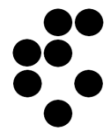




Representing Text – from characters to logic

Dunja Mladenić

Artificial Intelligence Laboratory,
J. Stefan Institute,
Slovenia





Outline

- Motivation - why representation matters
- Levels of text representation:
 - Lexical
 - Syntactic
 - Semantic
- Demos
 - SearchPoint
 - News visualization
 - AnswerArt
 - Enrycher
- ...to conclude

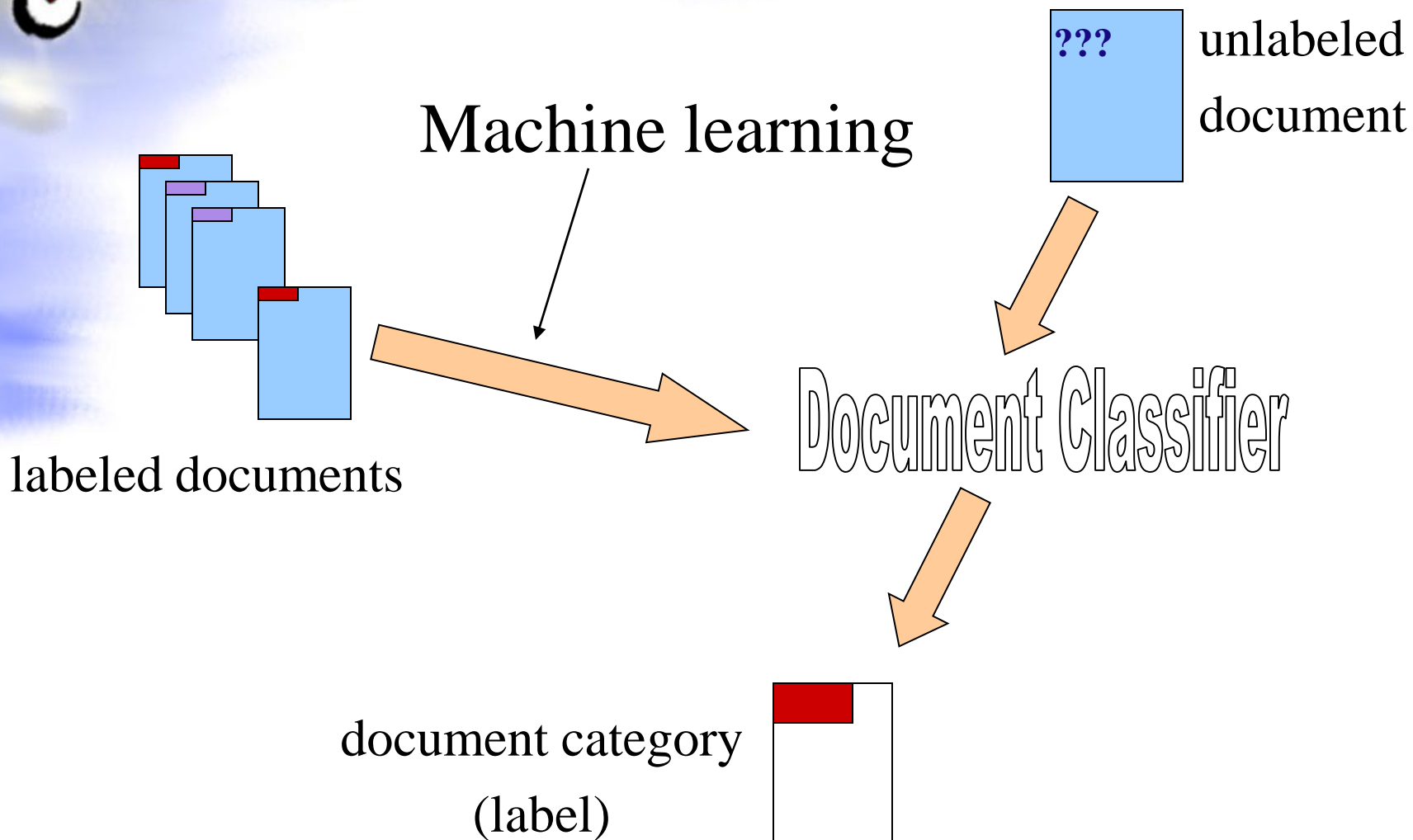


Motivation

- Machine learning – two major steps:
 - 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**



Document categorization





Easy decision problems require simple data representation

SVM - demo

Linear separation is enough to separate the data

Point class

PLUS MINUS

Clear

Learning properties

Cost: 100

Linear kernel

Polynomial kernel

Degree: 2

Gaussian kernel

Sigma: 10

Do the right stuff!

Made by:
Blaz Fortuna
<http://kt.ijs.si/blazf/index.html>

Jozef Stefan Institute, Slovenia
<http://www.ijs.si/>



Harder decision problems require better data representation

SVM - demo

Polynomial separation of data

Point class

PLUS MINUS

Clear

Learning properties

Cost: 100

Linear kernel

Polynomial kernel

Degree: 2

Gaussian kernel

Sigma: 10

Do the right stuff!

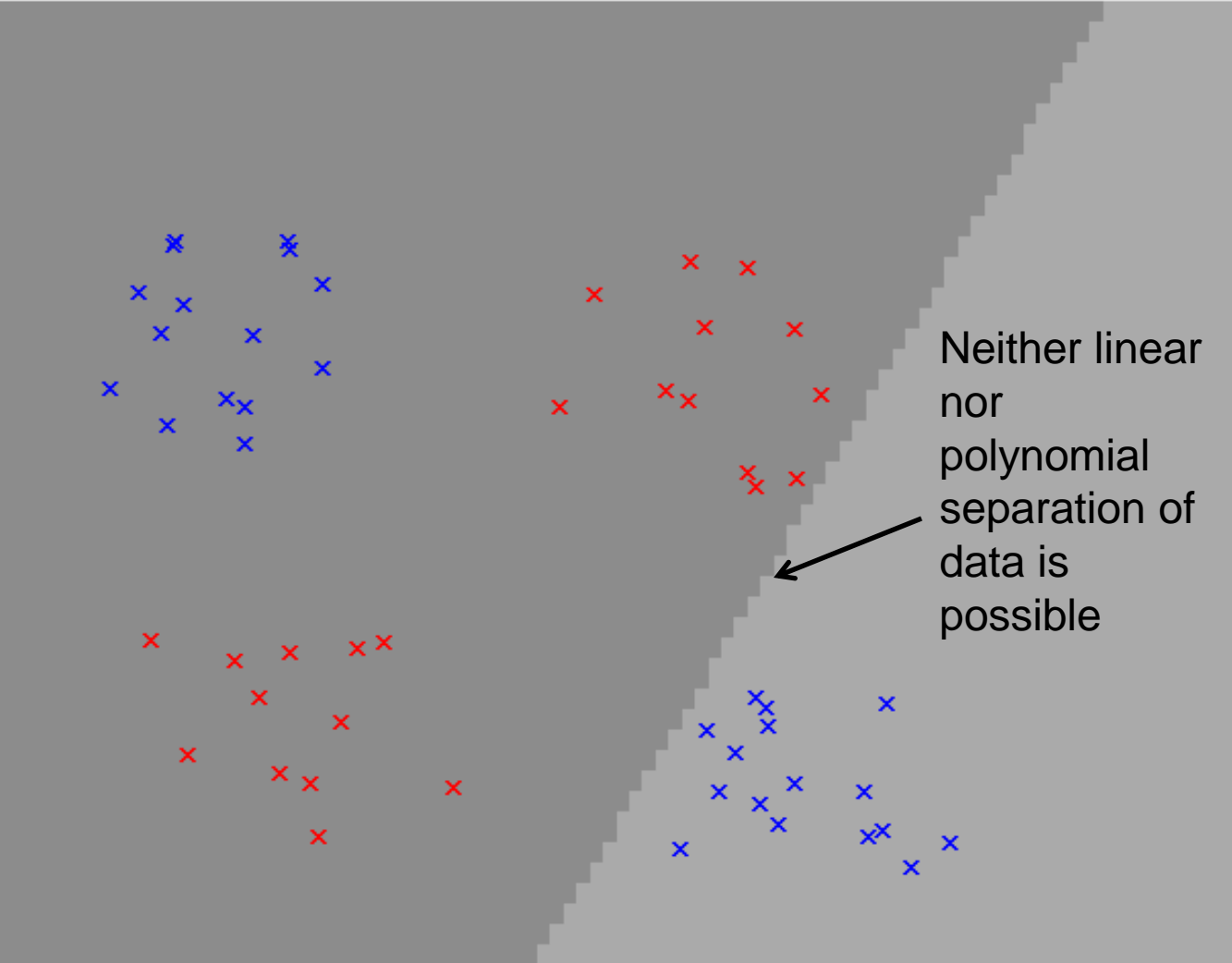
Made by:
Blaz Fortuna
<http://kt.ijs.si/blazf/index.html>

Jozef Stefan Institute, Slovenia
<http://www.ijs.si/>



Hard decision problems require sophisticated data representation

SVM - demo



Neither linear nor polynomial separation of data is possible

Point class

PLUS MINUS

Clear

Learning properties

Cost: 100

Linear kernel

Polynomial kernel

Degree: 2


Gaussian kernel

Sigma: 10

Do the right stuff!

Made by:
Blaz Fortuna
<http://kt.ijs.si/blazf/index.html>

Jozef Stefan Institute, Slovenia
<http://www.ijs.si/>





Hard decision problems require sophisticated data representation

SVM - demo

Gaussian kernel separates the data

Point class

PLUS MINUS

Clear

Learning properties

Cost: 100

Linear kernel

Polynomial kernel

Degree: 2

Gaussian kernel

Sigma: 10

Do the right stuff!

Made by:
Blaz Fortuna
<http://kt.ijs.si/blazf/index.html>

Jozef Stefan Institute, Slovenia
<http://www.ijs.si/>



Representing Text

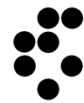
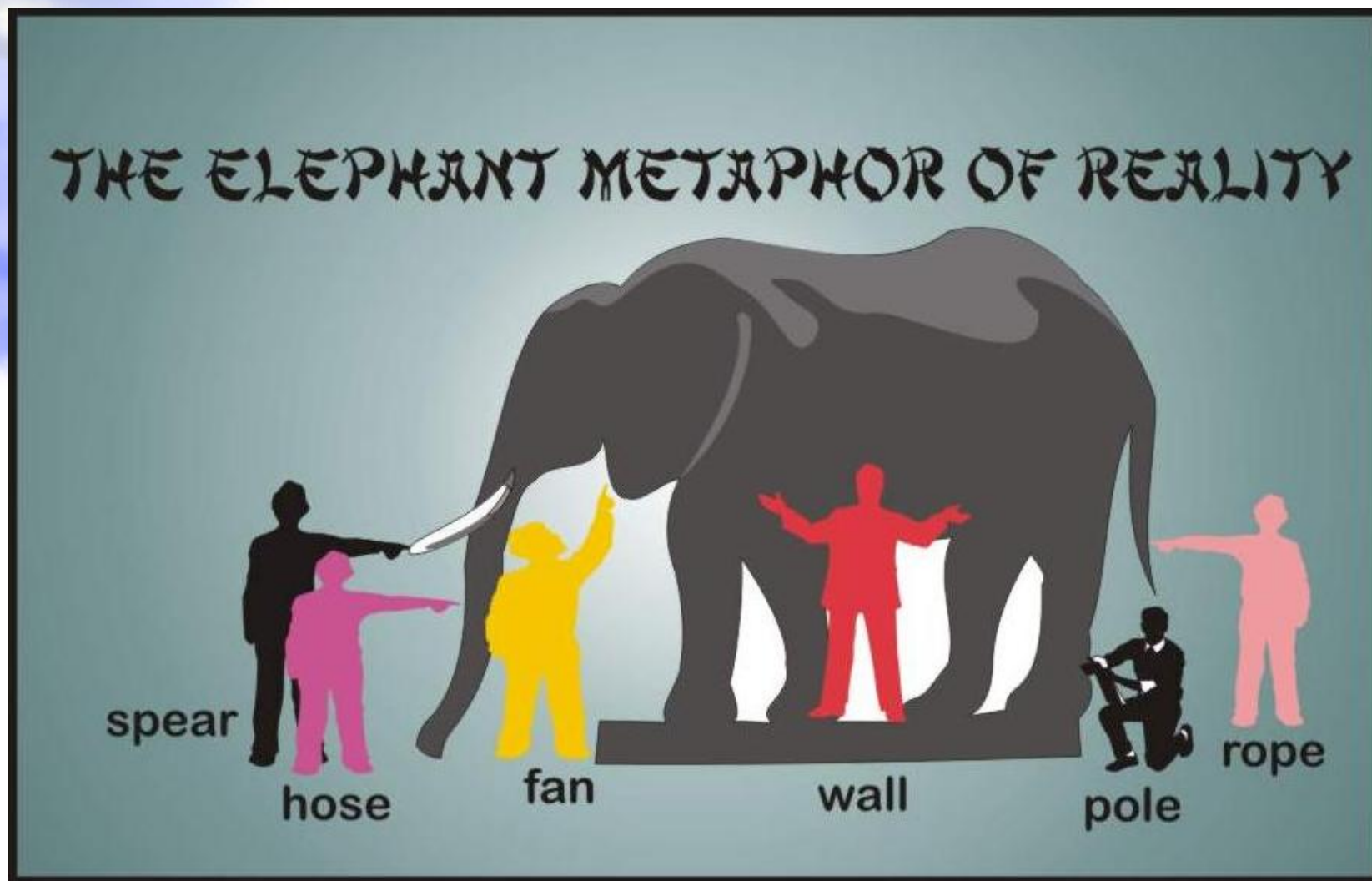


Key paradigms when dealing with data

- 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)

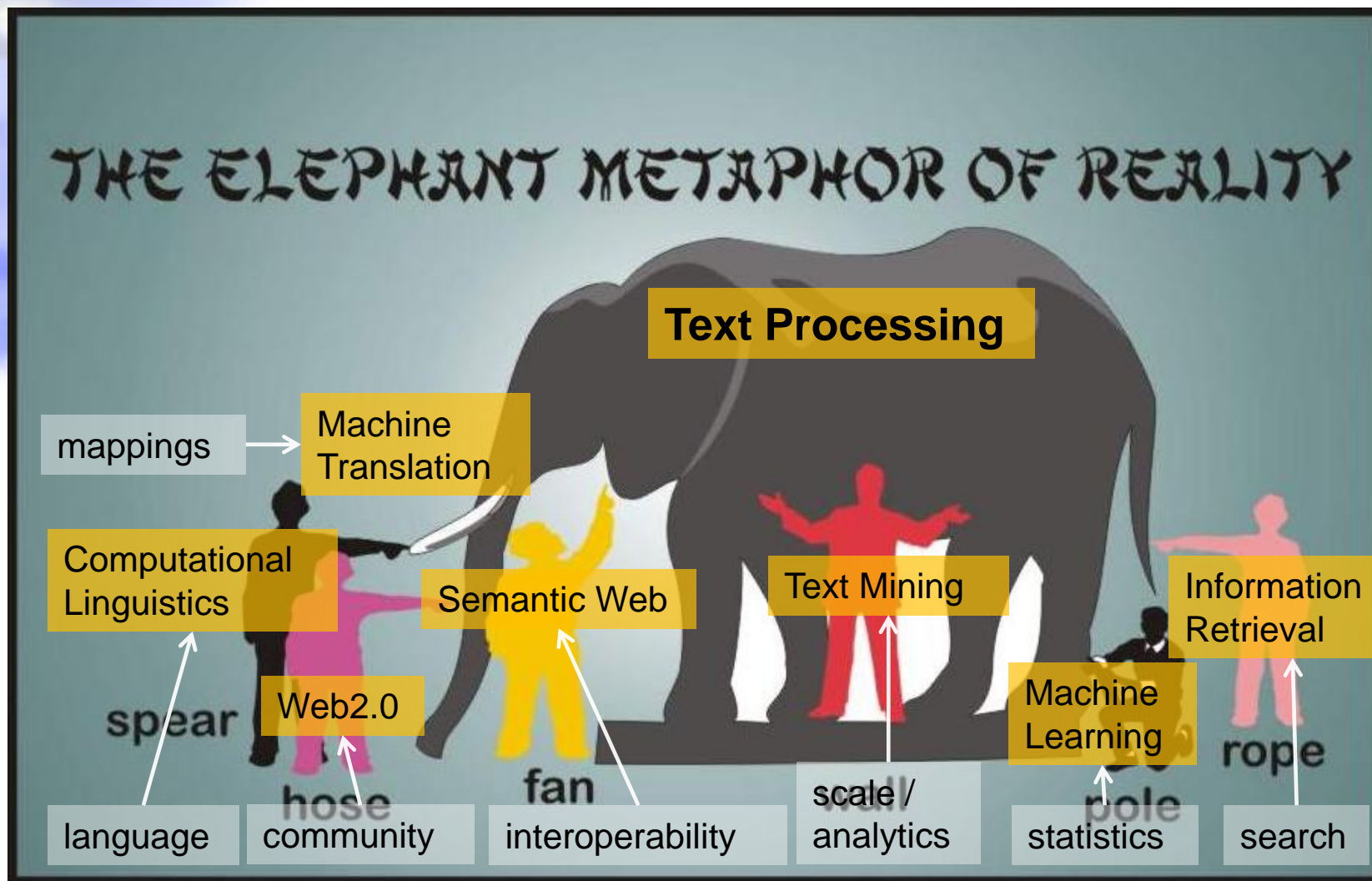


How different research areas approach text?





How different research areas approach text?





Levels of text representations

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



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)

Text categorization,
Clustering, Search,
Summarization, ...

Spam filtering, Machine translation

Syntactic

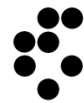
Multilingual search,
Associating text with images, ...

Data integration

Unifying semantics of data

Semantic

Reasoning,
Semantic Search





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

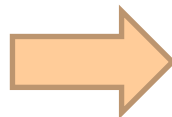
Semantic



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%



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.



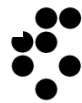
Bag-of-Words document representation

Each word is one element in a vector together with its frequency

Journal of Artificial Intelligence Research

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

0	learning
3	journal
2	intelligence
0	text
0	agent
1	internet
0	webwatcher
0	perl5
.	
.	
.	
1	volume





Similarity between BoW vectors

- 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 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)
 - ...



Boštjan Pajntar, Marko Grobelnik, Dunja Mladenić

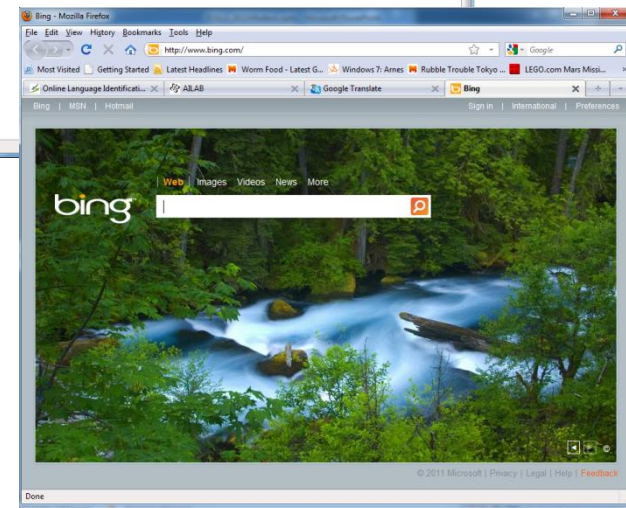
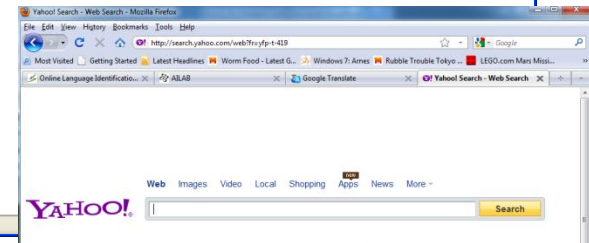
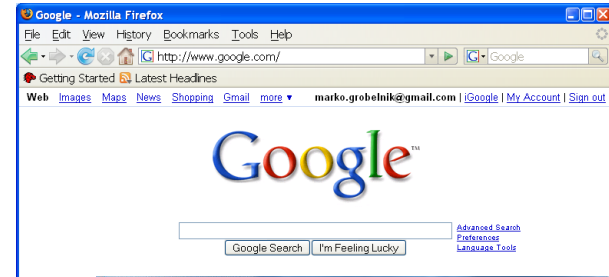
CONTEXT SENSITIVE SEARCH

<http://SearchPoint.ijs.si>



Search

- “Internet search” – one of the most common tasks involving text manipulation in everyday life
- ...but – how smart is search technology today?
 - ...not too smart!
 - It is sophisticated, but not smart



Example – Searching for “dolphin”



- Query “dolphin” has many meanings...
- ...but the first page of search engines doesn't provide us with many answers
- ...there are 131M more results

dolphin - Google Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.google.com/#sclient=psy&hl=en&source=hp&q=dolphin&pbx=1&oq=dolphin&aq=1

Most Visited Getting Started Latest Headlines Worm Food - Latest G... Windows 7: Arnes Rubble Trouble Tokyo ...

Online Language ... AILAB Google Translate dolphin - Search... AnswerArt - what... Enrycher

Web Images Videos Maps News Shopping Gmail more ▾

Google

dolphin

About 131,000,000 results (0.16 seconds) Advanced search

Everything

Images

Videos

News

Shopping

More

Any time

Past hour

Past 24 hours

Past 2 days

Past week

Past month

Past year

Custom range...

All results

Sites with images

Related searches

Dolphin - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Dolphin - Cached

Dolphins are marine mammals that are closely related to whales and porpoises. There are almost forty species of dolphin in 17 genera. ...

Bottlenose Dolphin - Marine mammal - Porpoise - Cetacean intelligence

Dolphin - Gamecube and Wii Emulator

www.dolphin-emulator.com/ - Cached

Dolphin is an open-source Nintendo GameCube and Wii Emulator for Microsoft Windows, Linux, and Mac OS X (Intel-based). It is the first emulator to ...

Dolphin Music - Guitars, Amps, Musical Instruments and Pro Audio ...

www.dolphinmusic.co.uk/ - Cached

5 hours ago – Buy all your musical instruments including, guitars, amps, and studio equipment, at **Dolphin Music** the UK's leading online musical instrument ...

Former Miami Dolphin Channing Crowder retires; Chad Henne struggles

MiamiHerald.com - 1 day ago

BY BARRY JACKSON and Jeff Darlington Former **Dolphins** linebacker Channing Crowder, 27, surprisingly announced his retirement in a radio interview Tuesday, ...

Business Insider 145 related articles

[Ex-Dolphin Channing Crowder retires](#) - ESPN

[More news for dolphin](#) »

Dolphin - BoonEx

Done



Context sensitive search

SEARCHPOINT

(1) [Dolphin - Wikipedia, the free encyclopedia](#)
 Dolphins are marine mammals that are closely related to whales and porpoises. There are almost forty species of dolphin in 17 genera. They vary in size from 1.2 m (4 ft ...
<http://en.wikipedia.org/wiki/Dolphin>

(48) [Dolphin Facts at Dolphinkind.com](#)
 Dolphin facts and information about dolphins and species of dolphins.
http://www.dolphinkind.com/dolphin_facts.html

(126) [In To Spirit -- Dolphins Hawaii -- Whale and dolphin watching ...](#)
 "Journey into Hawaiian waters and experience the dolphins and whales in their natural habitat which can be a life-changing experience."
<http://dolphinshawaii.com/>

(59) [Dolphin Essence](#)
 Dolphin Swims in Kona, Hawaii — Big Island and Humpback Whale Sightings Aloha!
 Welcome to Dolphin Essence where you can swim with dolphins in Kona, Big
<http://dolphinessence.com/>

(33) [ドルフィン\[DOLPHIN.NET\] - 横浜市立大学・関東学院 ...](#)
 Point Image
<http://ycudolphin.web.fc2.com/>

(22) [Manuel Antonio Quepos Costa Rica Planet Dolphin Catamaran Adventures](#)
 Boat tours of the offshore islands of Manuel Antonio, sea turtle, whale and dolphin watching.
<http://www.planetdolphin.com/>

(5) [Dolphin RFC](#)
 Dolphin RFC home site ... MESSAGE BOARD. Jersey Sponsorship – Full details see news item or under Tickets/Events
<http://dolphinrfc.com/>

Transferring data from searchpoint.ijs.si...

Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point



SearchPoint

"Contract"



SEARCHPOINT person Watson Ontology Search Search with topics Search with Dmoz

- (1) [xmlbase](#)
person
<http://www.daml.org/2003/06/owlConversion/examples/xmlbase>
- (2) [attribute2.owl](#)
person
<http://www.vistology.com/ont/tests/attribute2.owl>
- (3) [person1.owl](#)
person
<http://vistology.com/ont/tests/person1.owl>
- (4) [attribute3.owl](#)
person
<http://www.vistology.com/ont/tests/attribute3.owl>
- (5) [attribute1.owl](#)
person
<http://www.vistology.com/ont/tests/attribute1.owl>
- (6) [missing1.owl](#)
person
<http://www.vistology.com/ont/tests/missing1.owl>

Graphical User Interface



SABLE



METADATA
DOCUMENTS



Context
Discovery

Soft
Filtering

Graph
Drawing





Main advantages

- Generated clusters
 - (in contrast to predefined)
- User can search the whole cluster space and is not forced to select a single cluster
 - (Computer generated clusters are not necessarily what user has in mind)



News Visualization



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

News Analyser

Static Articles Static People Topics
Dynamic Articles Dynamic People

Query: clinton

Search Results

Date	Title
1997-04-09	USA: Senate probers seek Dole campaign...
1997-04-08	INDIA: Non-aligned ask for special UN me...
1997-04-08	EU: REUTER EC REPORT LONG-TER...
1997-04-08	INDIA: Egypt would mediate if Israel stalle...
1997-04-08	BELARUS: Lukashenko tells West not to...
1997-04-07	ISRAEL: Two Arabs wounded by Israeli se...
1997-04-06	USA: Netanyahu to pay US hospital visit t...
1997-04-04	USA: Democrats ask Cuban drug smuggle...
1997-04-04	UK: Ex-chief surgeon defends Gulf syndro...
1997-04-03	ISRAEL: Israel's Netanyahu to defend sett...
1997-04-03	ISRAEL: Netanyahu pledges to continue J...
1997-04-02	USA: Clinton authorizes disaster relief for T...
1997-04-01	JAPAN: PRESS DIGEST - Top Japan bus...
1997-04-01	USA: Clinton seeks FCC study on limiting li...
1997-04-01	USA: Clinton, King Hussein meet on Mide...
1997-04-01	MEXICO: Mexicans lambast new U.S. imm...
1997-03-31	USA: U.S. disapproves of Arab move on l...
1997-03-30	USA: Mrs. Clinton returns from two week A...
1997-03-30	JORDAN: Israel and Iraq set to dominate J...
1997-03-30	UK: GOLF-LEADING MONEY WINNERS ...
1997-03-27	USA: 23 million kids had gaps in health ins...
1997-03-27	UGANDA: Hillary promotes women's rights...
1997-03-27	USA: McVeigh's attorney seek delay over ...
1997-03-27	USA: U.S. cancer institute advises mamm...
1997-03-26	MOROCCO: Clinton envoy in Morocco to ...
1997-03-26	SYRIA: Egyptian premier in talks with Syria...
1997-03-26	MEXICO: Drugs lords kill "hundreds" of M...
1997-03-25	RUSSIA: Yeltsin meets Chinese, Indian le...
1997-03-25	UK: FEATURE - British spin doctors run c...
1997-03-24	USA: White House declines comment bef...
1997-03-24	BELGIUM: NATO will have no second-cla...
1997-03-23	USA: Clinton said warning to Gingrich bud...
1997-03-22	ITALY: Italy says Europe needs NATO, R...
1997-03-22	RUSSIA: Yeltsin foes blast summit outcom...
1997-03-22	OMAN: Arafat criticises U.S. Jerusalem res...

Topic Map

Selected group of news

Selected story

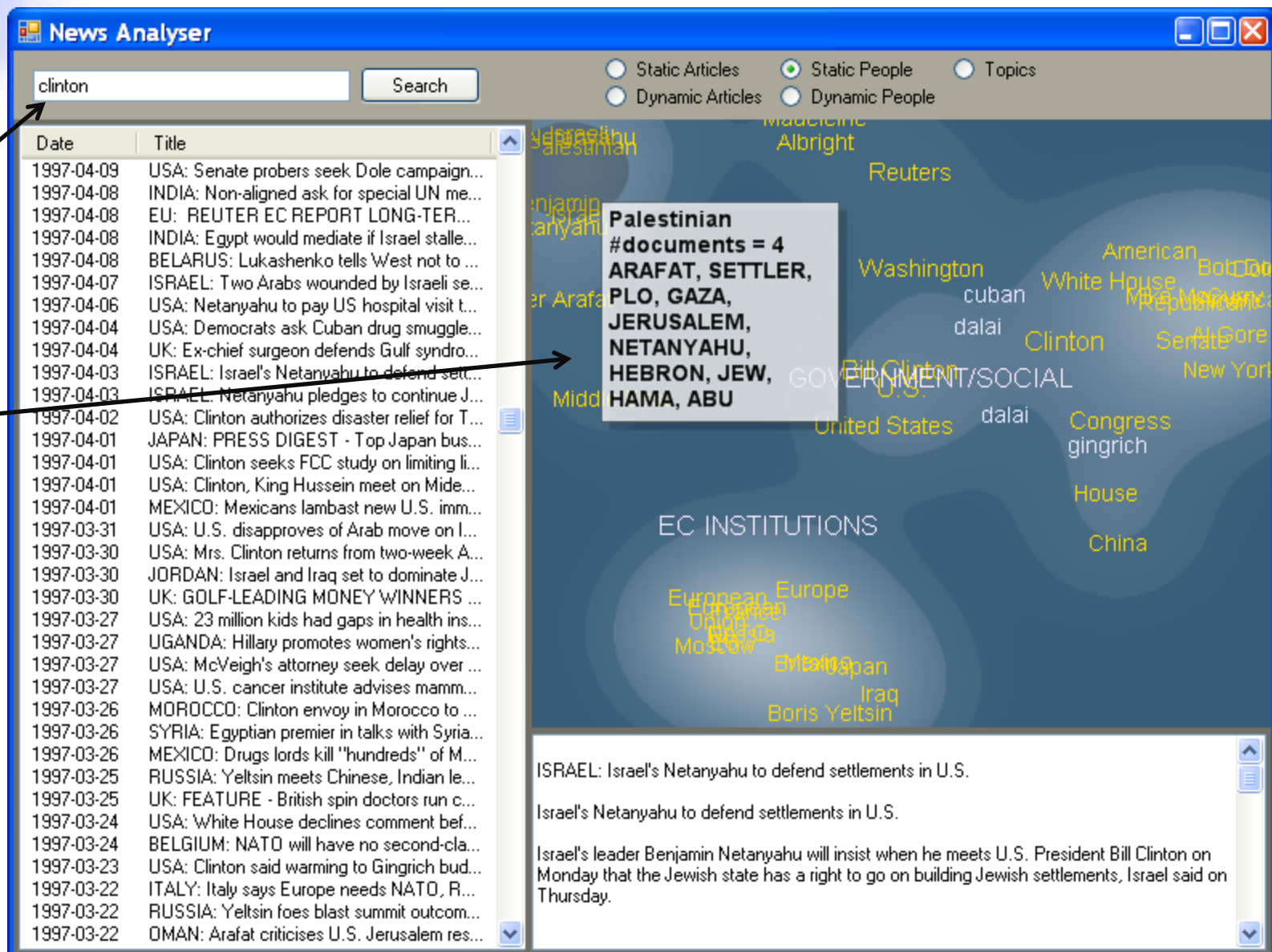
USA: U.S. will attend ...
#documents = 245
NATO, PALESTINIAN,
ISRA, PEAC, ISRAEL,
NETANYAHU,
YELTSIN, ARAFAT,
RUSSIA, SUMMIT

ISRAEL: Israel's Netanyahu to defend settlements in U.S.

Israel's Netanyahu to defend settlements in U.S.

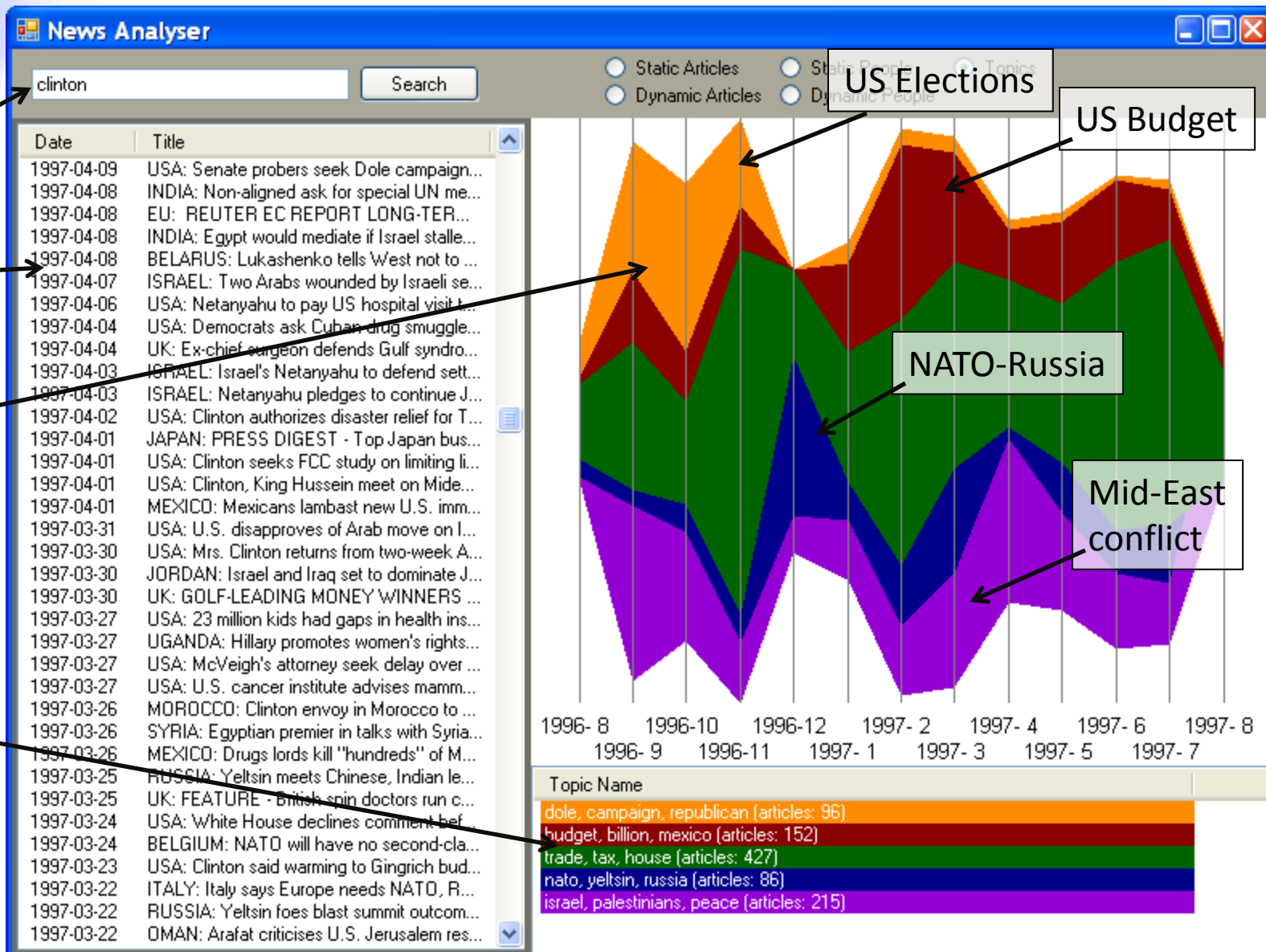
Israel's leader Benjamin Netanyahu will insist when he meets U.S. President Bill Clinton on Monday that the Jewish state has a right to go on building Jewish settlements, Israel said on Thursday.

Visualization of social relationships between "Clinton" and other entities





Topic Trends Tracking of the documents including "Clinton"



Query

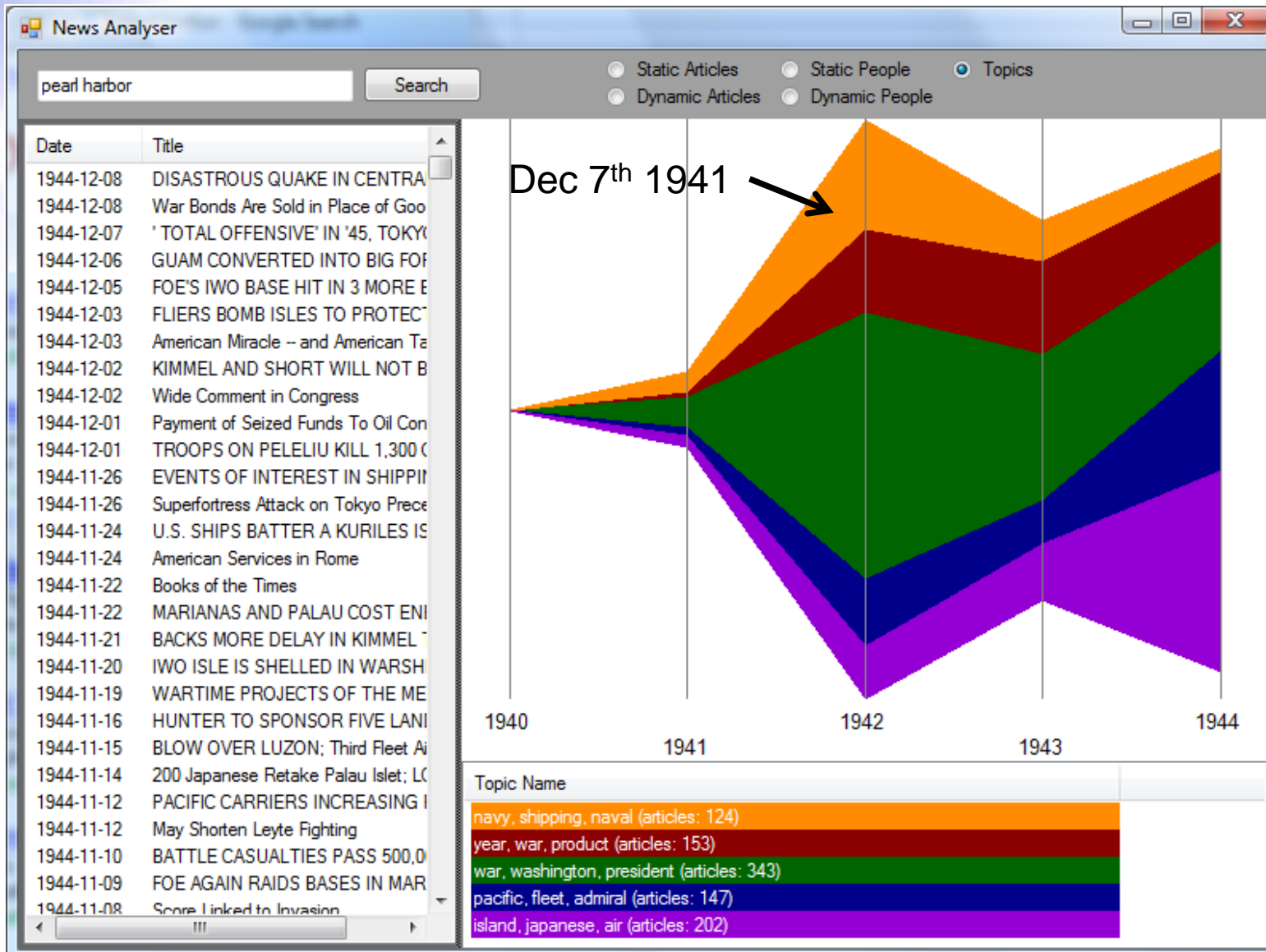
Result set

Topic Trends Visualization

Topics description

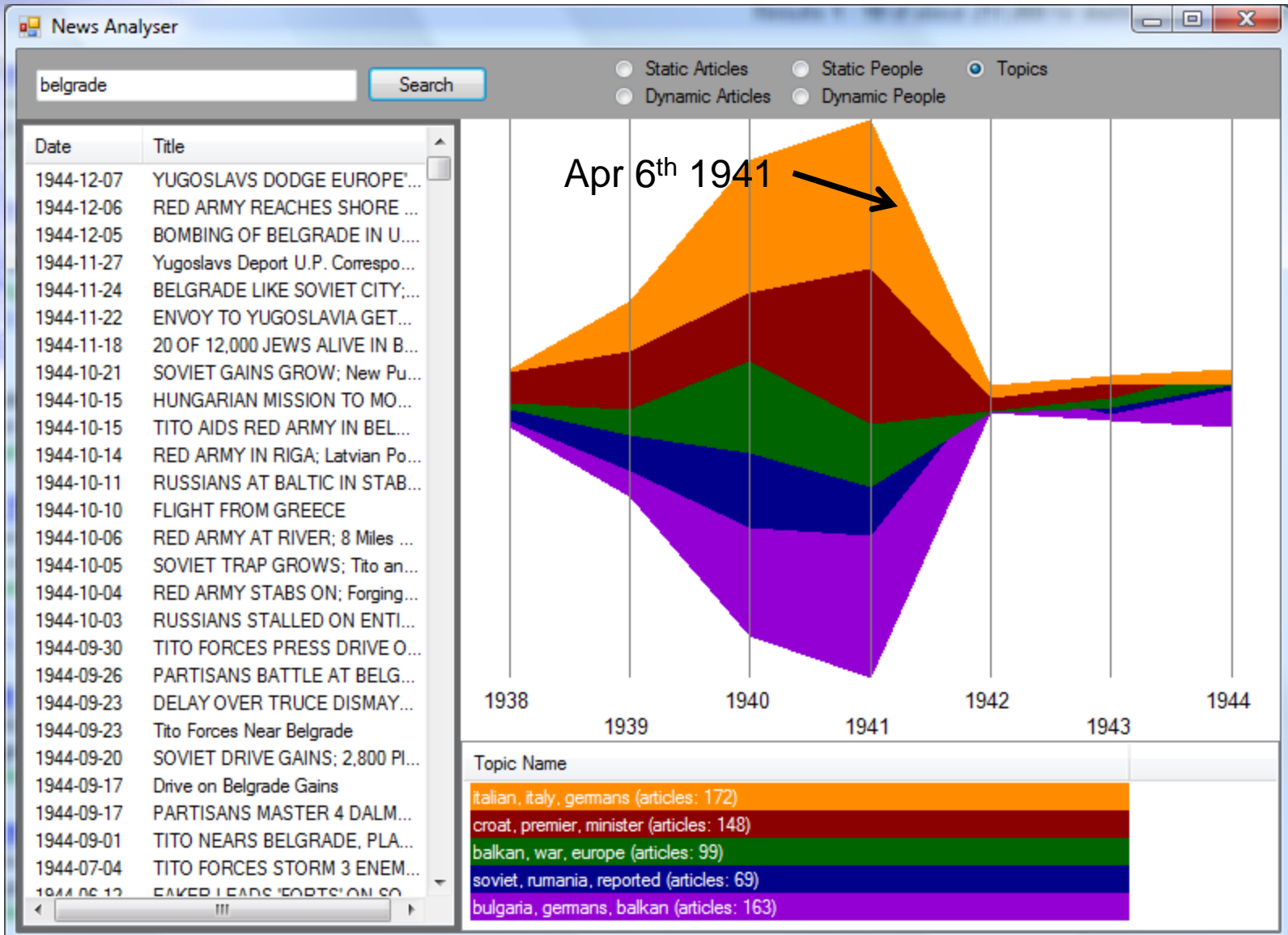


WW2 query “Pearl Harbor” into NYTimes archive



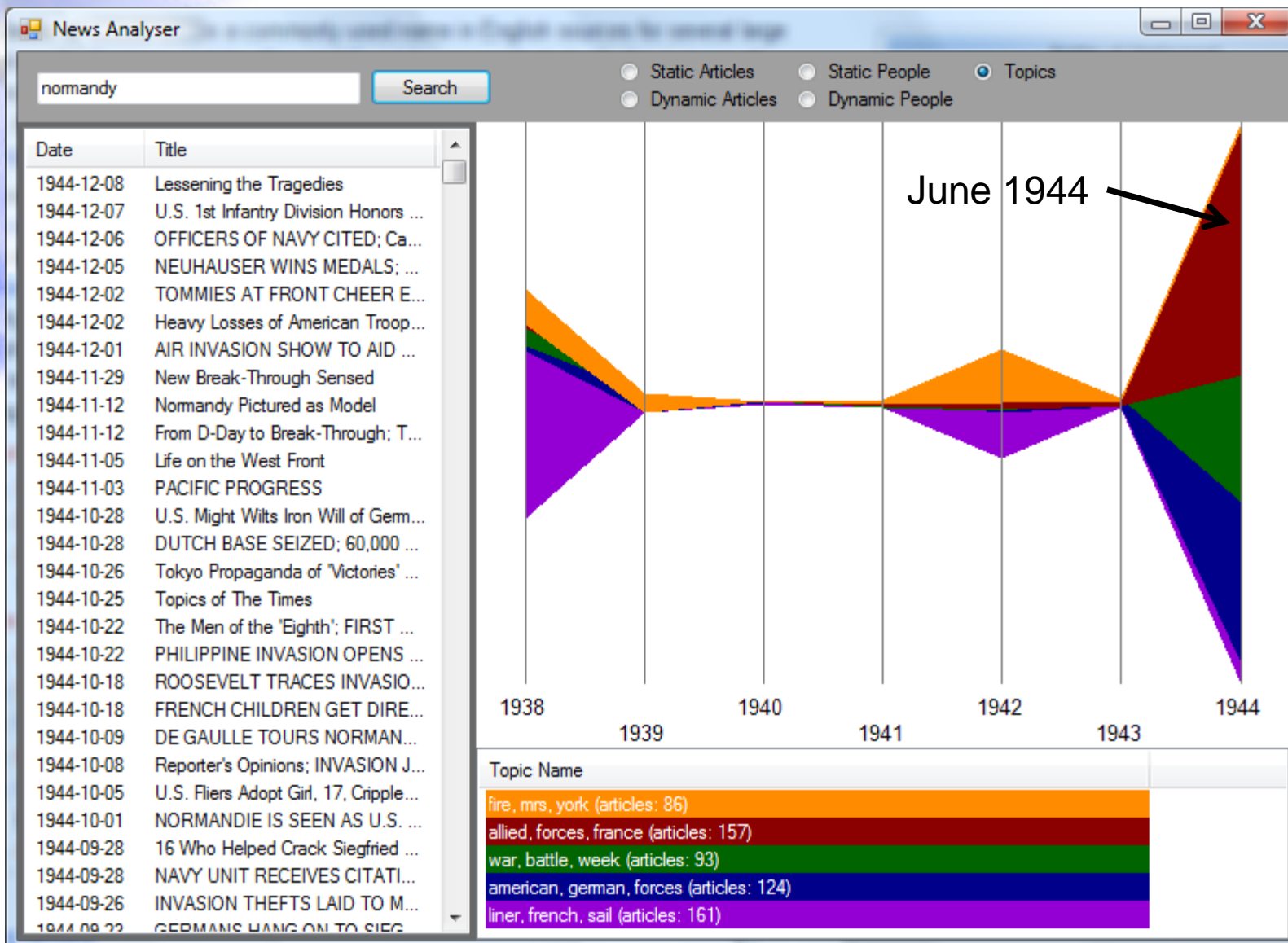


WW2 query “Belgrade” into NYTimes archive





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

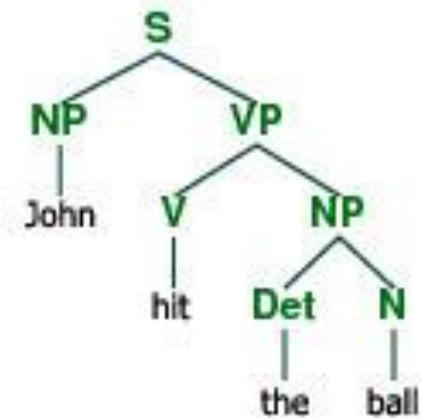
-
- Collaborative tagging / Web2.0
 - Linked Data
 - Templates / Frames
 - Ontologies / First order theories

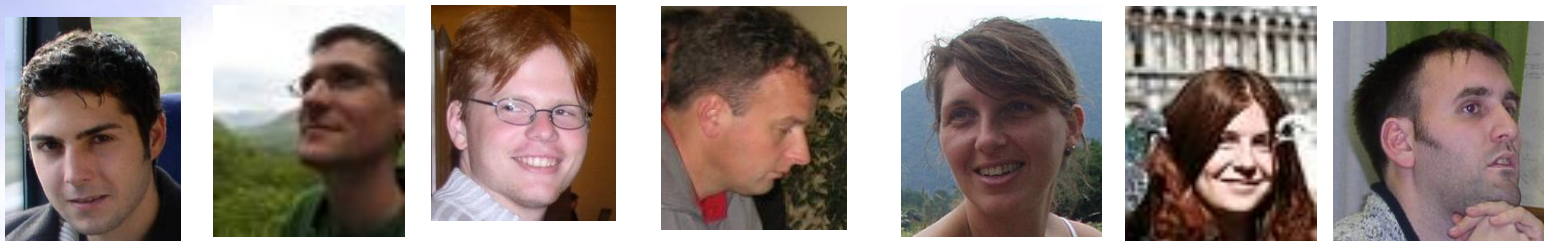
Semantic



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





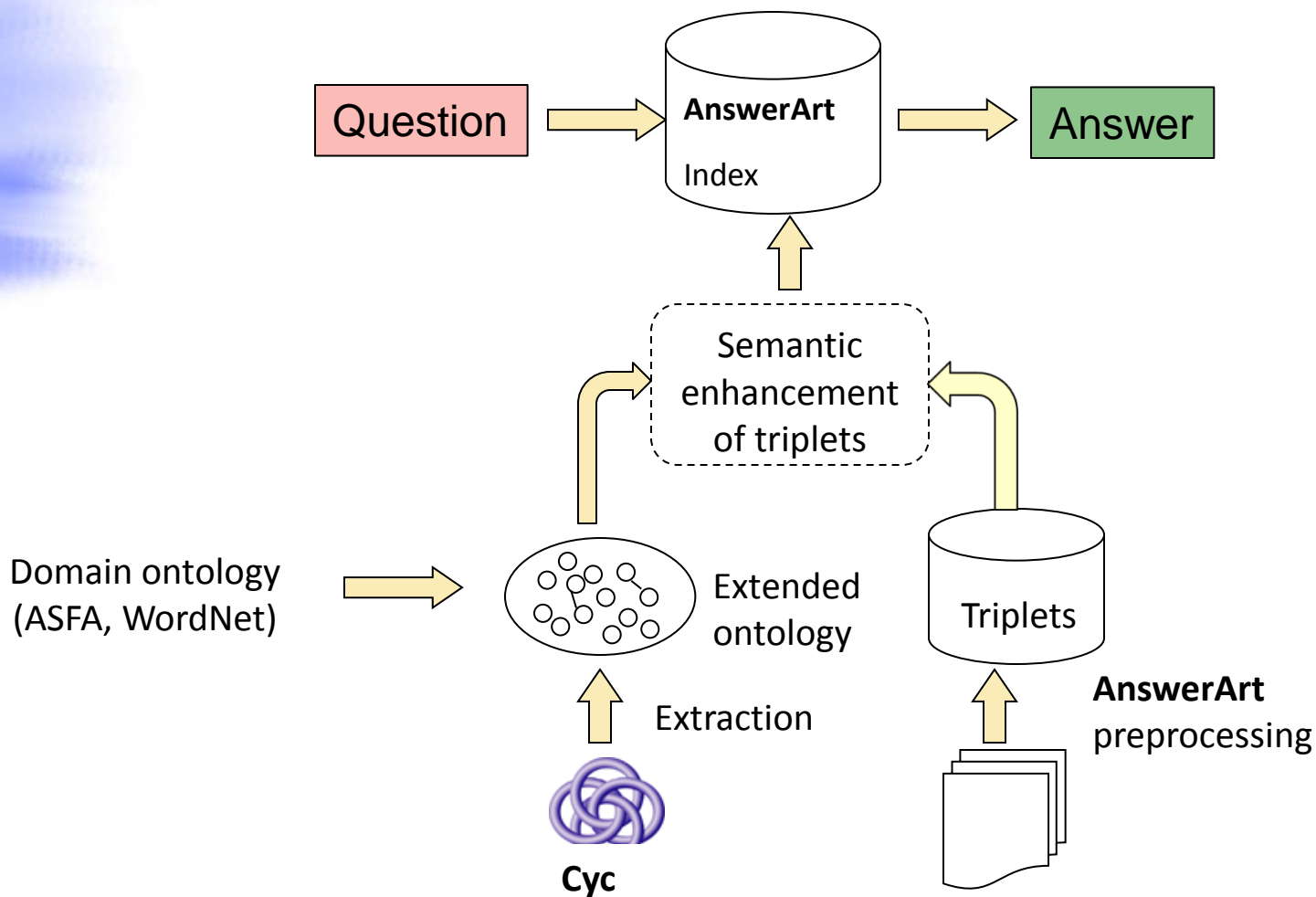
Luka Bradeško, Lorand Dali, Blaž Fortuna, Marko Grobelnik, Dunja Mladenčić, Inna Novalija, Boštjan Pajntar

ANSWER ART

<http://AnswerArt.net>





AnswerArt – System Architecture





AnswerArt using Medline



answer *Art* **NeOn**

Art of finding answers in document collection

Ask

search in:

Get the answers for your questions.

Try this:

what caused pollution
what could pollution have affected
what do sharks have
do sharks have teeth
where is water used

Broaden your knowledge with related information.

Try this:

where animals live
Australia
Australia developed
when is water used
salmon hatched

Make sense of a document contents in a second.

Try this:



AnswerArt using Medline

answer *Art*

who got injections

Show
document

Ask

We found that

the following	got	injections
rats	received	injections, injection
patients	received, receive	injection, injections
ovx rats	received	injections
monkey	received	injections
mapping technique	obtained	injection

Related documents

rats **Effects of naltrexone on the intake of ethanol and flavored solutions in All rats received injections** of naltrexone hydrochloride (10 mg/kg, i.p.) for 5 days after baseline intake measures and were monitored for a further 5 days (after-treatment phase).

rats **A serial MR study of cerebral blood flow changes and lesion development** Twenty-two **rats received** an intracerebral **injection** of ET-1 adjacent to the MCA.

rats **Triflusal posttreatment inhibits glial**



who got injections?

Ask

Show document
overview

 *Related documents*

Show Document Overview

A serial MR study of cerebral blood flow changes and lesion development

A serial MR study of cerebral blood flow changes and lesion development

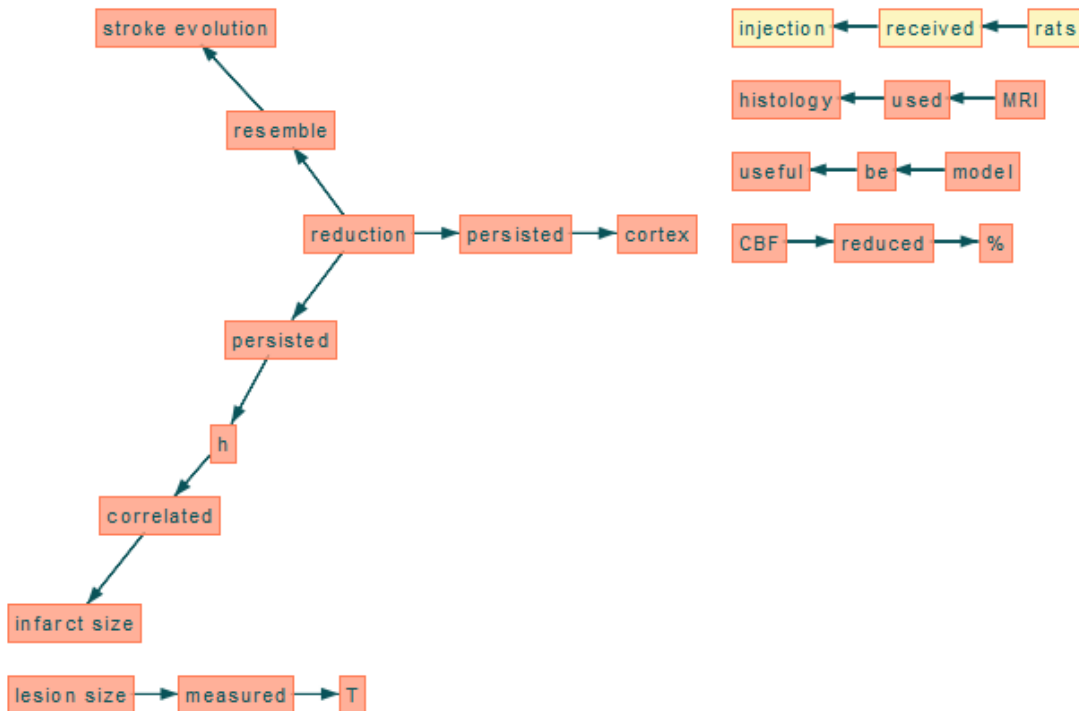
The vasoconstrictive peptide endothelin-1 (ET-1) has been used previously to transiently occlude the middle cerebral artery (MCA) in rats. However, the duration of the resulting reduction in cerebral blood flow (CBF) and the reperfusion characteristics are poorly understood. In this study perfusion and T(2)-weighted MRI were used together with histology to characterize the cerebral perfusion dynamics and lesion development following ET-1 injection. Twenty-two rats received an intracerebral injection of ET-1 adjacent to the MCA. CBF was reduced to 30-50% of control levels, and a significant reduction persisted for 16 h in the cortex and 7 h in the striatum. The lesion size measured by T(2)-weighted imaging at 48 h correlated with the final infarct size measured by histology at 7 d. The sustained reduction in CBF and the gradual development of the ischemic lesion resemble human stroke evolution, suggesting that this model may be useful for evaluating therapeutic agents, particularly when treatment is delayed.

Back





DOCUMENT OVERVIEW



FACTS

MRI used histology

rats received injection

CBF reduced %

reduction persisted h

reduction persisted cortex

lesion size measured T

h correlated infarct size

reduction resemble stroke evolution

model be useful

SUMMARY

In this study perfusion and T (2)- weighted MRI were used together with histology to characterize the cerebral perfusion dynamics and lesion development following ET- 1 injection. Twenty-two rats received an intracerebral injection of ET- 1 adjacent to the MCA. CBF was reduced to 30- 50% of control levels, and a significant reduction persisted for 16 h in the cortex and 7 h in the striatum. The lesion size measured by T (2)- weighted imaging at 48 h correlated with the final infarct size measured by histology at 7 d. The sustained reduction in CBF and the gradual development of the ischemic lesion resemble human stroke evolution, suggesting that this model may be useful for evaluating therapeutic agents, particularly when treatment is delayed.



AnswerArt using ASFA



answer Art **NeOn**

Art of finding answers in document collection

what do sharks have?

Ask

search in:
asfa

Get the answers for your questions.

Try this:

what caused pollution
what could pollution have affected
what do sharks have
do sharks have teeth
where is water used

Broaden your knowledge with related information.

Try this:

where animals live
Australia
Australia developed
when is water used
salmon hatched

Make sense of a document contents in a second.

Try this:



AnswerArt using ASFA



what do sharks have

Show document

Ask

We found that

sharks	have	the following
sharks	have	undergone declines
shark	has	tail
sharks	have	specialization
shark	has	skin
shark	has	meat
shark	has	manoeuvrability
shark	has	life history
shark	has	intention
shark	had	distribution
sharks	have	behavior patterns
shark	has	ability

Related documents

undergone declines Preliminary standardized catch rates for pelagic and large coastal sharks from logbook and observer data from the Northwest Atlantic Our results indicate that the hammerhead, white, and blue **sharks** may **have undergone declines** since 1986.

tail **Biomechanics: Hydrodynamic function of the shark's tail**
Biomechanics: Hydrodynamic function of the **shark's tail**

specialization **Steady swimming muscle dynamics in the leopard shark *Triakis semifasciata*** Thus, **sharks** such as *Triakis* may **have** no regional **specialization** in red muscle function like that seen in many teleosts, which may indicate that the evolution of differential muscle function along the body occurred after the divergence of cartilaginous and bony fishes.

skin **What is a shark doing in this pump?** The author suggests that the drag-reducing influence of longitudinal cutaneous riblets on a **shark's skin** could be adapted in pumps to increase efficiency.

meat **Shark data from Santos longliners fishery off Southern Brazil (1971-2000)** Since the beginning of this fishery, most of **shark's meat**



AnswerArt using ASFA

answer Art

DOCUMENT OVERVIEW

what do s

ring-within-a-ring vortex

Show document overview

have

contrast

branched-ring vortex

generated

angle

water flow patterns

quantify

we

fish tails

shed

rings

unclear

is

differs

lobe

has

tail

motion

FACTS

tail has lobe

differs is unclear

we quantify water flow patterns

they have ring-within-a-ring vortex structure

they have contrast

rings shed fish tails

branched-ring vortex generated angle

branched-ring vortex generated motion

Show Document

Biomech

Biom
sharl

The tail of mo
hydrodynamic
sharks and fir
vortex is gene
backwards an

Back

SUMMARY



Tadej Štajner, Delia Rusu, Lorand Dali, Blaž Fortuna,
Dunja Mladenčić, Marko Grobelnik

NATURAL LANGUAGE TEXT ENRICHMENT

<http://enrycher.ijs.si>



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

Enrycher - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://enrycher.ijs.si/

Enrycher

WHAT ARE YOU WRITING ABOUT? [home](#) [about](#) [api](#) [contact](#)

try out enrycher!

Examples:

- [Can Slovenia Win the World Cup? \(original\)](#)
- [Some Stem Cell Research Limits Lifted \(original\)](#)
- [Mexico Prepares to Lower Alert as Swine Flu Cases Ebb \(original\)](#)
- [Hotel Review: The Rough Luxe Hotel in London \(original\)](#)
- [A Walk in Calcutta \(original\)](#)
- [Bright Spot in Downturn: New Hire Is Robust \(original\)](#)

Slovenia's dramatic win over Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near domination from the likes of Brazil, Italy and Germany, international soccer is entering the era of the Cinderella. It may not happen this time around, but given the increasing flow of talent, training and information across borders, it's almost certain that a small upstart nation blessed with good athletes and better luck will make a legitimate run at the world's most coveted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Valter Birsa Wednesday.

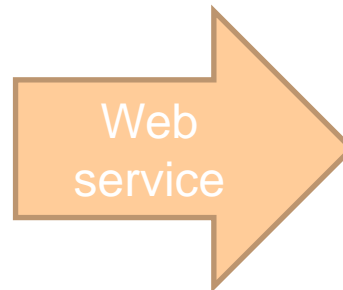
Wednesday near Paris, Ireland nearly pulled off the greatest win in its

HTML response

XML response

Enrych

Done



Enrycher result - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://enrycher.ijs.si/EnrycherFront/EnrycherServlet

Enrycher result

enrycher

THIS IS WHAT YOU'RE WRITING ABOUT: [home](#) [about](#) [api](#) [contact](#)

Network graph visualization showing relationships between entities like Michael Essien, Zlatan Ibrahimovic, and Yury Zhirkov.

interesting statements

- Slovenia has win
- Fox Soccer Channel entering era
- Slovenia has Valter Birsa Wednesday
- cross to William Gallas
- Ukraine ended scoreless tie
- Miralem Pjanic plays Ligue
- In France's 1 has Ligue
- Chebia has dielder droaga
- Fox Soccer Channel has names
- Russia has Yuri Zhirkov
- Irish exerted pressure
- Swedish referee Martin Hansson allowed score
- Ukraine has Andriy Shevchenko
- Bosnia has Edin Dzeko
- Salomon Kalou make opponent
- Internazionale has Samuel Eto'o

Done

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://enrycher.ijs.si/EnrycherFront/EnrycherServlet

http://enrycher.ijs.si/EnrycherServlet

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0" encoding="UTF-8" ?>
<enrych>
  <enrych>
    <enrych type="text/html" enrych="Topic on Soccer" ?>
      <enrych type="text/html" enrych="Topic on Soccer" ?>
        <enrych type="text/html" enrych="Topic on Soccer" ?>
          <enrych type="text/html" enrych="Topic on Soccer" ?>
            <enrych type="text/html" enrych="Topic on Soccer" ?>
              <enrych type="text/html" enrych="Topic on Soccer" ?>
                <enrych type="text/html" enrych="Topic on Soccer" ?>
                  <enrych type="text/html" enrych="Topic on Soccer" ?>
                    <enrych type="text/html" enrych="Topic on Soccer" ?>
                      <enrych type="text/html" enrych="Topic on Soccer" ?>
                        <enrych type="text/html" enrych="Topic on Soccer" ?>
                          <enrych type="text/html" enrych="Topic on Soccer" ?>
                            <enrych type="text/html" enrych="Topic on Soccer" ?>
                              <enrych type="text/html" enrych="Topic on Soccer" ?>
                                <enrych type="text/html" enrych="Topic on Soccer" ?>
                                  <enrych type="text/html" enrych="Topic on Soccer" ?>
                                    <enrych type="text/html" enrych="Topic on Soccer" ?>
                                      <enrych type="text/html" enrych="Topic on Soccer" ?>
                                        <enrych type="text/html" enrych="Topic on Soccer" ?>
                                          <enrych type="text/html" enrych="Topic on Soccer" ?>
                                            <enrych type="text/html" enrych="Topic on Soccer" ?>
                                              <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                  <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                    <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                      <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                        <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                          <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                            <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                              <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                  <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                    <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                      <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                        <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                          <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                            <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                              <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                  <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                    <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                      <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                        <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                          <enrych type="text/html" enrych="Topic on Soccer" ?>
                                                                                           ...

```

Done



Enrycher Service

Annotation Features:

Entity extraction

- People, locations, organizations, dates, percentages and money amounts

Entity resolution

- co-reference
- anaphora

Entity linkage to Linked Open Data (LOD)

Word Sense Disambiguation to LOD (WordNet 3.0 VUA)

Assertion extraction

- Subject – predicate – object sentence elements together with their modifiers

Categories – from the Open Directory and the Wikipedia category schema

The screenshot shows the Enrycher web interface. At the top, the logo 'enrycher' is displayed with the tagline 'THIS IS WHAT YOU'RE WRITING ABOUT:' and navigation links for 'home', 'about', 'api', and 'contact'. Below the header, there is a 'Show semantic graph' button. The main content area is titled 'interesting statements' and is divided into three columns: 'text', 'entities', and 'keywords'. The 'text' column contains a paragraph about children playing video games, with several words highlighted in green. The 'entities' column lists these highlighted words as links to external resources. The 'keywords' column lists a set of categories, also with links. The text in the screenshot is as follows:

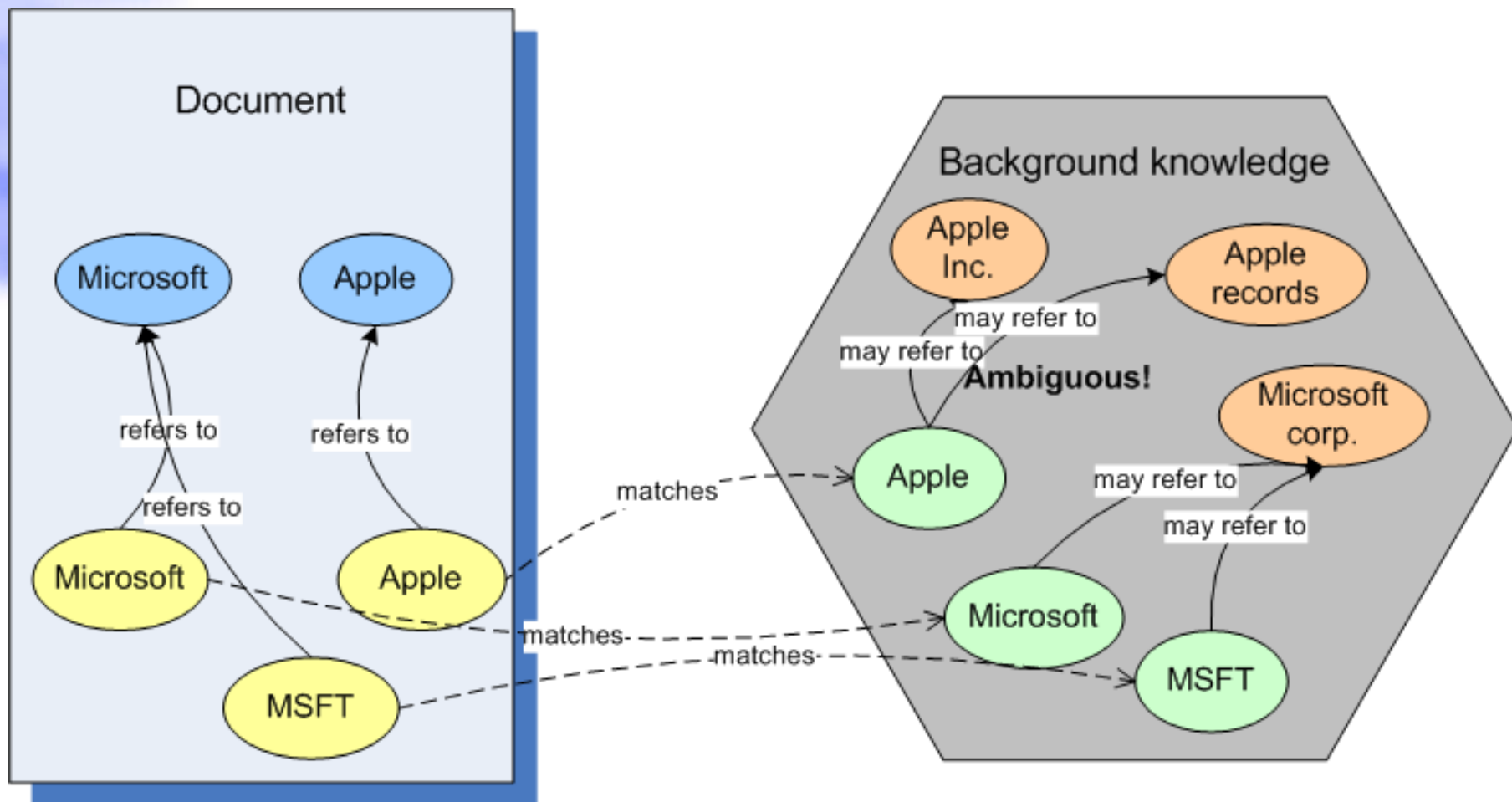
Show semantic graph

interesting statements

text	entities	keywords
About 9 percent of children play such long hours of video games that they are pathological gamers, increasing risks of anxiety, depression, bad grades and social phobia, a study in Singapore found.	<ul style="list-style-type: none">• Singapore• Pediatrics• Ames• Douglas Gentile	Health, Mental Health, Disorders, Mood, Games, Video Games, Depression, Recreation, Browser Based, Game Studies, Anxiety, Women, Society, Recreation and Sports,
The compulsive gamers played for a weekly average of 31 hours compared with 19 for kids not deemed pathological, according to research released today by the journal Pediatrics .	<ul style="list-style-type: none">• Iowa State University• The compulsive	categories <ul style="list-style-type: none">• Top/Health/Mental Health/Disorders/Mood/Depression• Top/Health/Mental Health/Disorders/Mood• Top/Games/Game Studies• Top/Health/Mental Health/Disorders/Anxiety• Top/Society/People/Women/Recreation_and_Sports• Top/Health/Medicine/Medical_Specialties/Pathology• Top/Games/Video Games/Recreation/Browser_Based• Top/Games/Video Games/History• Top/Games/Video Games/Recreation/Browser_Based/Board Games• Top/Health/Mental Health/Disorders
Overall, 83 percent of 3,034 children in the study played video games at least occasionally.		
Gamers are considered pathological when their playing interferes with everyday life, and their behavior is described as being similar to that of gambling addicts, according to background information in the paper.		
The gaming isn't merely a symptom of disorders such as depression, anxiety and social phobia, today's study found.		
Rather, gaming can cause and reinforce those maladies.		
Although children who are depressed may retreat into gaming, the gaming increases the depression, wrote the study authors, led by Douglas A. Gentile , a psychologist at Iowa State University , in Ames .		
The study, of children in grades 3, 4, 7 and 8, lasted two years.		
Kids who stopped being pathological gamers during the study period showed lower levels of depression, anxiety and social phobia compared with peers who didn't stop, the researchers said.		

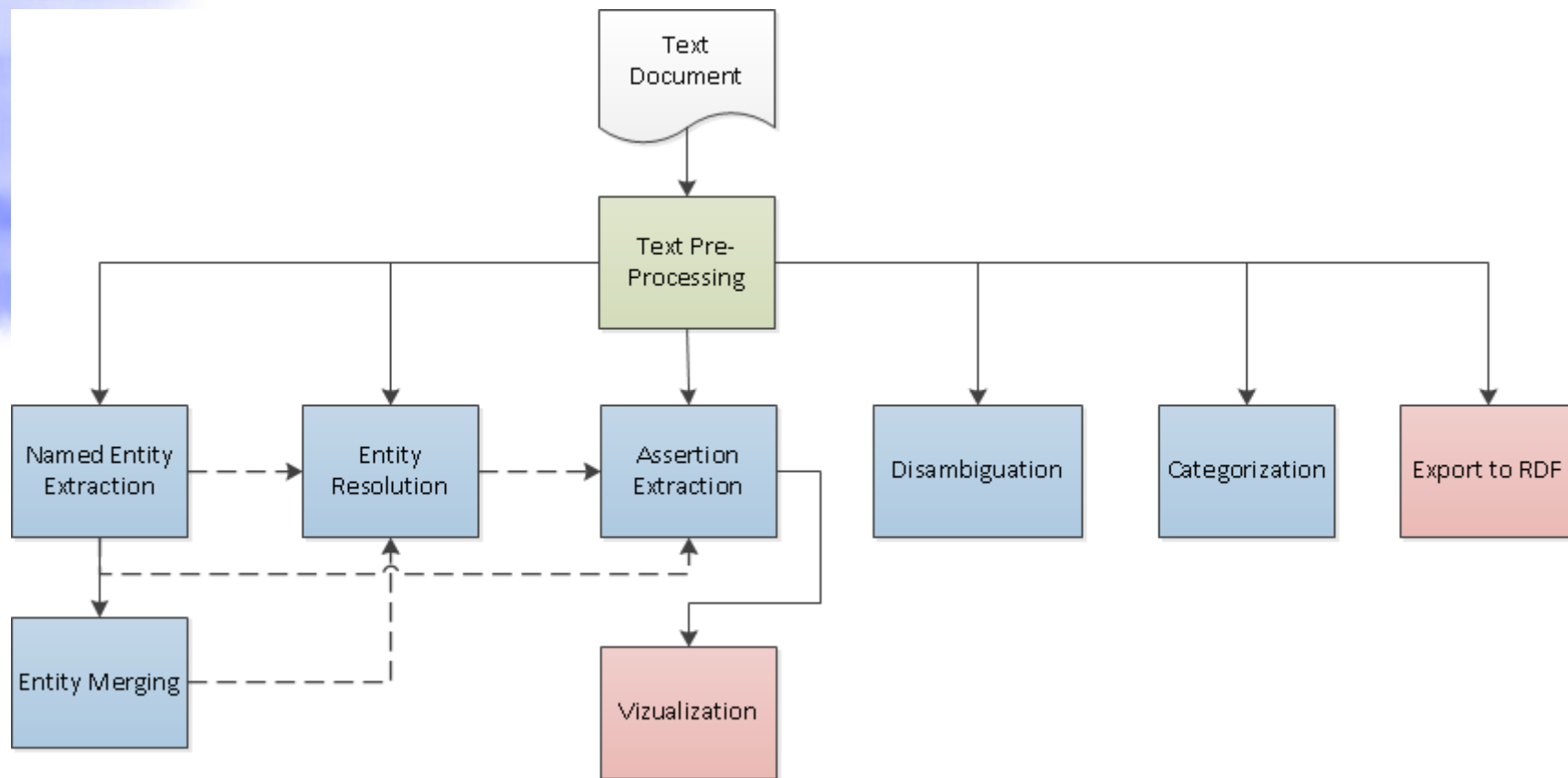


Entity resolution in text





Enrycher Service Dependencies



The dashed line marks dependencies between components that are optional, whereas the filled lines mark required dependencies



Enrycher - demo

enrycher

WHAT ARE YOU WRITING ABOUT?

[home](#) [about](#) [api](#) [contact](#)

try out enrycher!

Examples:

- [Can Slovenia Win the World Cup? \(original\)](#)
- [Some Stem Cell Research Limits Lifted \(original\)](#)
- [Mexico Prepares to Lower Alert as Swine Flu Cases Ebb \(original\)](#)
- [Hotel Review: The Rough Luxe Hotel in London \(original\)](#)
- [A Walk in Calcutta \(original\)](#)
- [Bright Spot in Downturn: New Hiring Is Robust \(original\)](#)

Dunja Mladenec is an expert on study and development of Machine Learning, Data/Text Mining, Semantic Technology techniques and their application on real-world problems. She is associated with the J. Stefan Institute since 1987 and she is currently leading the Artificial Intelligence Laboratory at the Institute. She got her MSc and PhD in Computer Science at University of Ljubljana in 1995 and 1998 respectively. Most of her research work is connected with the study and development of machine learning techniques and their application on real-world problems from



Enrycher - demo

Dunja Mladenic – Homepage

Dunja's Personal Web Site @ AiLab



RESEARCH BOOKS TUTORIALS COURSES VIDEOGRAPHY CONTACT DATA

Biographical Sketch

Dunja Mladenic is an expert on study and development of Machine Learning, Data/Text Mining, Semantic Technology techniques and their application on real-world problems. She is associated with the J. Stefan Institute since 1987 and she is currently leading the Artificial Intelligence Laboratory at the Institute. She got her [MSc](#) and [PhD](#) in Computer Science at University of Ljubljana in 1995 and 1998 respectively. Most of her research work is connected with the study and development of machine learning techniques and their application on real-world problems from different areas e.g., medicine, pharmacology, manufacturing, economy. Her current research focuses on using machine learning in data analysis, with particular interest in learning from text applied on the Web documents and intelligent agents. She was visiting School of Computer Science, [Carnegie Mellon University](#), Pittsburgh, PA, USA, as a visiting researcher in 1996-1997 and as a visiting faculty in 2000-2001 .



Dunja Mladenic is the Slovenian member of the Enwise Expert Group Promoting women scientists from the Central and Eastern European countries and the Baltic States to produce gender equality in science in the wider Europe. She was coordinating European Research and Development project "Data Mining and Decision Support for business competitiveness: A European virtual enterprise ([Sol-Eu-Net](#)) (2000-2003), involving 12 partners from 7 countries ([project homepage](#)). She is on the Management Board of several EU project including 5FP NoE project KNet – The European



Enrycher - demo

entities

text

entities

openCyc

keywords

Dunja Mladenic

• J. Stefan Institute

Computers, Artificial Intelligence, Conferences,

OpenCyc (Current): [<http://sw.opencyc.org/concept/Mx4rvVj9VJwpEbGdrcN5Y29ycA>]

OpenCyc (Versioned): [<http://sw.opencyc.org/2009/04/07/concept/Mx4rvVj9VJwpEbGdrcN5Y29ycA>]

Search



OpenCyc Individual: Ljublana

Unique ID: [[Mx4rvVj9VJwpEbGdrcN5Y29ycA](http://sw.opencyc.org/concept/Mx4rvVj9VJwpEbGdrcN5Y29ycA)]

English ID: [[CityOfLjubljanaSlovenia](#)]

English Aliases: ["Ljubljana", "Ljubljana, Slovenia"]

The capitalCity of Slovenia.

Instance of: [capital city](#)

Same as:

Related to (broader): [Earth](#), [Earth](#), [Eurasia](#), [Eurasia](#), [Europe](#), [Europe](#), [the Northern Hemisphere](#), [the Northern Hemisphere](#), [the Occident](#), [the Occident](#), [the Old World](#), [the Old World](#)

Related to (narrower):



Enrycher - demo

categories

interesting statements

[Dunja Mladenic](#) is [Enwise Expert Group](#)
[interest](#) applied [Web documents](#)

[d](#) [m](#) [o](#) [z](#) open directory project

In partn
Aol S

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [update listing](#) | [report abuse/sp](#)

text

entit

Search

the entire directory

[Dunja Mladenic](#) is an [expert](#) on
study and development of
Machine Learning, Data/[Text](#)
Mining, [Semantic Technology](#)
techniques and their
application on real-world
problems.
She is associated with the J.
Stefan Institute since 1987
and she is currently leading
the [Artificial Intelligence](#)
[Laboratory](#) at the Institute.
She got her MSc and PhD in

- [J. S.](#)
- [Inst](#)
- [Arti](#)
- [Inte](#)
- [Lab](#)
- [Ljul](#)
- [Car](#)
- [Mel](#)
- [Uni](#)
- [Pitt](#)
- [USA](#)
- [Enw](#)
- [Gro](#)
- [East](#)

[Top: Computers: Artificial Intelligence: Machine Learning](#) (246)

- [Belief Networks@](#) (48)
- [Case-Based Reasoning](#) (11)
- [Neural Networks@](#) (234)

- [Companies](#) (7)
- [Conferences](#) (28)
- [Datasets](#) (18)
- [Educational Resources](#) (3)
- [Mailing Lists](#) (3)
- [Publications](#) (18)
- [Research Groups](#) (39)
- [Software](#) (105)
- [Support Vector Machines@](#) (23)

See also:

- [Computers: Software: Databases: Data Mining](#) (166)
- [Science: Math: Statistics](#) (809)

This category in other languages:

[Japanese](#) (5)



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



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)
 - ...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



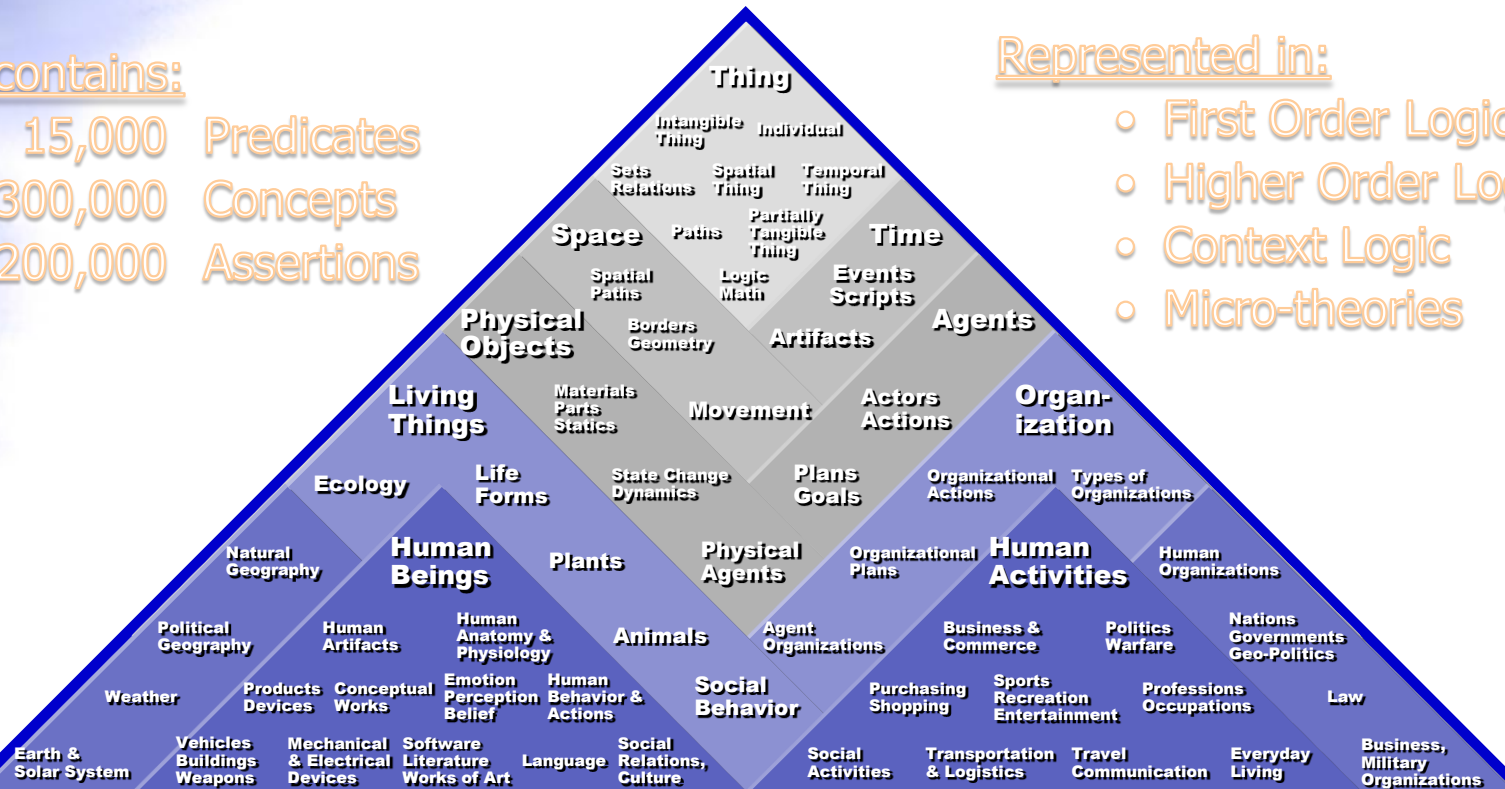
The Cyc Ontology

Cyc contains:

15,000 Predicates
300,000 Concepts
3,200,000 Assertions

Represented in:

- First Order Logic
- Higher Order Logic
- Context Logic
- Micro-theories

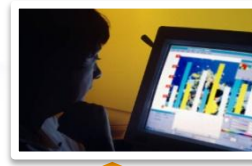


General Knowledge about Various Domains

Specific data, facts, and observations



Knowledge Users



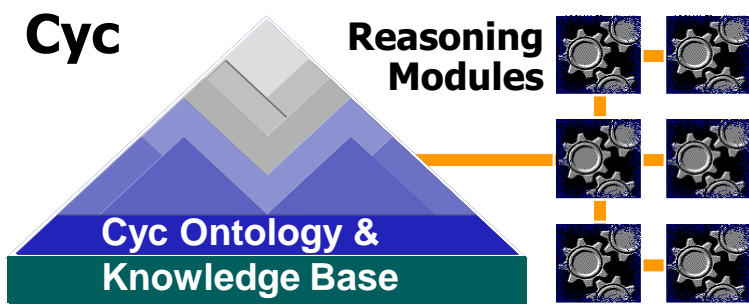
Knowledge Authors



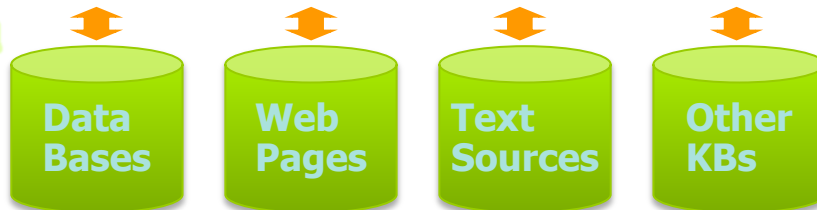
Knowledge Entry Tools

User Interface (with Natural Language Dialog)

Other Applications



External Data Sources



Cyc High-level Architecture



Cycorp © 2006

ailab.ijs.si





Cyc KB Extended w/Domain Knowledge

Thing

Intangible
Thing Individual

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))

Solar System

Buildings
Weapons

& Electrical
Devices

Literature
Works of Art

Language
Relations,
Culture

Social
Activities

Transportation
& Logistics

Travel
Communication

Energy
Living

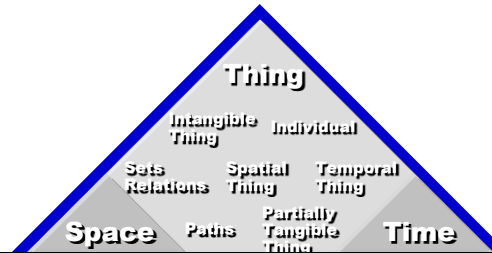
Military
Organizations

General Knowledge about Terrorism

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



Cyc KB Extended w/Domain Knowledge



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**

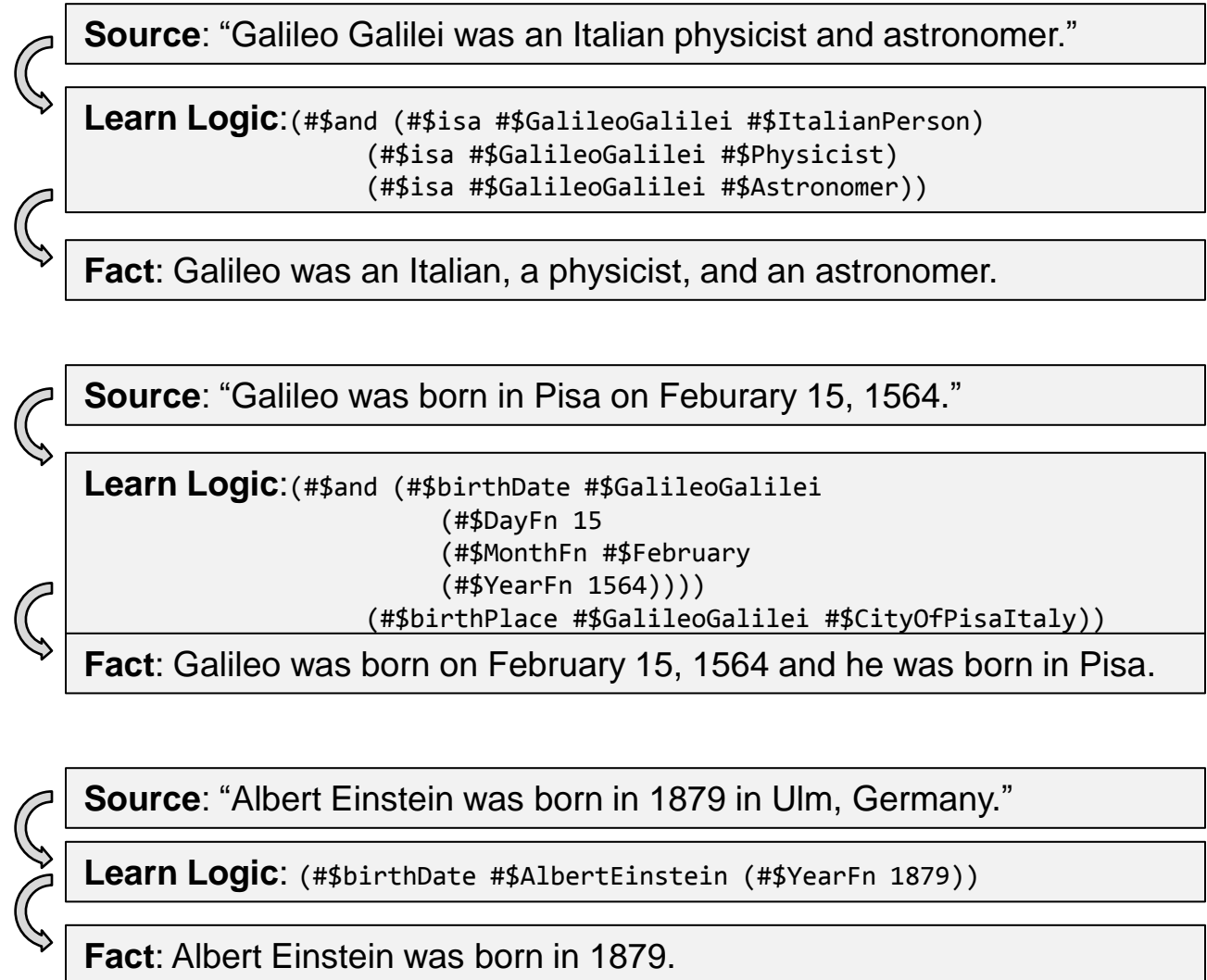


Example of automatic translating text into Cyc Logic

Translation happens with text-to-logic translation rules where CycKB is used as a background knowledge

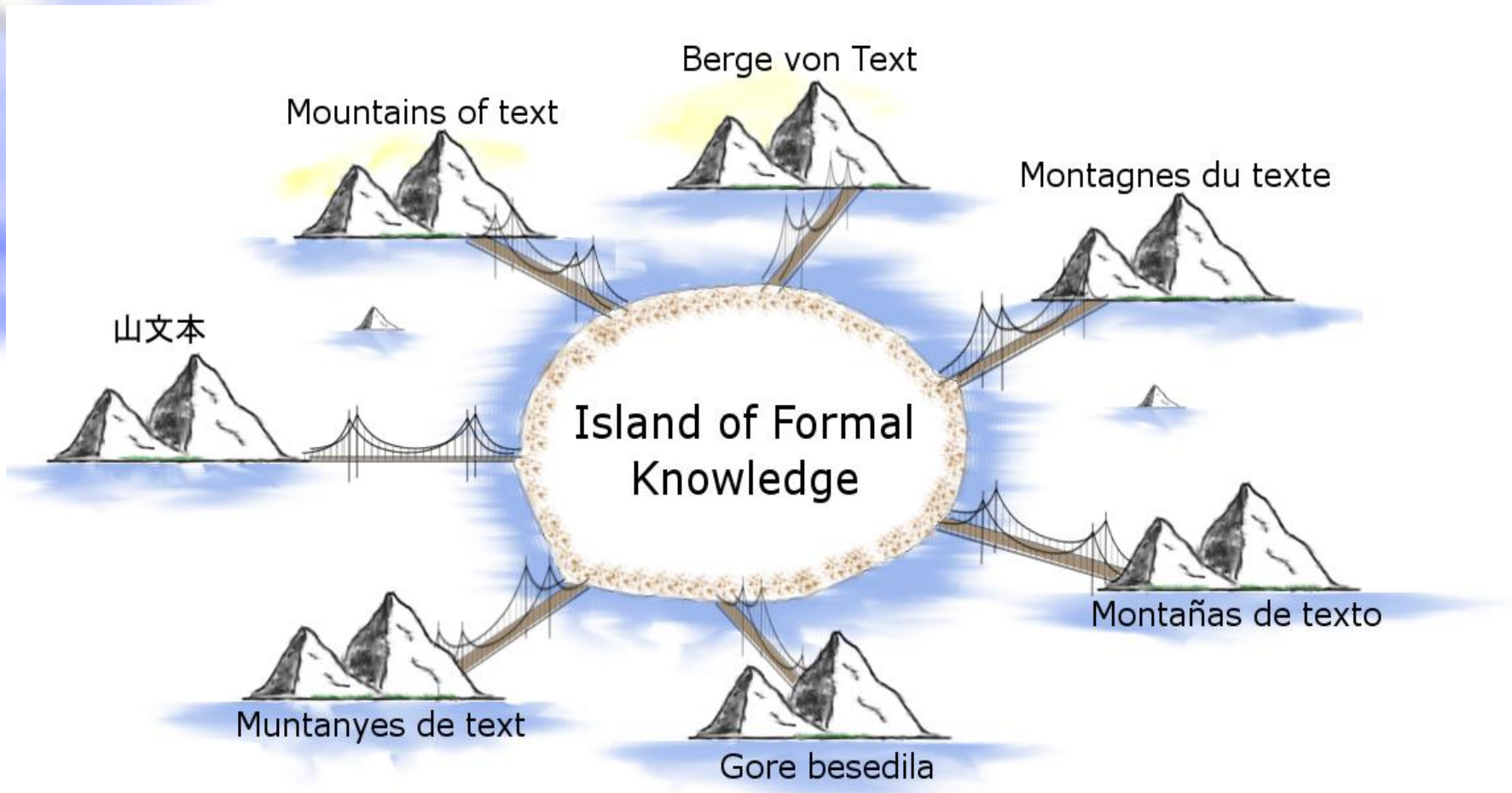
The goal is to automatize generation of translation rules for multiple languages

- ...the key is to reduce cost of creating translation rules to minimum





Representing knowledge in logic (X-Like project)

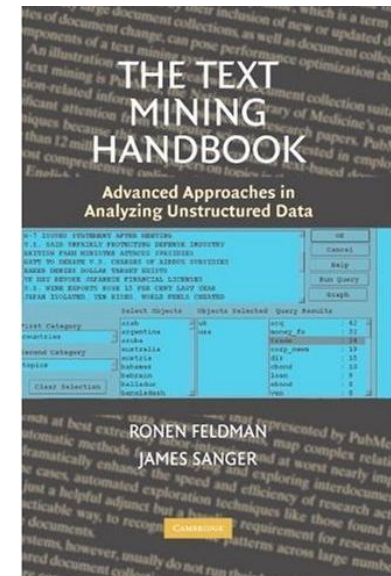
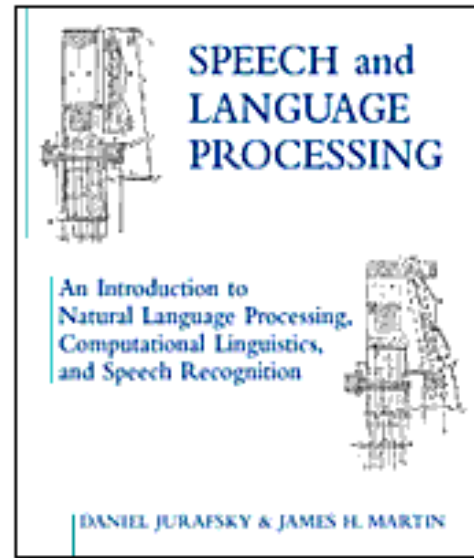
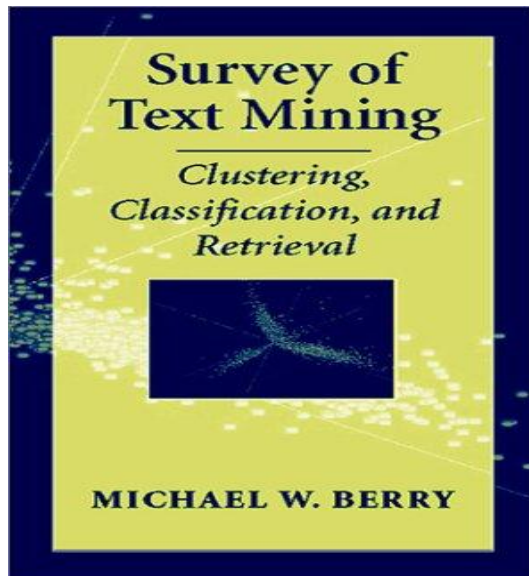
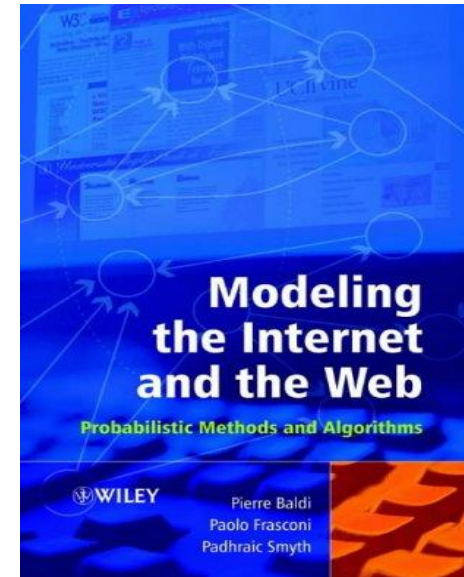
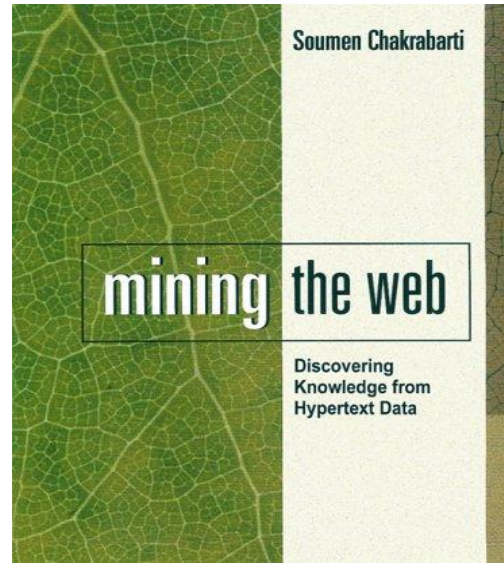
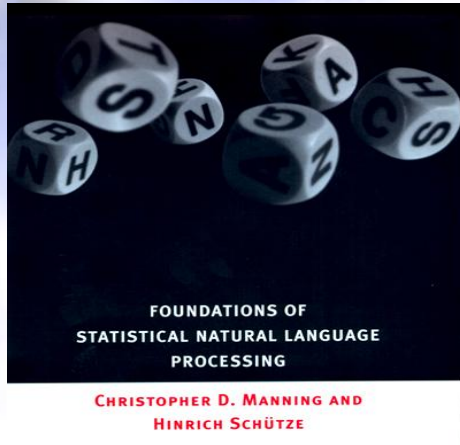




Further references...

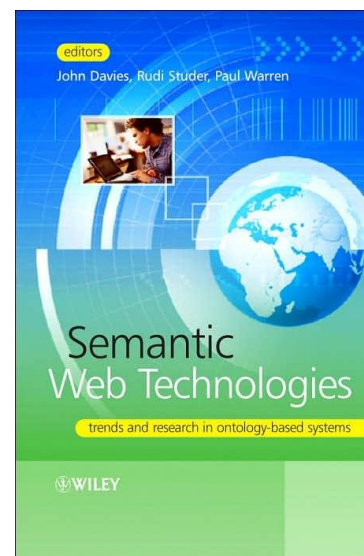
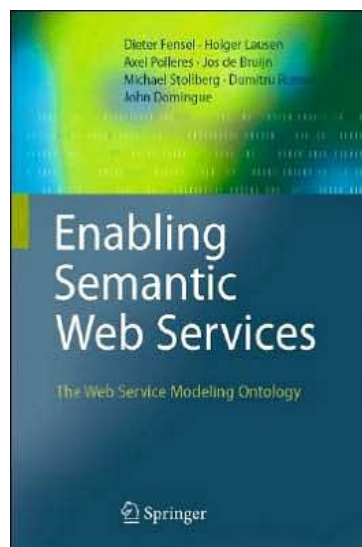
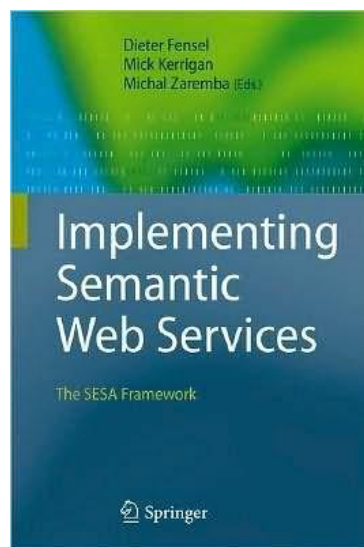
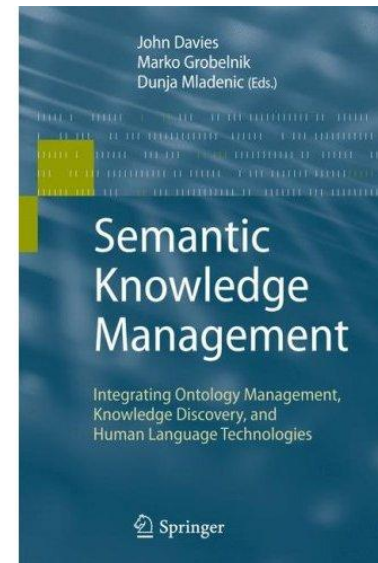
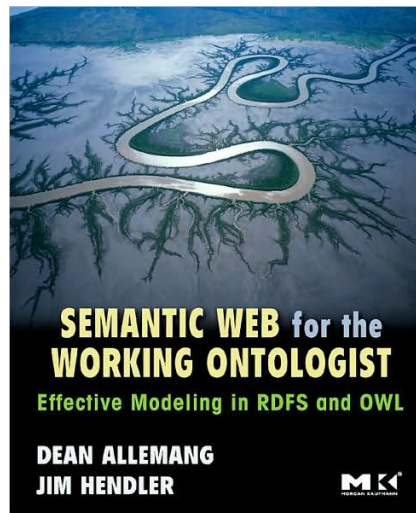
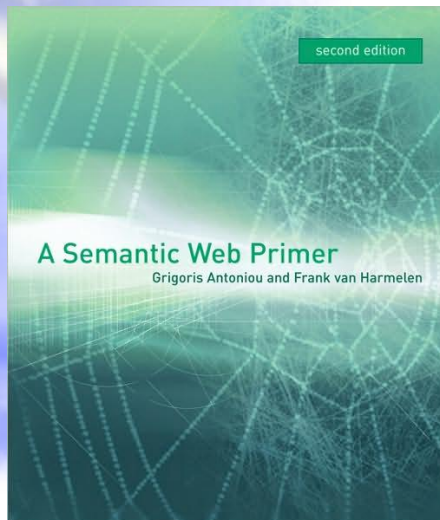


References to some Text-Mining books





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



Videolectures.net

622 events, 8910 authors, 11431 lectures,
13689 videos

videolectures.net
exchange ideas & share knowledge



SIGN IN

NEW USER
REGISTER

622 events, 8910 authors, 11431 lectures, 13689 videos

HOME • BROWSE LECTURES • PEOPLE • CONFERENCES • ACADEMIC ORGANISATIONS • EU SUPPORTED • ABOUT US • BLOG

SEARCH >>

Featured Lectures

MORE



Early Signs of Financial Crises
Siew Ann Cheong



Probabilities and Language Models
Jason Eisner



invited talk
Universal Access to Human Knowledge (Or Public Access to Digital ...
Brewster Kahle



Lecture 17: Entropy and disorder
Elizabeth Vogel Taylor



Effective teaching for disabled students
Alan Hurst



Computer Science and Human-Computer Interaction
Robert Jacob

CATEGORIES

- Agriculture (2)
- Architecture (83)
- Arts (151)
- Astronomy (39)
- Biology (198)
- Business (359)
- Chemistry (147)
- Computers (211)

NEWS

Interview with Tom Mitchell - head of ML@CMU July 21, 2011

We had a brief interview on our official Blog with Tom Mitchell, Chair of the Machine Learning Department at Carnegie Mellon University, who is one of the most prominent scholars in the field of Machine Learning. We talked about ML, education and what comes next. Enjoy the interview.

We teamed up with OpenStudy June 21, 2011

NEWSLETTER

Subscribe to our newsletter to receive digest of activity

Your Email:

RECENT EVENTS

WSDM 2011 - Hong Kong