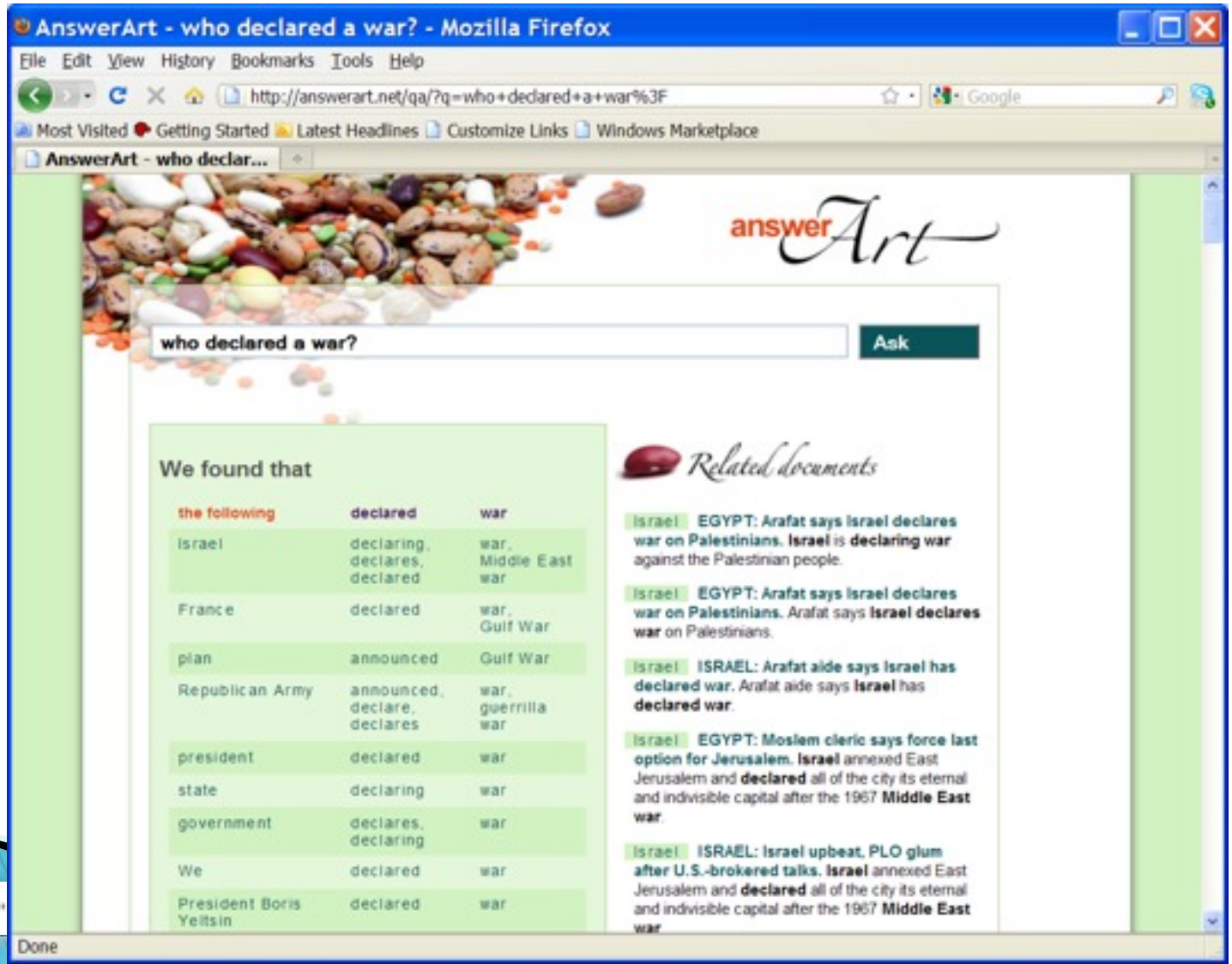
http://AnswerArt.net

"Jožef Stefan" Institute

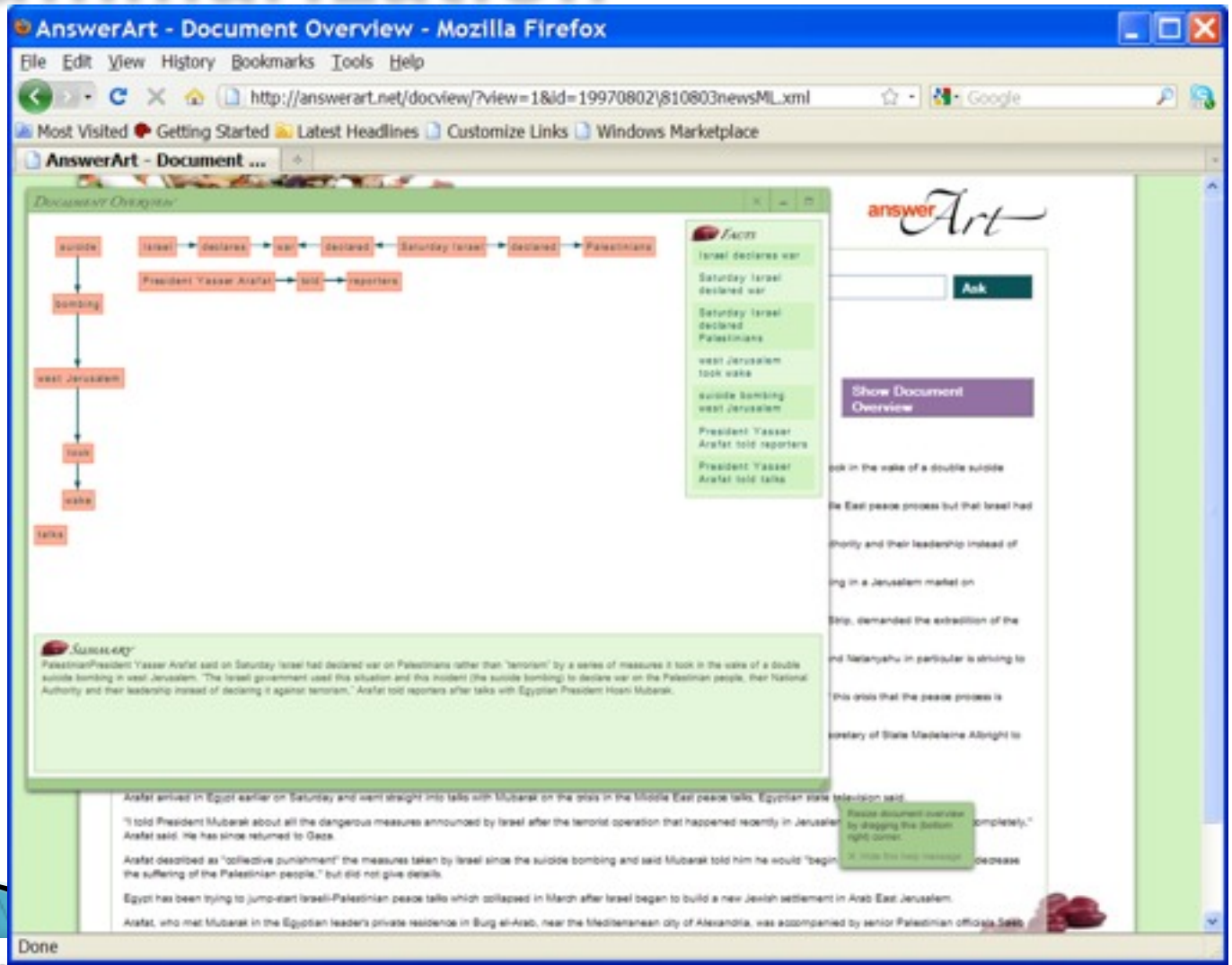
Asking a question in natural language



…result set is set of mentions of potential answers



Document browser with summarization



"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

Syntactic

- Cross-modality

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

Semantic

"Jožef Stefan" Institute

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

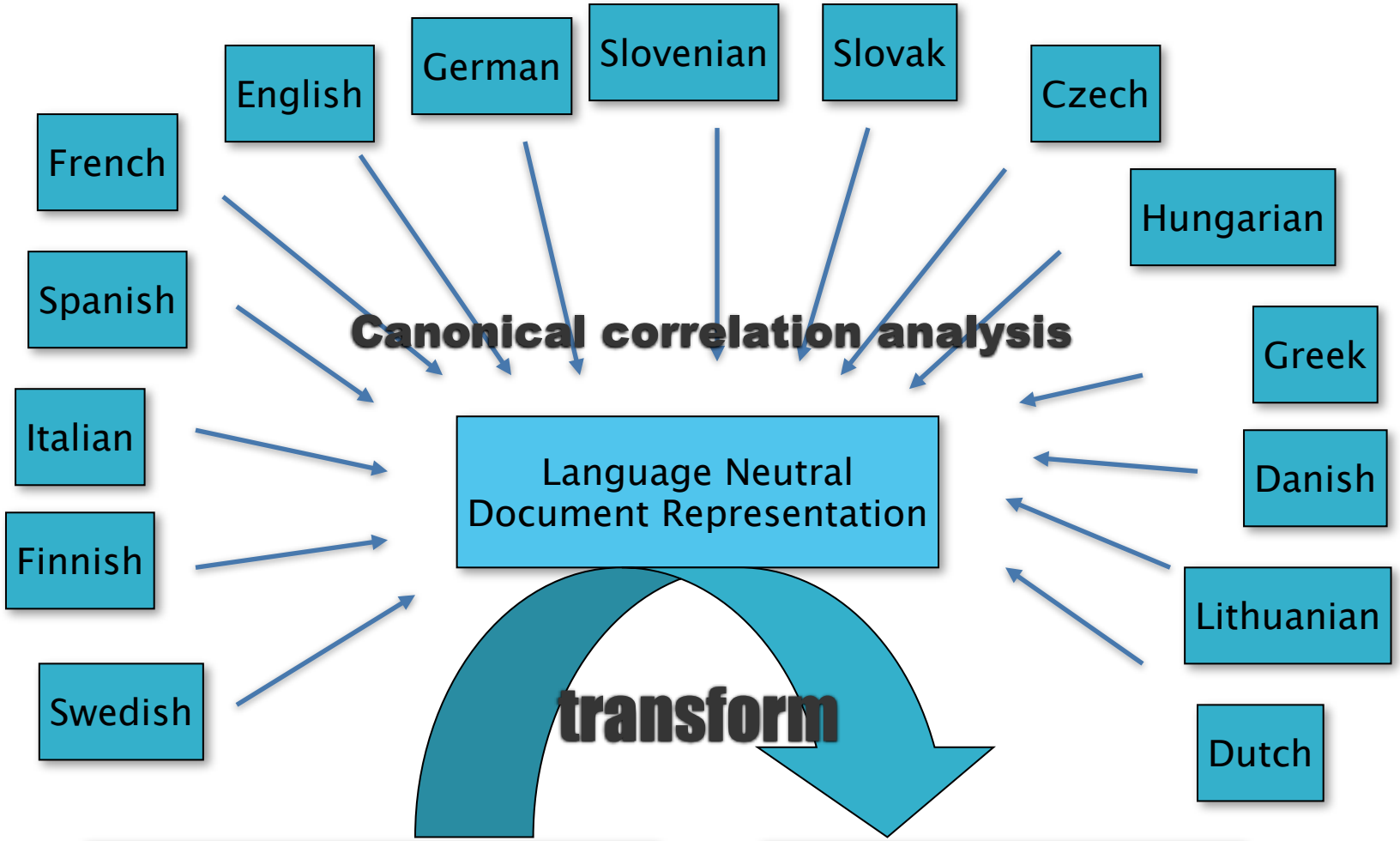
- …the question is how to create mappings between different representation so that we can benefit using more information about the same objects

"Jožef Stefan" Institute

Multilingual search

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

"Jožef Stefan" Institute



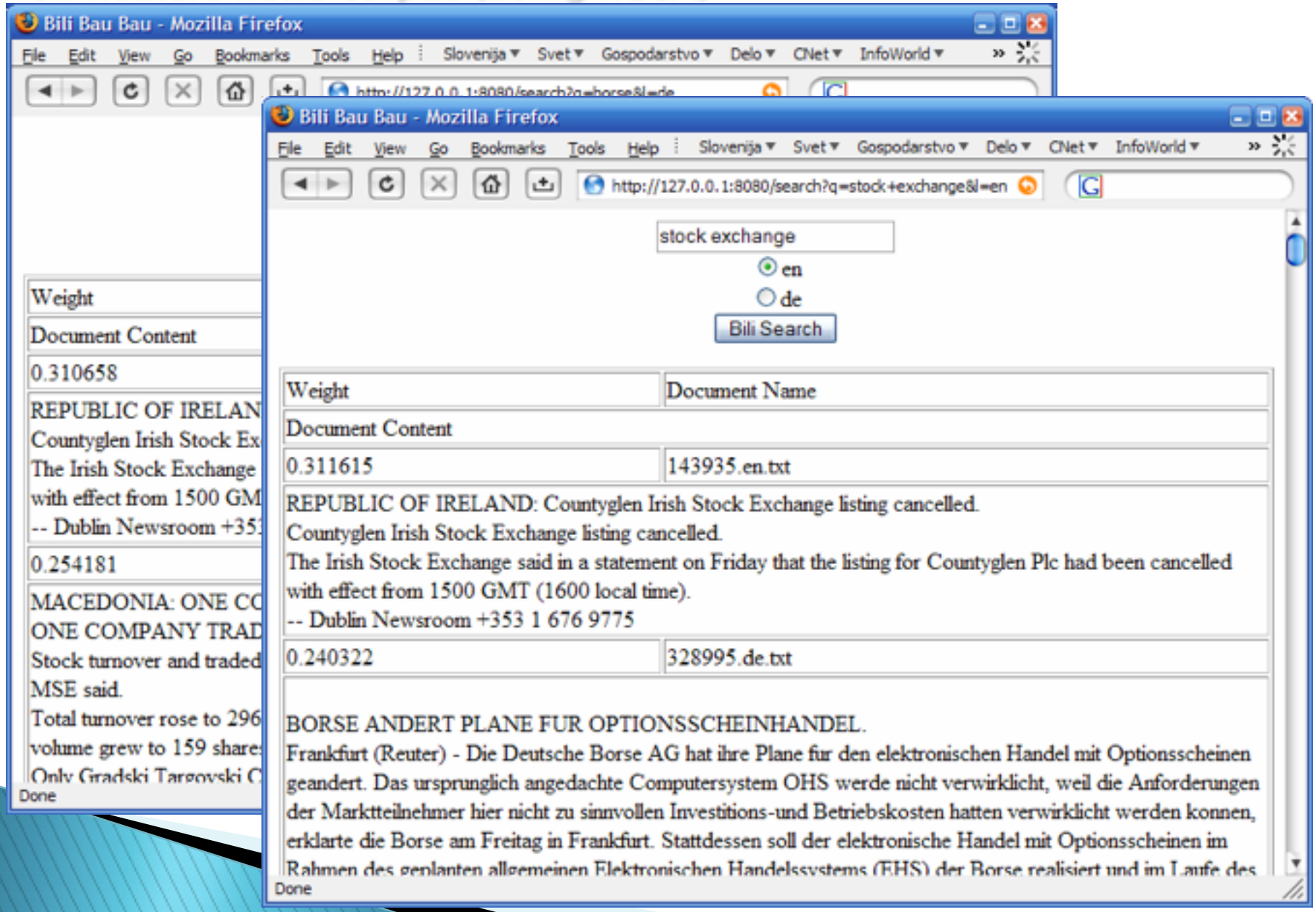
New document represented as text in **any** of the above languages

New document represented in **Language Neutral** way

…enables cross-lingual retrieval, categorization, clustering, …

SMART
Statistical Multilingual Analysis for Retrieval and Translation

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



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

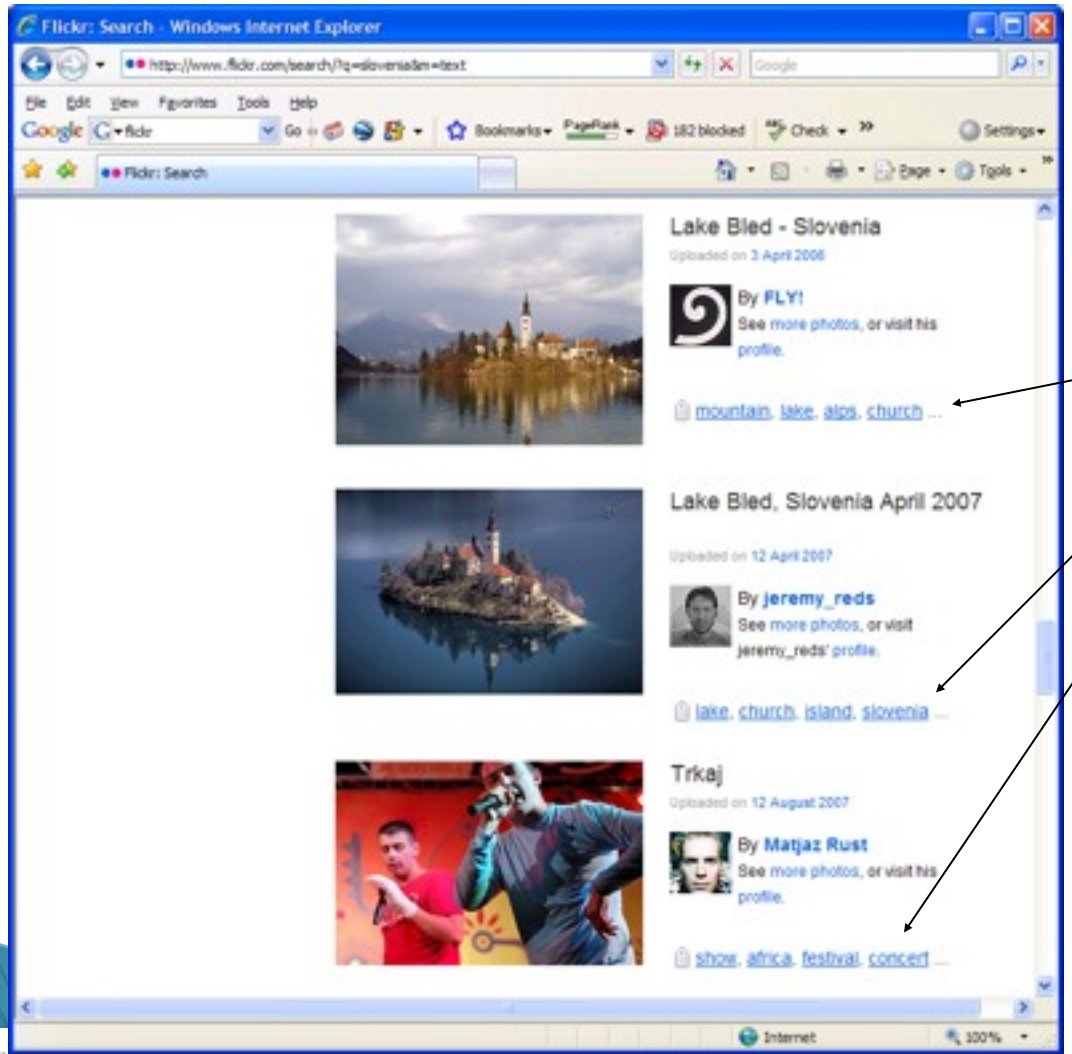
Semantic

"Jožef Stefan" Institute

Collaborative tagging

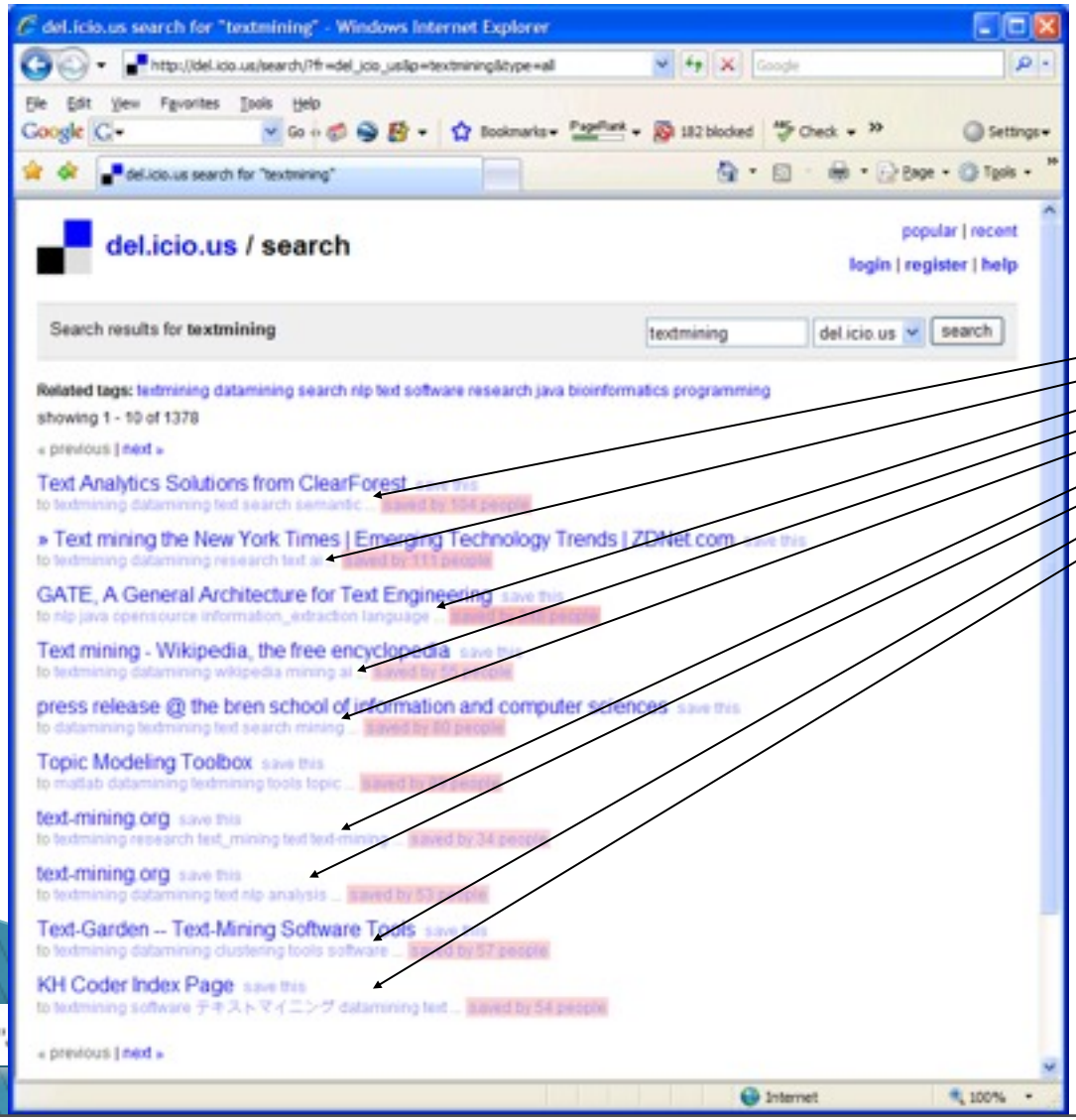
- Collaborative tagging is a process of adding metadata to annotate content (e.g. documents, web sites, photos)
  - …metadata is typically in the form of keywords
  - …this is done in a collaborative way by many users from larger community collectively having good coverage of many topics
  - …as a result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

Example: flickr.com tagging



Tags entered by users annotating photos

Example: del.icio.us tagging



Tags entered by users annotating Web sites

Levels of text representations

- Character (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

"Jožef Stefan" Institute

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

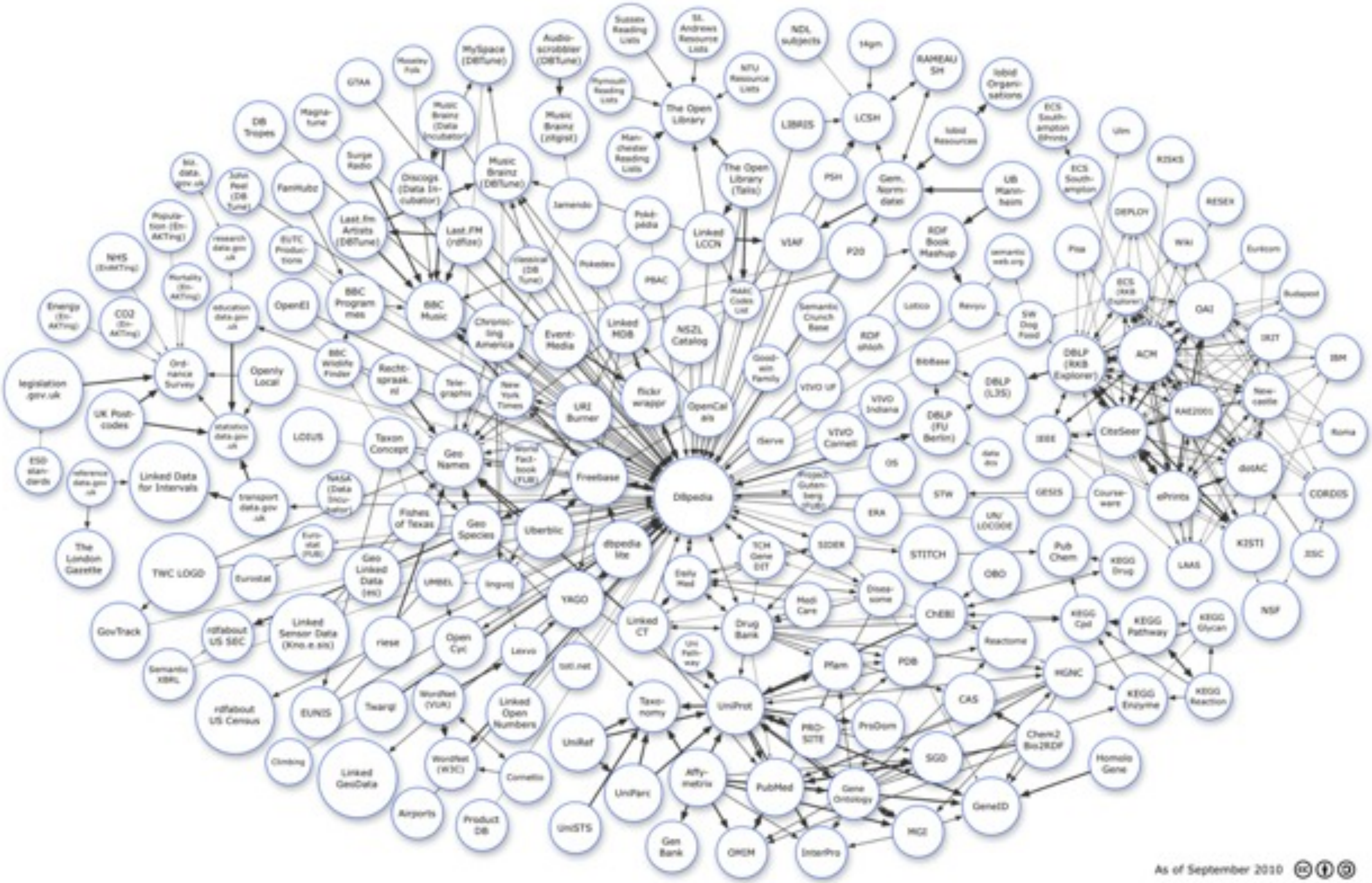
- The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data

- 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/

"Jožef Stefan" Institute

Linked Open Graph



http://richard.cyganiak.de/2007/10/lod/

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

"Jožef Stefan" Institute

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

"Jožef Stefan" Institute

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

"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

Syntactic

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

Semantic

"Jožef Stefan" Institute

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

"Jožef Stefan" Institute

Cyc Knowledge Base and Reasoning

"Jožef Stefan" Institute

The Cyc Ontology

Cyc contains:

15,000 Predicates

300,000 Concepts

3,200,000 Assertions

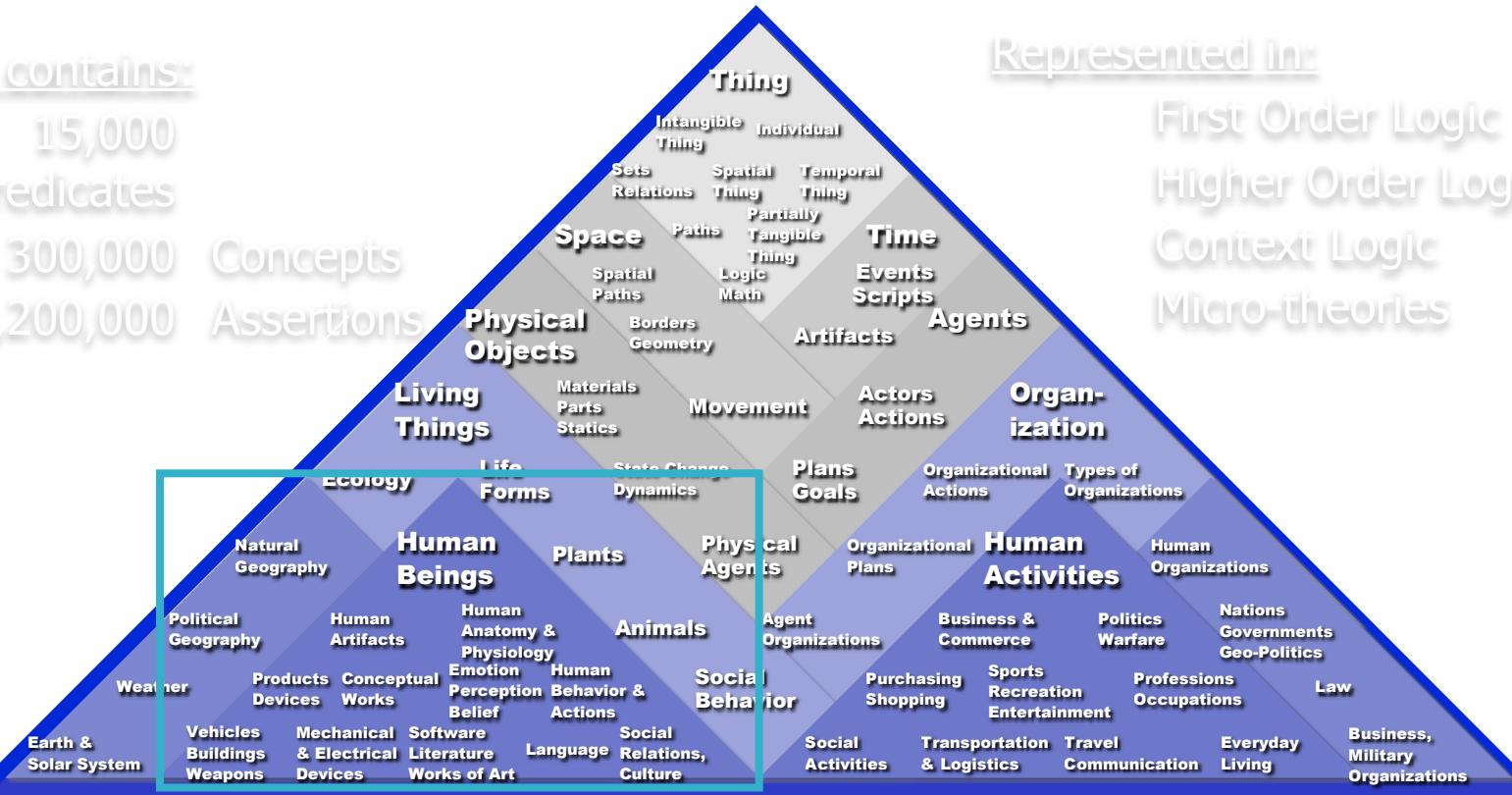
Represented in:

First Order Logic

Higher Order Logic

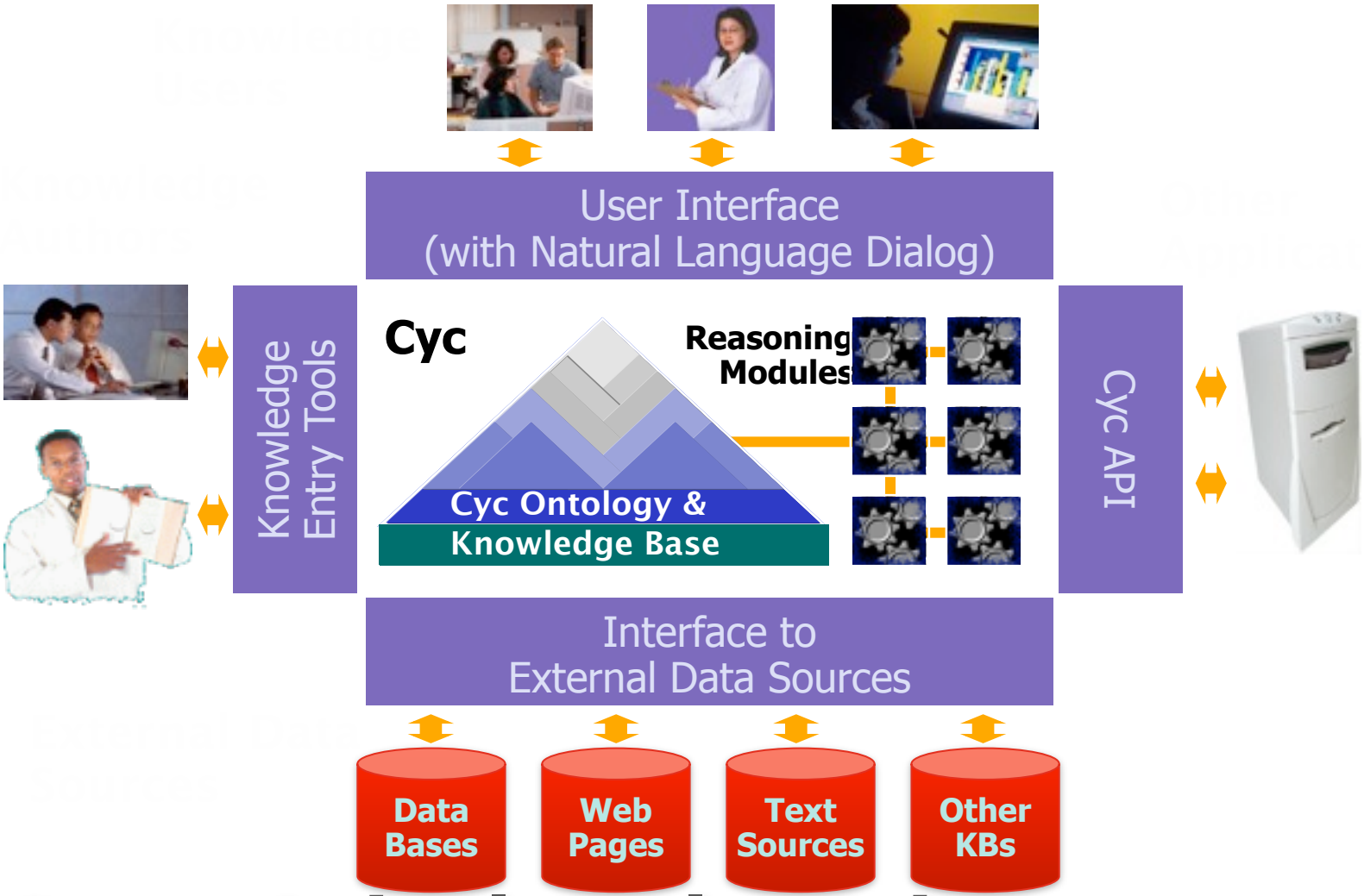
Context Logic

Micro-theories



General Knowledge about Various Domains

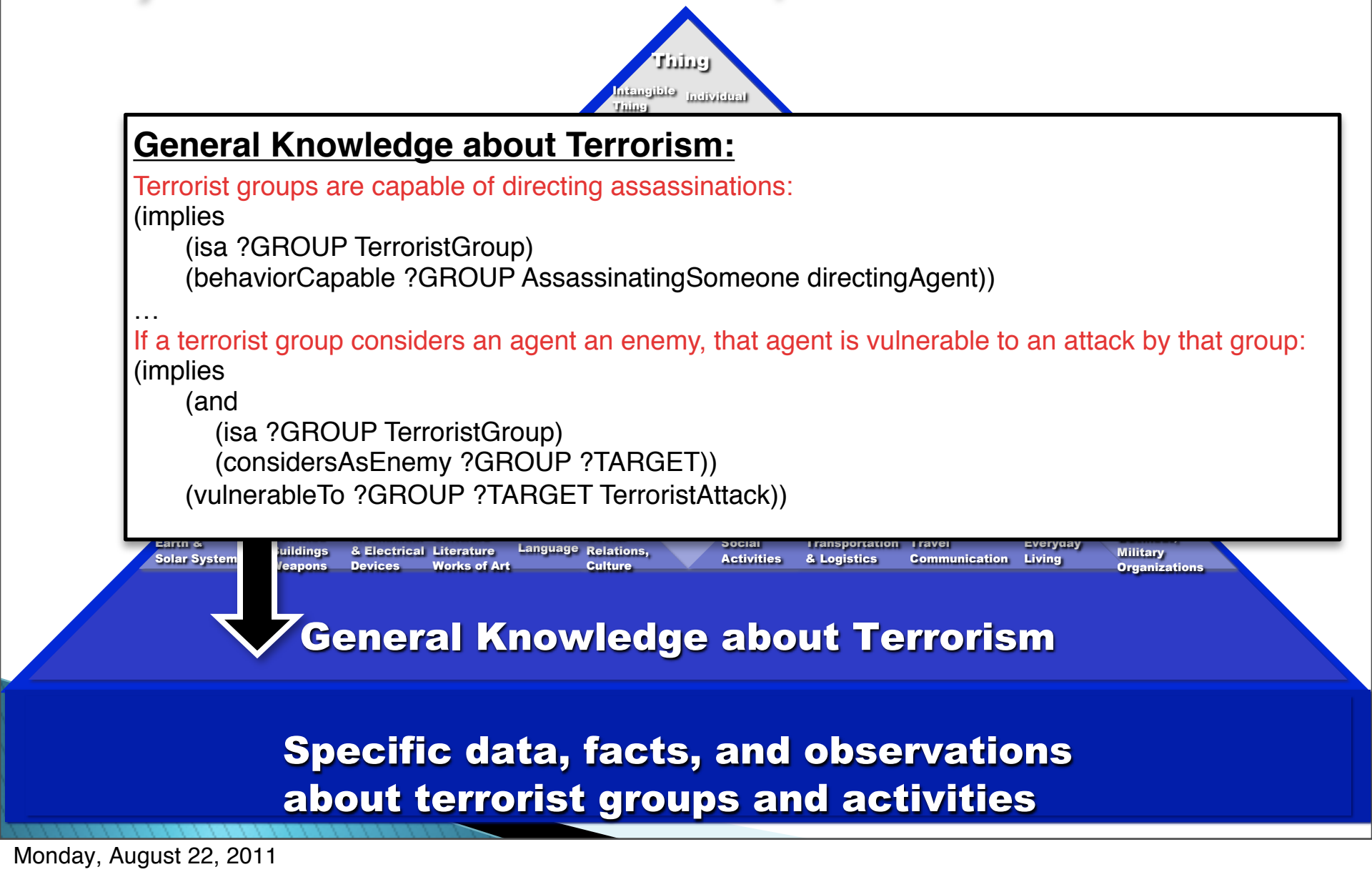
Specific data, facts, and observations



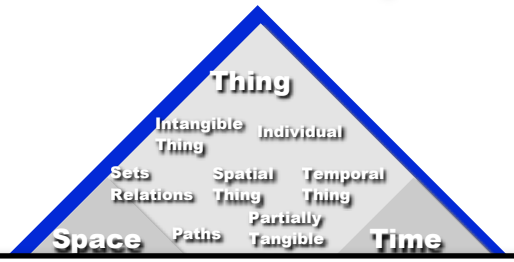
Cyc High-level Architecture

C. Matuszek, R.C. Kahlert
FACTory

Cyc KB Extended w/Domain



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 HarakatUlAnsar) 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

Source: "Galileo Galilei was an Italian physicist and astronomer."

Learn Logic:
```
(#$and (#$isa #$GalileoGalilei #$ItalianPerson)
       (#$isa #$GalileoGalilei #$Physicist)
       (#$isa #$GalileoGalilei #$Astronomer))
```

Fact: Galileo was an Italian, a physicist, and an astronomer.

Source: "Galileo was born in Pisa on Feburary 15, 1564."

Learn Logic:
```
(#$and (#$birthDate #$GalileoGalilei
                    (#$DayFn 15
                    (#$MonthFn #$February
                (#$YearFn 1564))))
       (#$birthPlace #$GalileoGalilei #$CityOfPisaItaly))
```

Fact: Galileo was born on February 15, 1564 and he was born in Pisa.

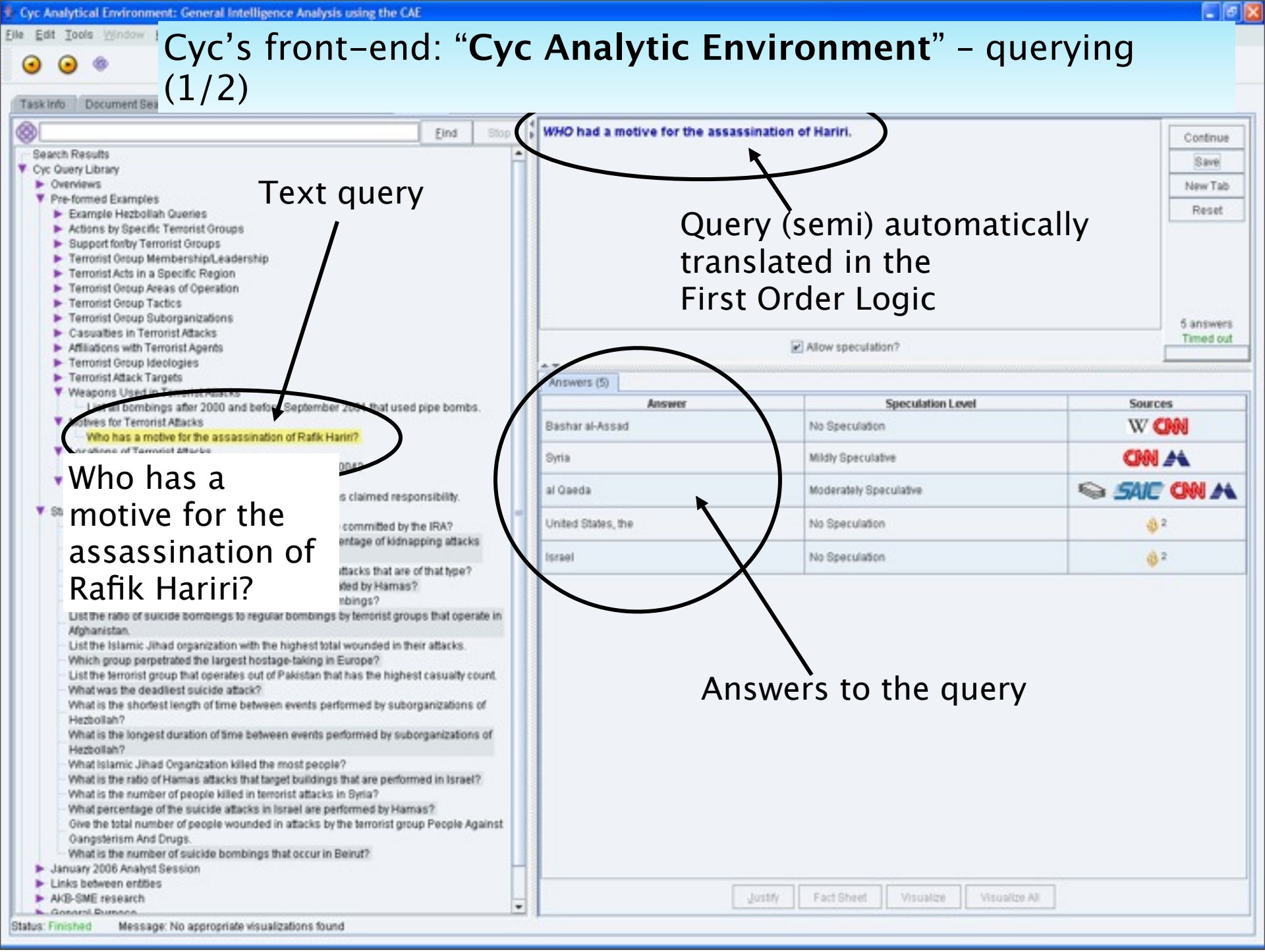
Source: "Albert Einstein was born in 1879 in Ulm, Germany."

Learn Logic: (#$birthDate #$AlbertEinstein (#$YearFn 1879))

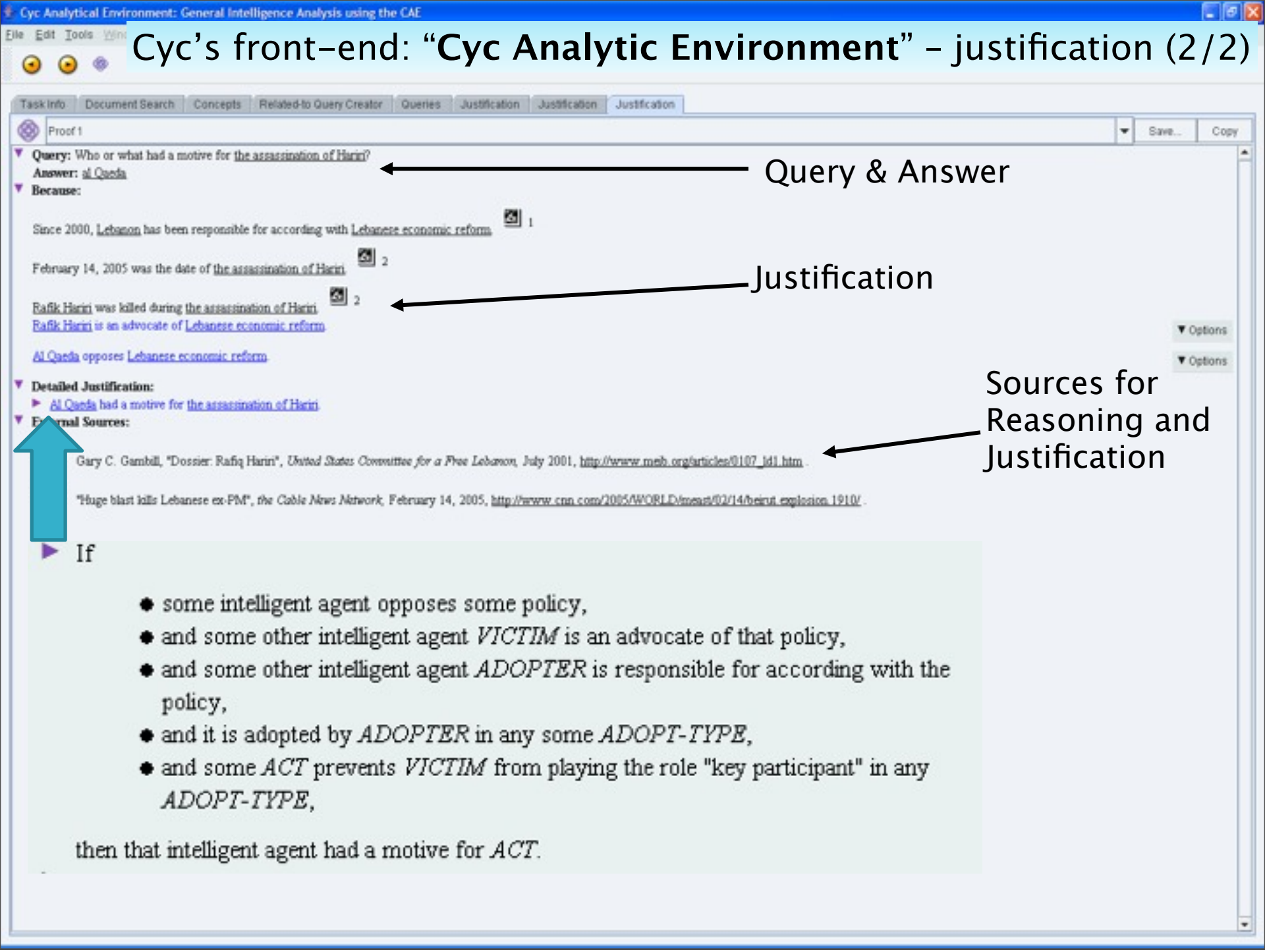
Fact: Albert Einstein was born in 1879.

"Jožef Stefan" Institute

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



Cyc's front-end: "**Cyc Analytic Environment**" – justification (2/2)



If

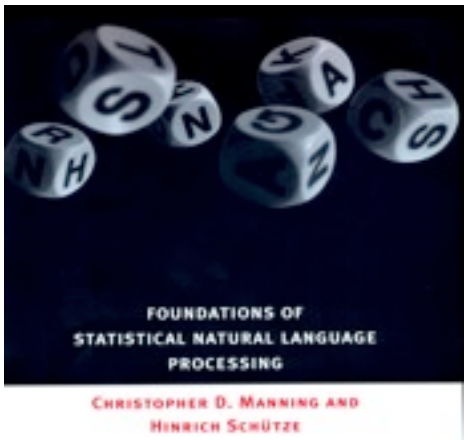
- some intelligent agent opposes some policy,
- and some other intelligent agent *VICTIM* is an advocate of that policy,
- and some other intelligent agent *ADOPTER* is responsible for according with the policy,
- and it is adopted by *ADOPTER* in any some *ADOPT-TYPE*,
- and some *ACT* prevents *VICTIM*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

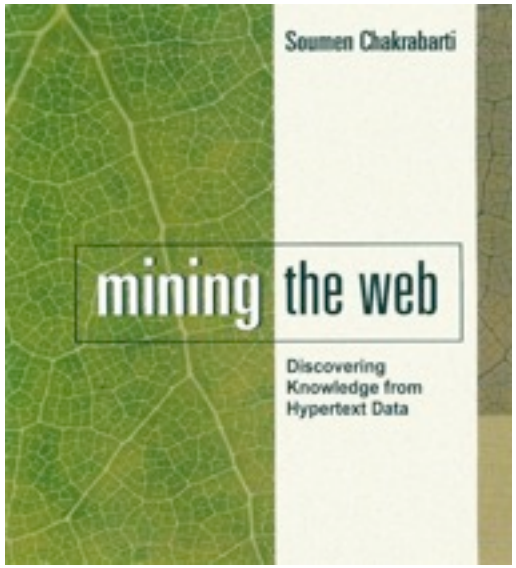
then that intelligent agent had a motive for *ACT*.

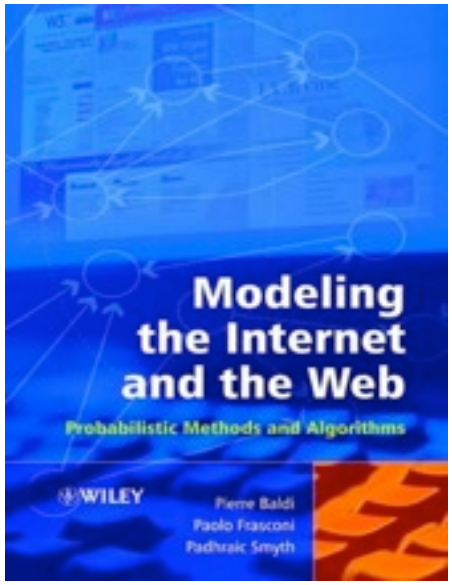
Further references…

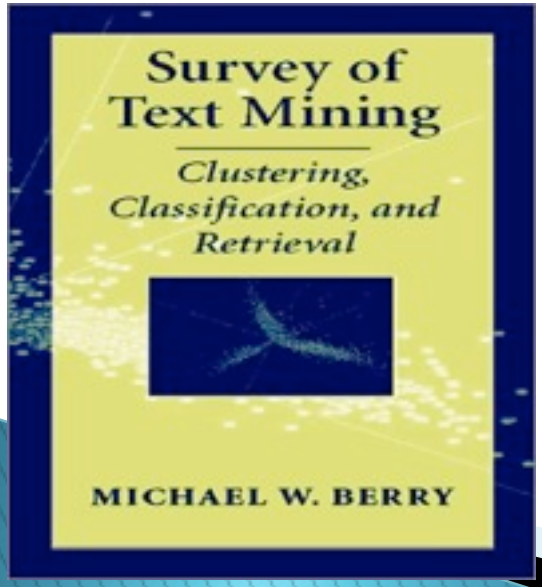
"Jožef Stefan" Institute

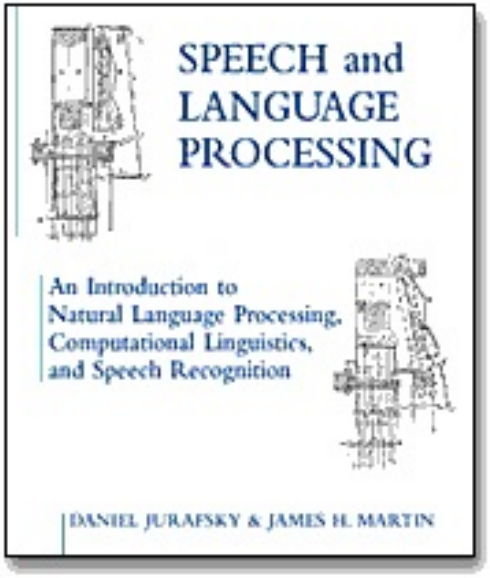
References to some Text-Mining books

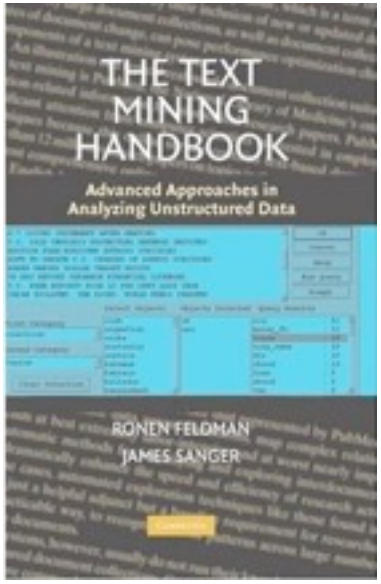




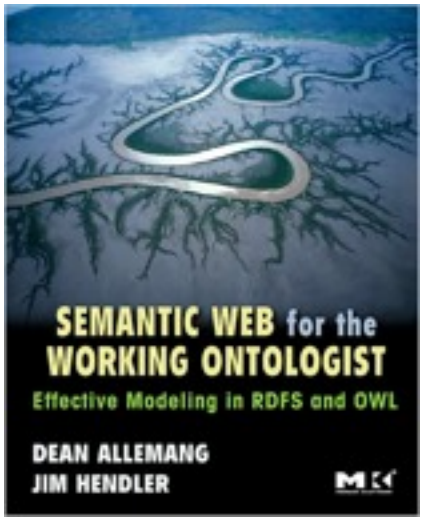


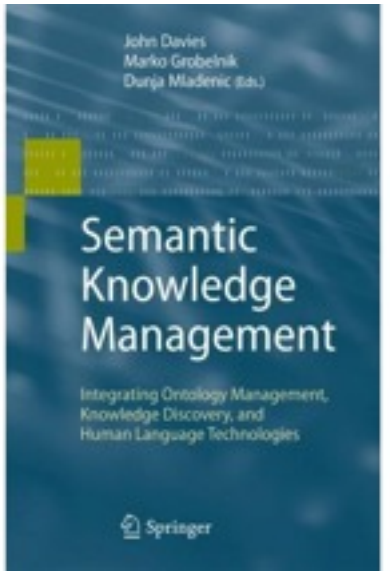


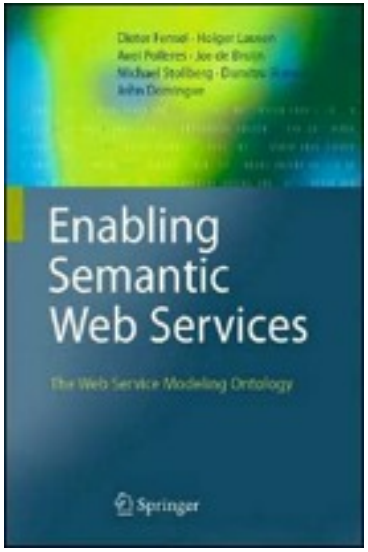


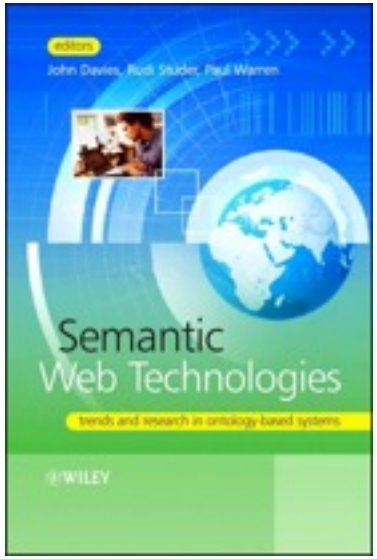


Books on Semantic Technologies









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

"Jožef Stefan" Institute