MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Data Mining and Knowledge Discovery Part III - Text, web and multimedia mining

Prof. Dr. Dunja Mladenić

Information and Communication Technologies (ICT2), 2017/2018

www.mps.si

Requirements

- Attendance at the lectures
- Reading Homework
  - presentations 30.1.2018
- Report on the results of the project work to be sent via e-mail by 22.01.2018 to Janez.Brank@ijs.si
  - 5-10 pages report
- Presentation of the project on 30.01.2018
  - 5-10 slides presentation (10-15 minutes presentation)
- Oral exam on 30.01.2018
  - based on the material presented at the lectures, the lectures slides, additional reading/video material
  - demonstrate understanding of the material including its usage in practical research and application settings beyond the lectured settings

Our intelligence, our sophistication, is the key to our living!... Old age without wisdom, youth without success and childhood without smiles are worthless. [Bhajan, 2001]

Outline

I. Introduction
- finding regularities
- processing text
- finding statistical artifacts instead of evidence

II. Representation
- lexical, syntactic, semantic

III. Tasks
- extracting triplets from text
- learning document extracts
- user modeling
- communication analysis

IV. Techniques
- supervised , semi-supervised, unsupervised learning

V. Handling data size
- atypical operators
- storing big data

Jožef Stefan Institute, Artificial Intelligence Laboratory

Jozef Stefan Institute (JSI) is the leading Slovene research institution for natural sciences (900+ people) in the areas of computer science, physics, chemistry, ecology

Artificial Intelligence Laboratory has over 40 people working in various areas of artificial intelligence (machine learning, data mining, social network analysis, semantic technologies, computational linguistics, logic)

Spinoff-s: Quintlligence, Cyc-Europe, LiveNetLife, ModroOko, Envigence

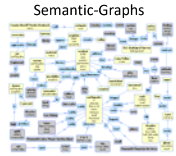
Academic Partners: Carnegie Mellon, Cornel, Stanford, MIT, Uni. Maryland, KIT, UCL,…

Business Clients: Accenture Labs, Bloomberg, British Telecom, Google Labs, Microsoft Research, New York Times, Siemens, Wikipedia

Selection of Portals and Products:

- ❖Text-Garden (http://www.textmining.net)
- ❖Enrycher (http://enrycher.ijs.si/)
- ❖VideoLectures.NET (http://videolectures.net/)
- ❖IST-World (http://www.ist-world.org/)
- ❖Search-Point (http://searchpoint.ijs.si/)
- ❖OntoGen (http://ontogen.ijs.si/)
- ❖Document-Atlas (http://docatlas.ijs.si/)
- ❖Contextify (http://contextify.net/)
- ❖NewsFeed (http://newsfeed.ijs.si/)
- ❖ DiversiNews (http://aidemo.ijs.si/diversinews/)
- ❖EventRegistry (http://eventregistry.org/)
- ❖Twitter Observatory (http://twitterobservatory.net/)

Semantic-Graphs



Document-Atlas

Event Registry

SearchPoint

VideoLectures.net

Enrycher

Selection of Projects (Integrated Projects and Networks of Excellence only):

Coordinating: **XLike** Cross-lingual Knowledge Extraction; **Toposys** Topological Complex Systems; **NRG4Cast** Energy Forecasting

H2020: MSCA **RENOIR** Reverse EngiNeering of sOcial Information pRocessing, MSCA **BigDataFinance, OPTIMUM** Multi-source Big Data Fusion Driven Proactivity for Intelligent Mobility, **AQUASMART** Aquaculture Open Data Cloud Innovation, CSA **EDSA** European Data Science Academy

IP: **ACTIVE, COIN, EURIDICE, NeOn, ECOLEAD, SEKT**

NoE: **PlanetData, PASCAL2, MetaNet, Multilingual Web, LT-Web**

ailab.ijs.si

Research Areas

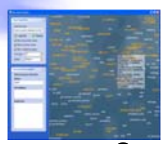
- Artificial Intelligence, Machine Learning,
- Data-Mining, Text-Mining, Web-Mining,
- Semantic Technologies, Social network Analysis,
- Language Technologies, Natural Language Processing, Multi-lingual, Cross-lingual technologies,
- Scalability, Real-time data analysis,
- Data visualization,
- Knowledge management,
- Knowledge Reasoning, Sensor Networks

Technologies

Deep Semantics & Reasoning (Cyc)

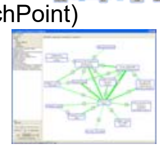
Light-Weight Semantic Technologies (OntoGen)

Social Computing/Web2.0 (LiveNetLife)

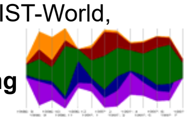


Computational Linguistics (Enrycher, AnswerArt)

Complex Data Visualization (DocAtlas, NewsExplorer, SearchPoint)



Graph/Social Network Analysis (GraphGarden/SNAP, IST-World, FPIntelligence)



Data/Web/Text/Stream-Mining (TextGarden Suite of tools)

Statistical Machine Learning

ailab.ijs.si

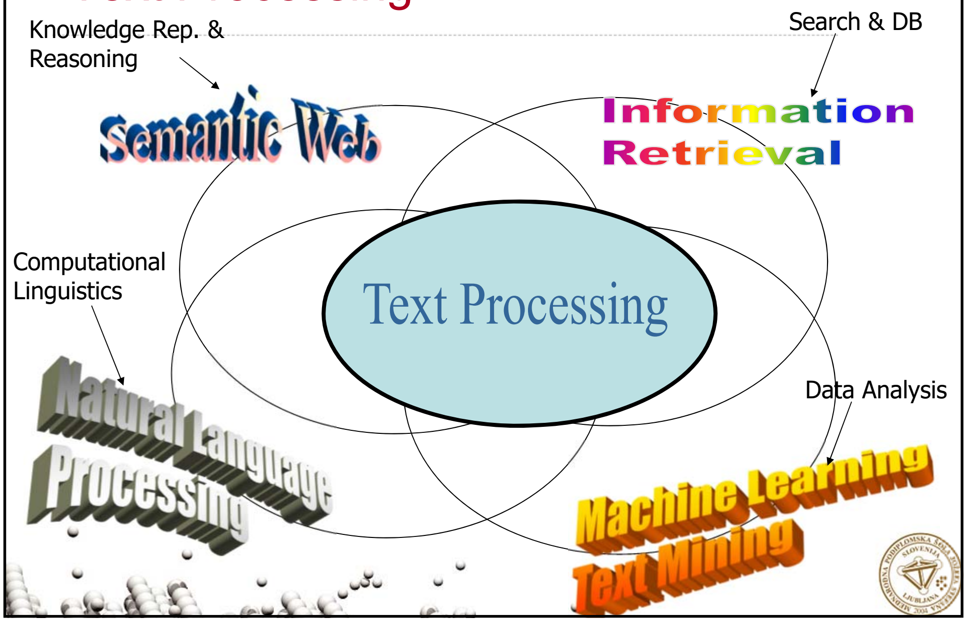
What are we talking about?

Data as a starting point – applying algorithms to data

Text, Web, Multimedia – potentially large datasets

- Goal: "…finding **interesting** regularities in large **text, web or multimedia** data…" [Usama Fayad, adapted]
  - …where **interesting** means: non-trivial, hidden, previously unknown and potentially useful
- Find semantic and abstract information from the raw data
  - surface form of text, bitmap of photos, graph structure
- Find regularities in web-structure, -logs, -content

Research areas contributing to Text Processing



Big Data
Truth or a random phenomenon?

Risk with "Big Data mining"

- we can "discover" patterns that occur by chance
- …if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Bonferroni's princi
we find a statistica
we are looking for
– a patter is there by
of the phenomeno

"*...truth is simple, straight and with a smile. You don't have to remember it. You have to say it. You know it and then you have to live it. It is so simple.* " [Y.Bhajan]

Meaningfulness of Analytic Answers

Calculate the expected number of occurrences of the pattern under the assumption that the data is random

Illustrative example

- Find (unrelated) people who at least twice have stayed at the same hotel on the same day (can be different hotel each day)
  - 10^9 people being tracked
  - 1000 days
  - each person stays in a hotel 1% of the time (1 day out of 100) – probability of staying in a hotel is 0.01
  - there are 10^5 hotels, capacity of a hotel is 100 people

If everyone behaves randomly (i.e., no conspiracy) will the data mining (by chance) detect anything suspicious?

Example taken from: Rajaraman, Ullman: Mining of Massive Datasets

Calculation of patterns detected by chance

Event/pattern: 2 people on 2 days stay in the same hotel

- 2 people at the same day go to a hotel
  - a person stays in a hotel 1% of the time, $0.01 * 0.01 = 10^{-4}$
- 2 people at the same day go to the same hotel (10^5 hotels)
  - probability $= 10^{-4} * 10^{-5} = 10^{-9}$
- 2 people at the same day go to the same hotel, occurs twice
  - probability $= 10^{-9} * 10^{-9} = 10^{-18}$

$$\binom{n}{2} = \frac{n*(n-1)}{2} = \frac{n^2-n}{2} \approx \frac{n^2}{2}$$

Random behavior

- Choose 2 people from $\mathbf{10^9}$ and choose 2 days from $\mathbf{10^3}$
  - ways to choose: $10^{18}/2 * 10^6/2 = 5 * 10^{17} * 5 * 10^5 = 25 * 10^{22}$
- Event probability - expected number of "suspicious" pairs of people in random data (out of 10^9 people) = 250 000 (!)
  - $25 * 10^{22} * 10^{-18} = 25 * 10^4 = 250\ 000$
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

Example taken from: Rajaraman, Ullman: Mining of Massive Datasets

Variation: 10^7 people being tracked instead of 10^9

Random behavior

- Choose 2 people from **10^7** and choose 2 days from 10^3
  – ways to choose: $10^{14}/2 * 10^6/2 = 5 * 10^{13} * 5 * 10^5 = 25 * 10^{18}$

- Event probability - expected number of "suspicious" pairs of people in random data of 10^7 people
  – $25 * 10^{18} * 10^{-18} = 25$

Example taken from: Rajaraman, Ullman: Mining of Massive Datasets

Text/Data Analytics

Three major dimensions:

- Representations
  - from character-level over word level to first-order theories
- Tasks
  - from search over (un-, semi-) supervised learning, to visualization, summarization, translation …
- Techniques
  - from manual work over learning to reasoning

Handling Data Size - Big Data

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Representing Text Data

www.mps.si

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

Lexical

- Language models
- Full-parsing
- Cross-modality

Syntactic

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Semantic

Levels of text representations

- Character (character n-grams and sequences)
- Words (stop-words, stemming, lemmatization)
- Phrases (word n-grams, proximity features)
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

- Language models
- Full-parsing
- Cross-modality

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- 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

Multilingual search, Associating text with images, …

Unifying semantics of data

Reasoning, Semantic search

Lexical

Syntactic

Semantic

Levels of text representations

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

- Language models
- Full-parsing
- Cross-modality

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Lexical

Syntactic

Semantic

Character level

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

Good and bad sides

- Representation has several important strengths:
  - …it is very robust since avoids language morphology
    - (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

Levels of text representations

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

Lexical

- Language models
- Full-parsing
- Cross-modality

Syntactic

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Semantic

Word level

- The most common representation of text used for many techniques
  - …there are many tokenization software packages which split text into the words
- Important to know:
  - Word is well defined unit in western languages – e.g. Chinese has different notion of semantic unit

Words Properties

- Relations among word surface forms and their senses:
  - **Homonomy**: same form, but different meaning (e.g. bank: river bank, financial institution)
  - **Polysemy**: same form, related meaning (e.g. bank: blood bank, financial institution)
  - **Synonymy**: different form, same meaning (e.g. singer, vocalist)
  - **Hyponymy**: one word denotes a subclass of 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 a process of transforming a word into its stem
  - (universe, university, universities, university's, universal) –> univers
- Lemmatization transforms word into its normalized form
  - universe -> universe, (university, universities, university's) -> university, universal –> universal
- …stemming provides an inexpensive mechanism to merge words with similar meaning

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
- Words
- **Phrases**
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

Lexical

- Language models
- Full-parsing
- Cross-modality

Syntactic

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Semantic

Phrase level

- Instead of having just single words we can deal with phrases
  - artificial intelligence, text mining, word for windows
- We use two types of phrases:
  - Phrases as frequent contiguous word sequences
  - Phrases as 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-grams

• September 2006, Goolge released n-grams (sequences of up to n words)

Length of n-gram	Number of different n-grams
1	13,588,391
2	314,843,401
3	977,069,902
4	1,313,818,354
5	1,176,470,663
no. sentences	95,119,665,584
no. words	1,024,908,267,229

passive smoking increased the risk
cow eats grass
humans currently reside on earth
iraq declared war
ship docked in the port
we use this a lot
for all the examples </S>
15th Century Book of Hours
170USD go thread (1
1395 0 BEA171 H 19

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html#links

Example: 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
ceramics company , 133

- 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

Levels of text representations

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

- Language models
- Full-parsing
- Cross-modality

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Lexical

Syntactic

Semantic

Part-of-Speech level

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

Part-of-Speech Table

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.

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
- Words
- Phrases
- Part-of-speech tags
- **Taxonomies / thesauri**
- Vector-space model

Lexical

- Language models
- Full-parsing
- Cross-modality

Syntactic

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Semantic

Taxonomies/thesaurus level

- Thesaurus has a main function to connect different surface word forms with the same meaning into one sense (synonyms)
  - …additionally we often use hypernym relation to relate general-to-specific word senses
  - …by using synonyms and hypernym relation we compact the feature vectors
- The most commonly used general thesaurus is WordNet which exists in many languages (e.g. EuroWordNet)
  - http://www.illc.uva.nl/EuroWordNet/

WordNet – database of lexical relations

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

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

Lexical

Syntactic

Semantic

Vector-space model level

- The most common way to deal with documents is first to transform them into **sparse numeric vectors** and then deal with them with **linear algebra operations**
  - …by this, we forget everything about the linguistic structure within the text
  - …this is sometimes called “structural curse” because this way of forgetting about the structure doesn’t harm efficiency of solving many relevant problems
  - This representation is referred to also as “Bag-Of-Words” or “Vector-Space-Model”
  - Typical tasks on vector-space-model are classification, clustering, visualization etc.

Representing documents as vectors

Having a set of documents, represent each as a feature vector:

1. divide text into units (eg., words), remove punctuation, (remove stop-words, stemming,…)
2. each unit becomes a feature having numeric weight as its value (eg., number of occurrences in the text - referred to as term frequency or TF)

Commonly used weight is TFIDF:

$$TFIDF(w) = tf(w) * \log\left(\frac{N}{df(w)}\right)$$

- tf(w) – term frequency (no. of occurrences of word *w* in document)
- df(w) – document frequency (no. of documents containing word *w*)
- N – no. of all documents

Example of document representation

Bob the builder is a children animated movie on a **character** **Bob** and his friends that include several vehicle **characters**. They face challenges and jointly solve them, such as, repair a roof or save **Bob's** cat from a tall tree…

Pixar has several short **animated** **movies** suitable for **children**. Locomotion is one of them showing train engine and a train wagon as two characters that face a challenge of crossing a half-broken bridge…

…

Simpson family provokes a smile on many adult and **children** faces showing everyday life of a family of four…

bob	builder	children	animated	movie	character	friend	vehicle	---	…
3	1	1	1	1	2	1	1	…	…
0	0	1	1	1	1	0	0	…	…
…	…	…	…	…	…	…	…	---	…
0	0	1	0	0	0	0	0	…	…

Similarity between document vectors

- 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)

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

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

Lexical

- **Language models**
- Full-parsing
- Cross-modality

Syntactic

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Semantic

Language model level

- Language modeling is about determining probability of a sequence of words
  - The task typically gets reduced to the estimating probabilities of a next word given two previous words (trigram model):

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

Frequencies of word sequences

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

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

Lexical

- Language models
- **Full-parsing**
- Cross-modality

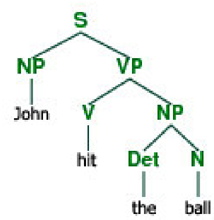
Syntactic

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- 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



Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

- Language models
- Full-parsing
- **Cross-modality**

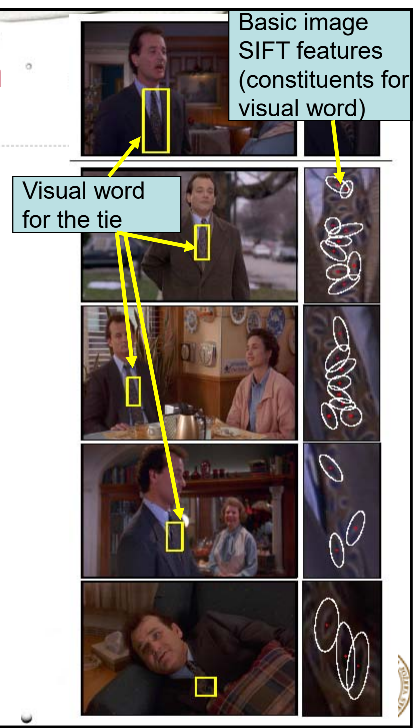
- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Cross-modality level

- It is very often the case that objects are represented with different data types:
  - Text documents
  - Multilingual texts documents
  - Images
  - Video
  - Social networks
  - Sensor networks
- …the question is how to create mappings between different representation so that we can benefit using more information about the same objects

Example: Aligning text with audio, images and video

- The word “**tie**” has several representations (http://www.answers.com/tie&r=67)
  - Textual
  - Multilingual text
    - (tie, kravata, krawatte, …)
  - Audio
  - Image:
    - http://images.google.com/images?hl=en&q=necktie
  - Video (movie on the right)
- Out of each representation we can get set of features and the idea is to correlate them
  - KCCA (Kernel Correlation Analysis) method generates mappings between different representations into “**modality neutral**” data representation



Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

Lexical

- Language models
- Full-parsing
- Cross-modality

Syntactic

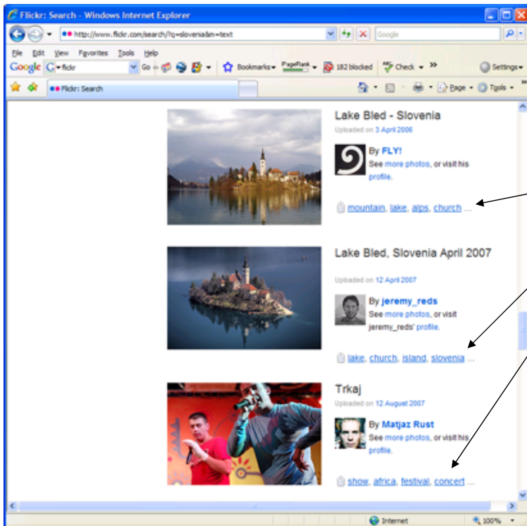
- **Collaborative tagging / Web2.0**
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

Semantic

Collaborative tagging

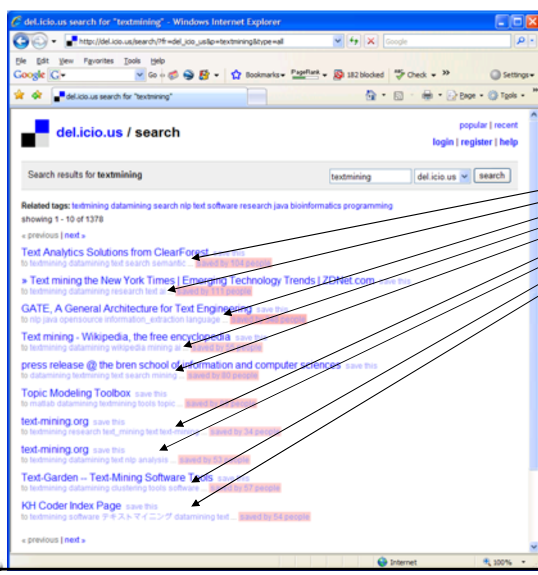
- Collaborative tagging is a process of adding metadata to annotate content (e.g. documents, web sites, photos)
  - …metadata is typically in the form of keywords
  - …this is done in a collaborative way by many users from larger community collectively having good coverage of many topics
  - …as a result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

Example: flickr.com tagging



Tags entered by users annotating photos

Example: del.icio.us tagging



Tags entered by users annotating Web sites

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

Lexical

- Language models
- Full-parsing
- Cross-modality

Syntactic

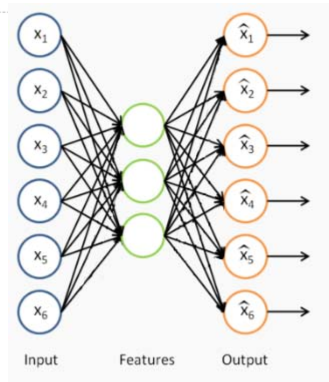
- Collaborative tagging / Web2.0
- Learning Features – word embedding
- Templates / Frames
- Ontologies / First order theories

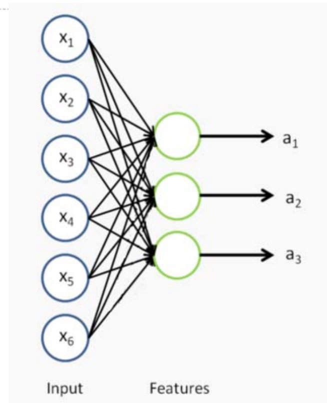
Semantic

Learning Features

- Generate new features from unlabeled data using machine learning methods
- Deep learning on text
  - features taken from hidden layers of a deep neural network that was trained on the original features
  - captures latent structure in text

Learning Features



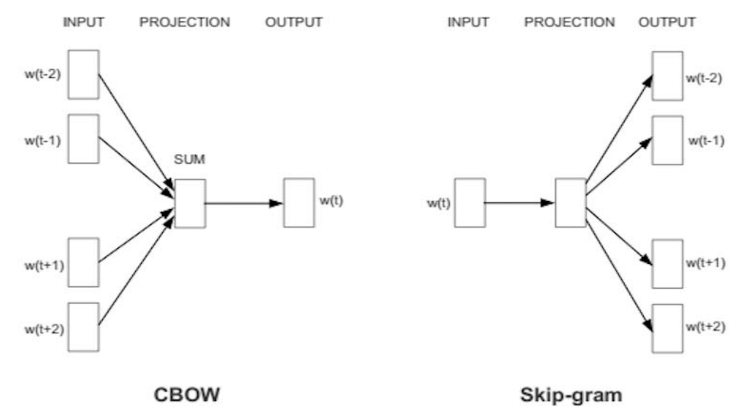


Idea:
learn features from unlabeled data, take the hidden units as new features

Learning Features

- Generate new features from unlabeled data using machine learning methods
- Deep learning on text
  - features taken from hidden layers of a deep neural network that was trained on the original features
  - captures latent structure in text
- Neural word embedding
  - two-layered neural network
  - captures co-occurrences of words

Learning Features



Neural word embedding:

- Continuous bag-of-words – using context to predict the word
- Skip-grams – using word predict its context
- …

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

- Language models
- Full-parsing
- Cross-modality

- Collaborative tagging / Web2.0
- Learning Features – word embedding
- **Templates / Frames**
- Ontologies / First order theories

Template / frames level

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

Examples of templates of KnowItAll system

- Generic approach of extracting is described in
  - *Unsupervised named-entity extraction from the Web: An experimental study [Oren Etzioni et al]*
- KnowItAll system uses the following generic templates:
  - NP "and other" <class1>
  - NP "or other" <class1>
  - <class1> "especially" NPList
  - <class1> "including" NPList
  - <class1> "such as" NPList
  - "such" <class1> "as" NPList
  - NP "is a" <class1>
  - NP "is the" <class1>
- …each template represents specific relationship between the words appearing in the variable slots
- From template patterns KnowItAll bootstraps new templates

Levels of text representations

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

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 relations) 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

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Example tasks

www.mps.si

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

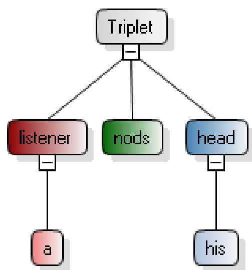
Extracting triplets from text

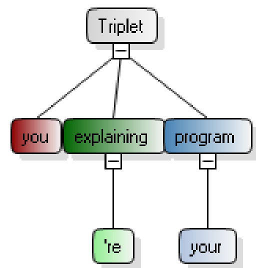
RUSU, Delia, FORTUNA, Blaž, GROBELNIK, Marko, MLADENIĆ, Dunja. Semantic graphs derived from triplets with application in document summarization. *Informatica (Ljublj.)*, 2009, vol. 33, no. 3, pp. 357-362.

www.mps.si

Task

- Extract (*subject, predicat, object*) triplets from text
- Example:
  - If a listener nods his head while you're explaining your program; wake him up.



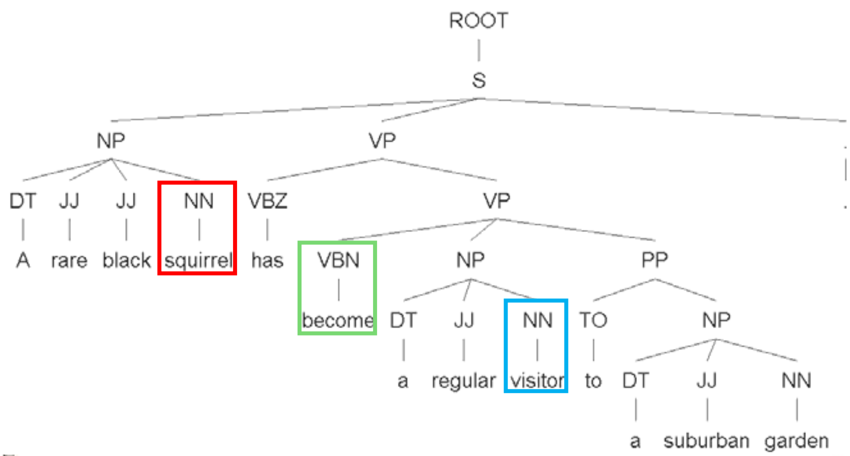


Extraction of triplets using parsers

- Approach description:
  - Parse the sentence with a deep parser
  - Determine subject, object and predicate from the parse tree
- Advantage:
  - Many freely available parsers
- Disadvantage:
  - Solves much harder problem (deep parsing) in order to extract triplets

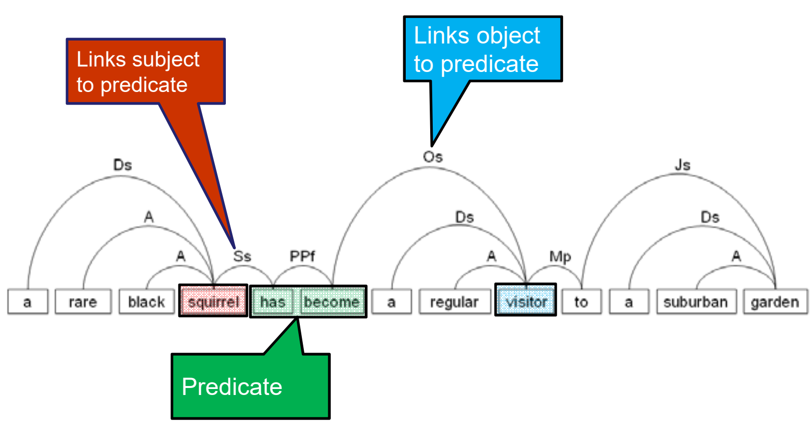
Using OpenNLP

A rare black squirrel has become a regular visitor to a suburban garden.



Using Linked Parser

A rare black squirrel has become a regular visitor to a suburban garden.



Machine learning approach

- Triplet extraction can be defined as a binary classification problem
  - Set of tree words from a sentence can be positive (an actual triplet) or negative (not a triplet).
  - Classification algorithms, such as SVM, can be naturally applied to this task

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Learning Document Extracts

- Leskovec, J., Milic-Frayling, N., Grobelnik, M. (2005), Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts, In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI), 1069-1074, July 2005, Pittsburgh, Pennsylvania.*
- Leskovec, Grobelnik, Milic-Frayling, LinkKDD 2004 (Learning Sub-structures of Document Semantic Graphs for Document Summarization)
- Rusu, Fortuna, Grobelnik, Mladenić, Informatica 2009 (Semantic Graphs Derived From Triplets With Application In Document Summarization)

www.mps.si

Document Extracts

- Document
  - too small to count on statistics
  - identify and use linguistic and semantic structure
- Data from "Document Understanding Conference"
  - set of documents and their summaries
- Approach
  - extract semantic network from a document and identify relevant parts to represent summary
- Experimental results
  - 70% recall of and 25% precision on extracted Subject-Predicate-Object triples

Document (plain text format)

↓

Text preprocessing

↓

- Named entity extraction
- Co-reference resolution
- Pronominal anaphora resolution

↓

Subject – Predicate – Object triplet extraction

↓

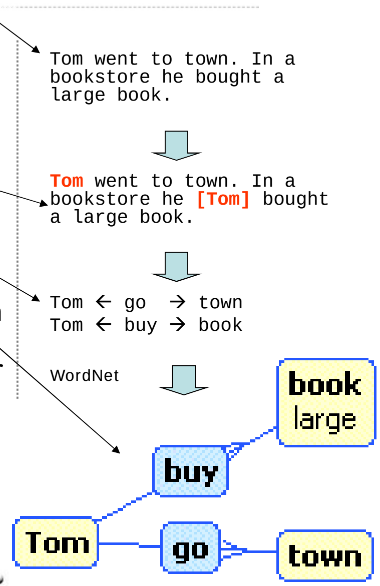
Triplet enhancement
- linking triplets to named entities
- semantic normalization

↓

Semantic graph

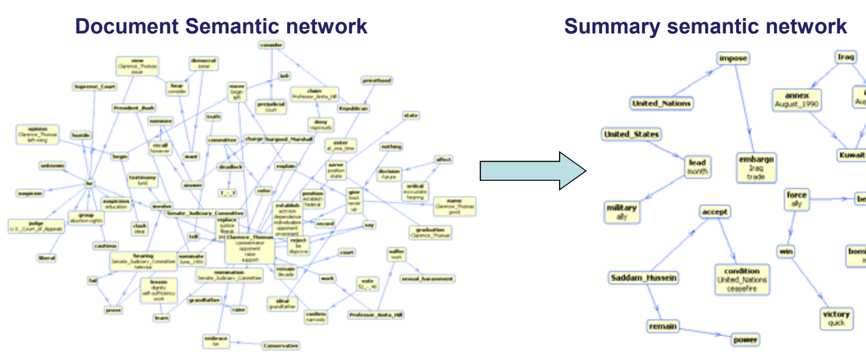
Knowledge Rich Summarization Example

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

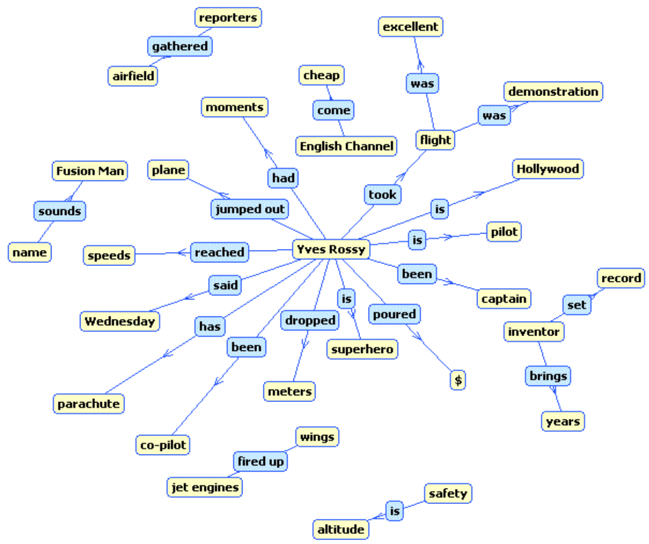


Training of summarization model

- A model was trained deciding which **Subject-Predicate-Object** triple belongs into the target summary
- For training was used Support Vector Machine (SVM) on 400 statistic, linguistic and graph topological features



Semantic graph



MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

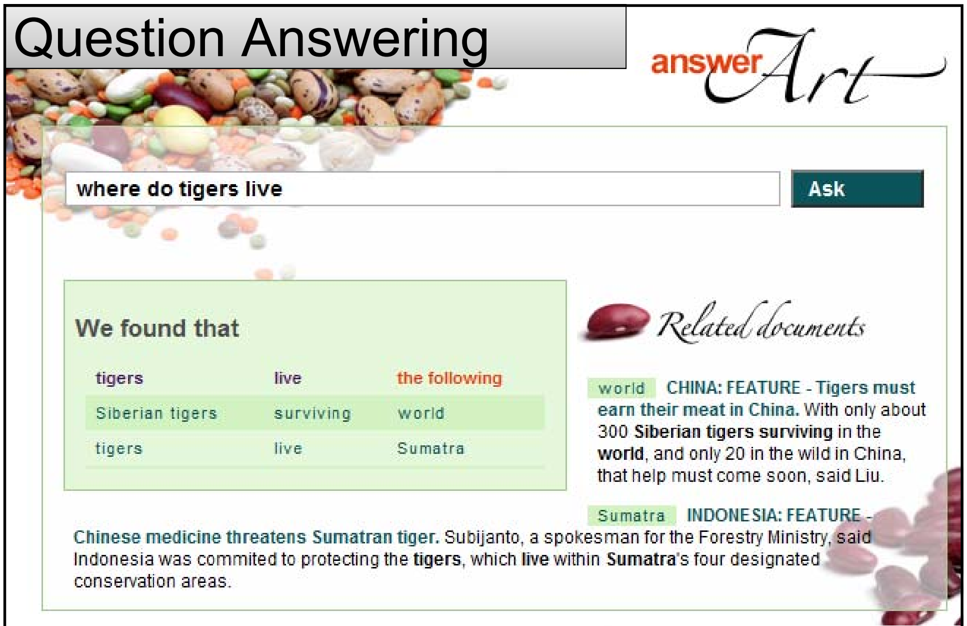
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Question Answering

• DALI, Lorand, RUSU, Delia, FORTUNA, Blaž, MLADENIĆ, Dunja, GROBELNIK, Marko. *Question answering based on semantic graphs. WWW-2009 Workshop on Semantic Search.*
• DALI, Lorand, RUSU, Delia, FORTUNA, Blaž, MLADENIĆ, Dunja, GROBELNIK, Marko. AnswerArt - contextualized question answering. *ECML PKDD 2010, 2010,* pp. 579-582.
• BRADEŠKO, Luka, DALI, Lorand, FORTUNA, Blaž, GROBELNIK, Marko, MLADENIĆ, Dunja, NOVALIJA, Inna, PAJNTAR, Boštjan. Contextualized question answering, ITI-2010.

www.mps.si

Question Answering



http://answerart.net/

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Multi-lingual and cross-lingual

- FORTUNA, Blaž, RUPNIK, Jan, PAJNTAR, Boštjan, GROBELNIK, Marko, MLADENIĆ, Dunja. Cross-lingual search over 22 European languages. ACM SIGIR 2008, 2008, pp 883.
- TOMAŠEV, Nenad, RUPNIK, Jan, MLADENIĆ, Dunja. The role of hubs in cross-lingual supervised document retrieval. *Advances in knowledge discovery and data mining : PAKDD 2013,* INCS 7819, pp. 185-196.

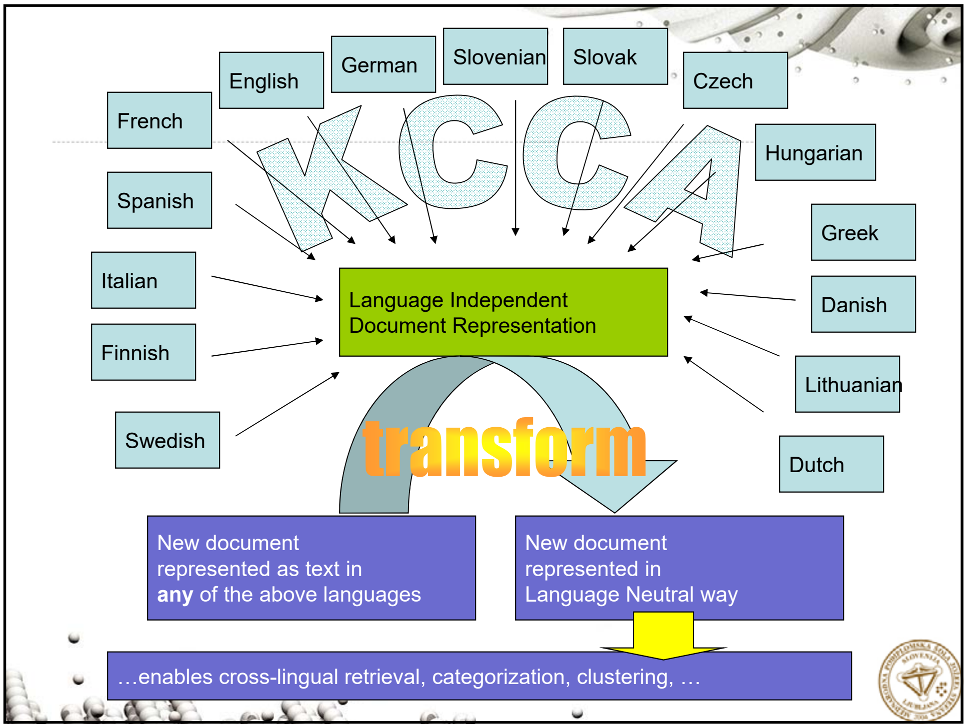
www.mps.si

Multilingual data

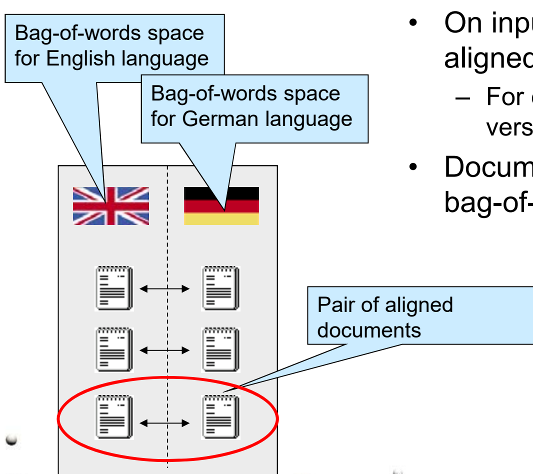
- Text in several natural languages
- Perform machine learning and retrieval on textual data regardless the language differences
- Approach:
  - Machine Translation (on sentence level)
  - Multilingual lexicon (on word level)
  - Mapping into semantic space (on word level, eg., KCCA)

KCCA to handle multilingual data

- KCCA enables representing documents in a "**language neutral way**"
- Intuition behind KCCA:
  1. Given a parallel corpus (such as Acquis)…
  2. …first, we automatically identify language independent semantic concepts from text,
  3. …then, we re-represent documents with the identified concepts,
  4. …finally, we are able to perform cross language statistical operations (such as retrieval, classification, clustering…)



Input for KCCA



- On input we have set of aligned documents:
  - For each document we have a version in each language
- Documents are represented as bag-of-words vectors

The Output from KCCA

- **The goal:** find pairs of *semantic dimensions* that co-appear in documents and their translations with high correlation
  - *Semantic dimension* is a weighted set of words.
- These pairs are pairs of vectors, one from e.g. English bag-of-words space and one from German bag-of-words space.

The Algorithm – Theory

Formally the KCCA solves:

max$_{(x,y)}$ Corr(<*x*,>, <*y*,>)

- *x, y* – semantic directions for English and German
- (,) is a pair of aligned documents

Examples of Semantic Dimensions from Acquis corpus: English-French (1/2)

Most important words from semantic dimensions automatically generated from 2000 documents:

DIRECTIVE, DECISION, VEHICLES, AGREEMENT, EC, VETERINARY, PRODUCTS, HEALTH, MEAT
DIRECTIVE, DECISION, VEHICULES, PRESENTE, RESIDUS, ACCORD, PRODUITS, ANIMAUX

NOMENCLATURE, COMBINED, COLUMN, GOODS, TARIFF, CLASSIFICATION, CUSTOMS
NOMENCLATURE, COMBINEE, COLONNE, MARCHANDISES, CLASSEMENT, TARIF, TARIFAIRES

EMBRYOS, ANIMALS, OVA, SEMEN, ANIMAL, CONVENTION, BOVINE, DECISION, FEEDINGSTUFFS
EMBRYONS, ANIMAUX, OVULES, CONVENTION, SPERME, EQUIDES, DECISION, BOVINE, ADDITIFS

SUGAR, CONVENTION, ADDITIVES, PIGMEAT, PRICE, PRICES, FEEDINGSTUFFS, SEED
SUCRE, CONVENTION, PORC, ADDITIFS, PRIX, ALIMENTATION, SEMENCES, DECISION

EXPORT, LICENCES, LICENCE, REFUND, VEHICLES, FISHERY, CONVENTION, CERTIFICATE, ISSUED
EXPORTATION, CERTIFICATS, CERTIFICAT, PECHE, VEHICULES, ..., CONVENTION

Veterinary, Transport
Customs
Export Licences
Agriculture
Veterinary

Examples of Semantic Dimensions from Acquis corpora: English-Slovene (2/2)

Most important words from semantic dimensions automatically generated from 2000 documents :

OLIVE, OIL, AID, SUGAR, PRICE, STATE, MILK, LICENCES, OR, EXPORT, INTERVENTION
OLJA, OLJCNEGA, POMOCI, SLADKORJA, POMOC, OLJK, SLADKOR, ALI, DOVOLJENJA

NOMENCLATURE, COLUMN, COMBINED, GOODS, TARIFF, CLASSIFICATION, ST, ANNEXED, INVOKED
NOMENKLATURO, STOLPCU, NOMENKLATURE, KOMBINIRANO, KOMBINIRANE, CARINSKI, BLAGA

QUOTAS, TARIFF, SEED, CUSTOMS, COLUMN, ENERGY, INVOKED, ATOMIC, QUOTA, OPENING
KVOT, TARIFNE, SEMENA, KVOTE, TARIFNIH, CARINSKI, ATOMSKO, ENERGIJO, ODPRTJU

DESIGNATIONS, GEOGRAPHICAL, INDICATIONS, EURATOM, PROTECTED, ECSC, NAMES, ORIGIN
OZNACB, EURATOM, GEOGRAFSKI, POREKLA, ESPJ, ZASCITENIH, OZNACBE, IMEN, REGISTER

WINE, WINES, ALCOHOL, DRINKS, ...TILLATION, POULTRYMEAT, ICEWINE, ANALYSIS
VINO, VINA, VIN, ...NSKEM, VINS..., ALKOHOL, NAMIZNEGA, ...STILACIJO, DESTILACIJE

Agriculture
Customs
Wine
Agriculture protection
Energy

Applications of KCCA

- **Cross-lingual document retrieval**: retrieved documents depend only on the meaning of the query and not its language.
- **Automatic document categorization:** only one classifier is learned and not a separate classifier for each language
- **Document clustering**: documents should be grouped into clusters based on their content, not on the language they are written in.
- **Cross-media information retrieval**: in the same way we correlate two languages we can correlate text to images, text to video, text to sound, …

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

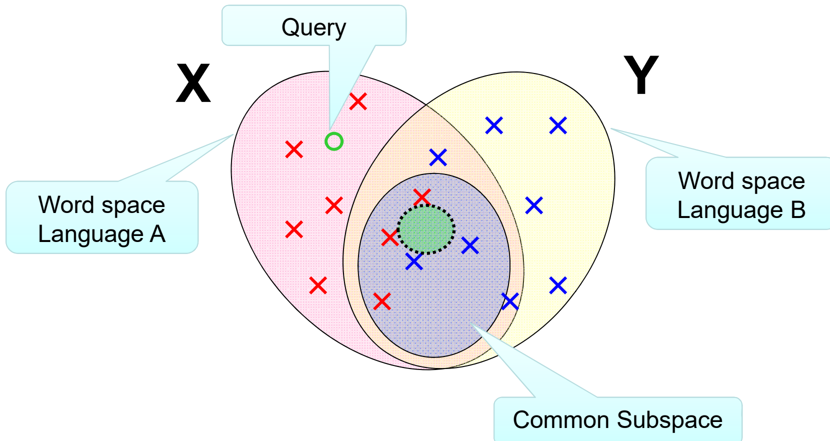
Cross-lingual Similarity

• JanRupnik, Andrej Muhič, Primož Škraba, Multilingual Document Retrieval through Hub Languages, SiKDD2012, 2012.

• RUPNIK, Jan, MUHIČ, Andrej, ŠKRABA, Primož. Cross-lingual document retrieval through hub languages. *NIPS 2012, Neural Information Processing Systems Worshop,* Neural Information Processing System Foundation, 2012.

www.mps.si

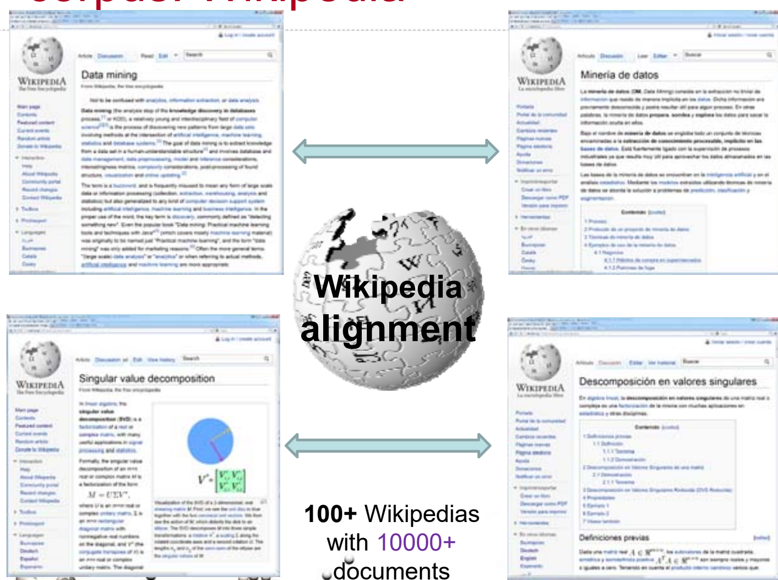
Cross-lingual information retrieval



Cross-lingual similarity function

Similarity computation: 40 ms

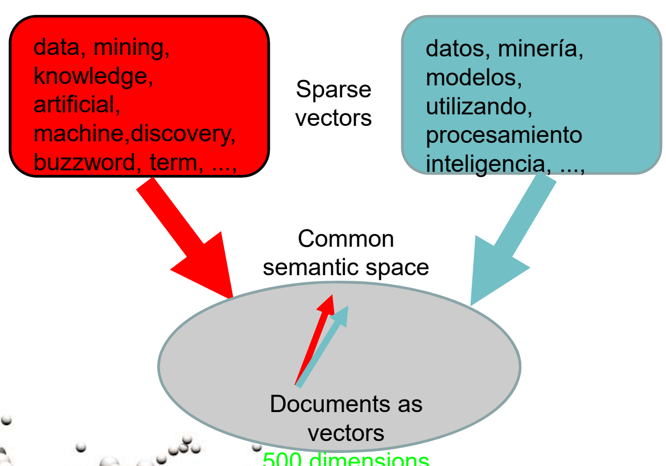
Comparable multilingual corpus: Wikipedia



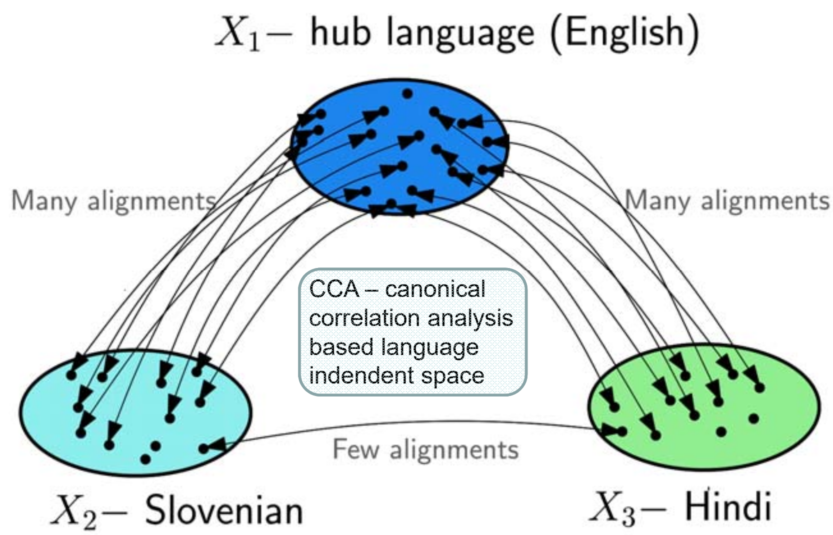
Bag of words vector spaces

Minimal preprocessing:

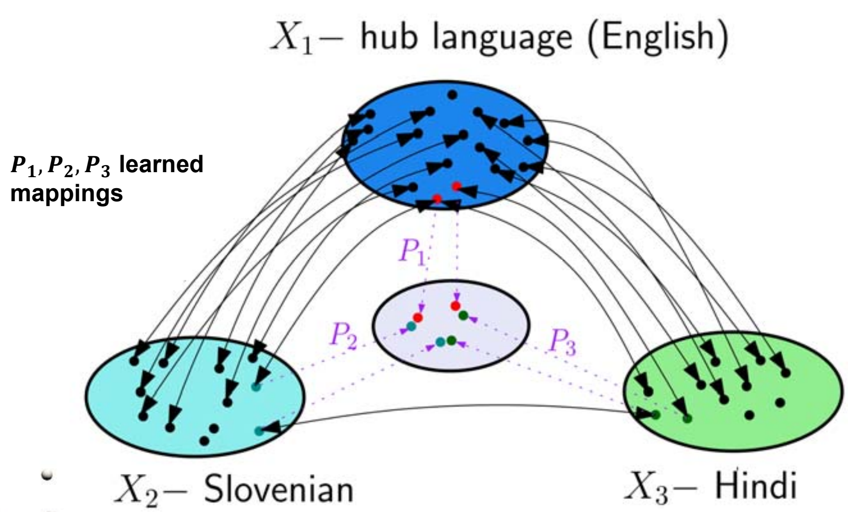
- Remove rare words
- TFIDF weighting
- Vector normalization



Hub language and CCA



Language independent representation



Cross-lingual similarity function

- Large comparable corpora based on Wikipedia for 50+ major languages
- Use of hub languages and CCA in the case of low direct alignment information enables information retrieval
- Demo at http://xling.ijs.si
  - Web page (eg., ailab.ijs.si) – Google translate
  - Wikipedia – the same concept in different languages

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Cross-lingual Named Entity Extraction

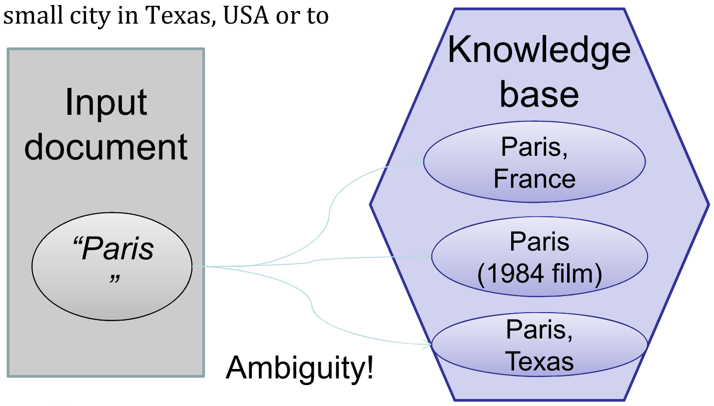
- ŠTAJNER, Tadej, MLADENIĆ, Dunja. Cross-lingual named entity extraction and disambiguation. 4th Jožef Stefan International Postgraduate School Students Conference, 2012, pp. 176-181.
- ŠTAJNER, Tadej, NOVALIJA, Inna, MLADENIĆ, Dunja. Informal multilingual multi-domain sentiment analysis. *Informatica*, ISSN 0350-5596, 2013, 37:4, pp. 373-380.

www.mps.si

Named entity disambiguation

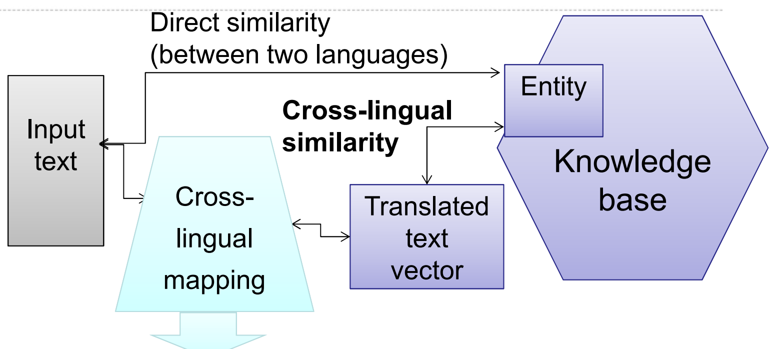
"Paris" can refer to

- a city in France but it can also refer to
- a 1984 film directed by Wim Wenders having title Paris, Texas.
- a small city in Texas, USA or to



Depends on the context – compare Input document and Knowledge base

Cross-lingual named entity disambiguation



a) Machine translation
94% performance of a monolingual baseline (requires a machine translation system)

b) Cross-lingual dictionary
context-independent dictionary constructed from looking at anchor texts from non-English to English Wikipedia pages

c) CCA regression vector space mapping
map the input text into the target language

Experimental evaluation

- Cross-lingual context-similarity using CCA gives better results than Directly calculating similarity
  - Topic of the trained mapping should overlap with the topic of the source text
- Not certain whether it compares favourably to a machine translation based system

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Techniques for Data Modeling

www.mps.si

First steps to data modeling

- Data representation in a suitable format
  - feature vectors are commonly used
  - for each data point (example), each feature has one value from a predefined set of possible values
  - features generation and feature selection may be applied

transformation or combination

feature subset selection

Illustrative example – cartoon descriptions

Bob the builder

Vehicles characters = yes

Human characters = yes

Feature vector = [1, 1]

Features:

- vehicle characters [yes, no]
- human characters [yes, no]

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Supervised Learning

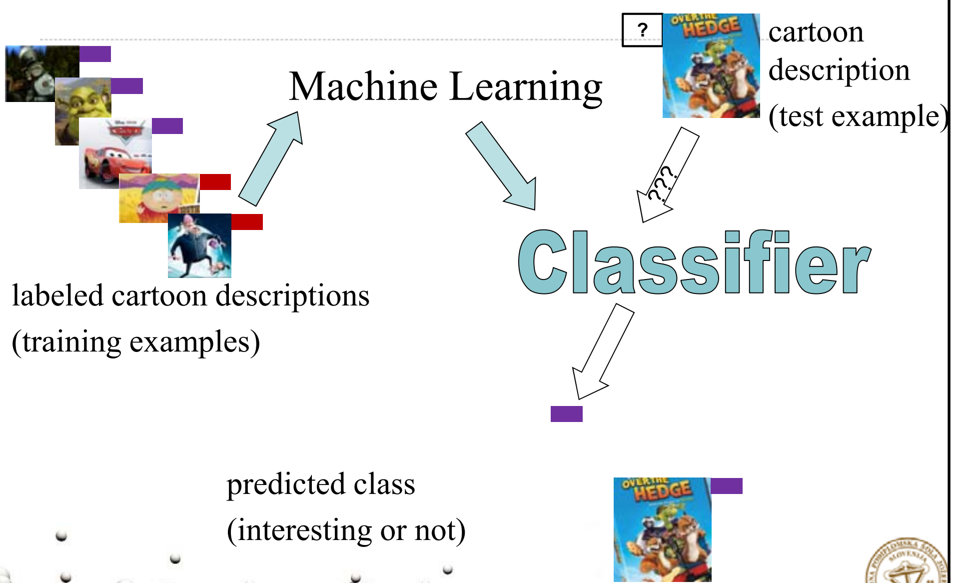
www.mps.si

Supervised learning

Assign an object to a given finite set of classes:

- Document categorization
  - …assign content categories to a text document
- Image classification
  - …decide if an image is relevant for a user
- Spam filtering of e-mails
  - …decide if an email is a spam or a regular email
- Recommending articles in a newspaper
  - …decide if an article fits the user profile
- Semantic/linguistic annotation
  - …assign semantic or linguistic annotation to a word or phrase

Recommending cartoons



Supervised learning

Given: a set of labeled examples represented by feature vectors

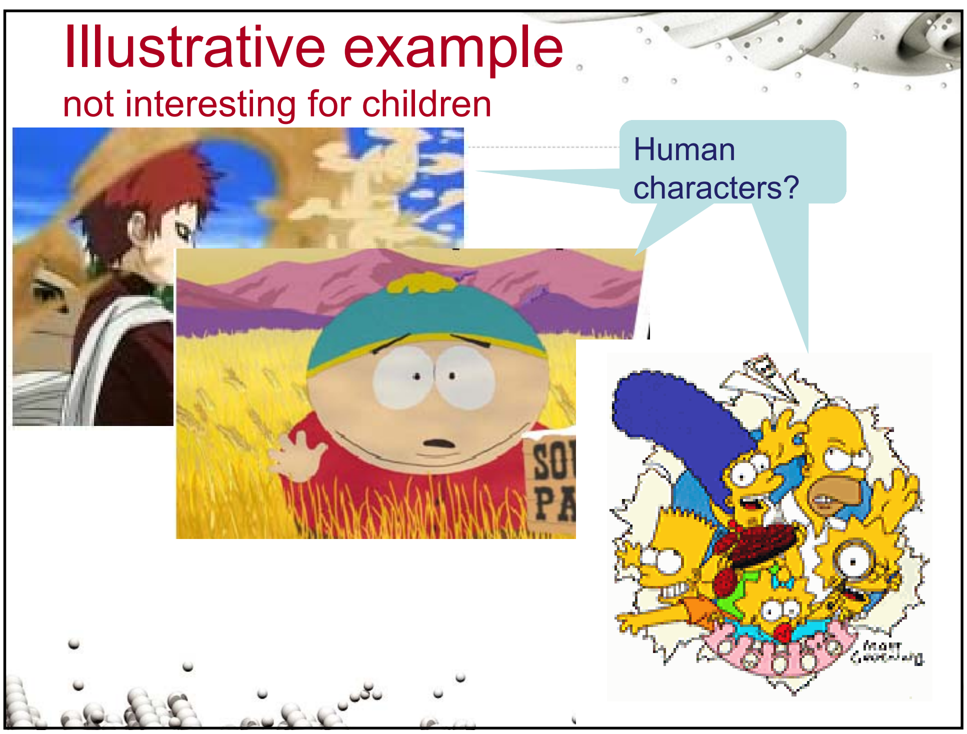
Goal: build a model approximating the target function which would automatically assign right label to a new unlabeled example

- Feature values:
  - discrete (eg., word_occurs ∈ {yes, no}, eyes_color ∈ {brown, blue, green})
  - continuous (eg., age ∈ [0..200])
  - ordered (eg., size ∈{small, medium, large})
- Values of the target function – labels:
  - discrete (classification) or continuous (regression)
  - exclude each other (eg., medical diagnosis) or not (eg., a single document content can talk about several topics)
  - have some predefined relations (taxonomy of document categories, e.,g., DMoz)

The target function can be

- represented in different ways (storing examples, symbolic, numerical, graphical,…)
- modeled by using different algorithms


Short?
strative example
re co mending cartoon for childre
Long &
animals?
Vehicles?
OVER THE
HEDG
Disney · PI AR
Cars
Animals OR
Vehicles


Illustrative example
not interesting for children
Human
characters?

Illustrative example

Recommending cartoon for children

Title	Characteristic words	Duration
Bob the builder	vehicles, human, Bob,..	10 mins
Pixar-Locomotion	vehicles, locomotive,…	5 mins
Ice age	animals, squirrel, ice,…	90 mins
Over the hedge	animals, neighborhood,.	60 mins
Cars	vehicles, car, race,…	90 mins

Target function

There is a trade-off between the expressiveness of a representation and the ease of learning

- The more expressive a representation, the better it will be at approximating an arbitrary function; however, more examples will be needed to learn an accurate function

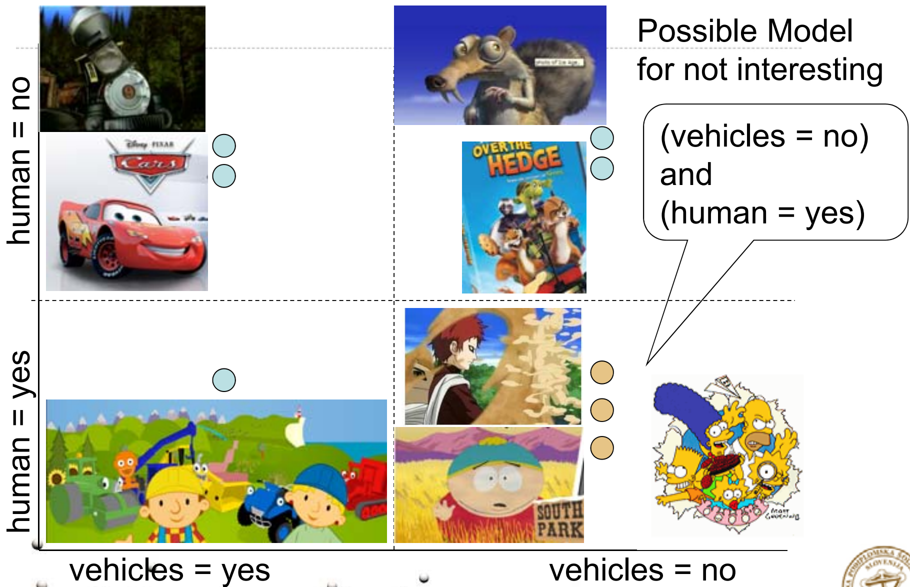
Illustrative example

- Values of the target function: discrete labels (classification), exclude each other

Cartoon interesting for children:

yes	no

Possible data visualization



Generalization

- Model must generalize the data to correctly classify yet unseen examples (the ones which don’t appear in the training data)
- Lookup table of training examples is a consistent model that does not generalize
  - An example that was not in the training data can not be classified

Occam’s razor:

- Finding a *simple* model helps ensure generalization

Algorithms for learning classification models

Storing examples
- Nearest Neighbour

Symbolic
- Decision trees
- Rules in propositional logic or first order logic

Numerical
- Perceptron algorithm
- Winnow algorithm
- Support Vector Machines
- Logistic Regression

Probabilistic graphical models
- Naive Bayesian classifier
- Hidden-Markov Models

Nearest neighbor

- Storing training examples without generating any generalization
  - Simple, requires efficient storage
- Classification by comparing the example to the stored training examples and estimating the class based on classes of the most similar examples
  - Similarity function is crucial

Also known as:
- Instance-based, Case-based, Exemplar-based, Memory-based, Lazy Learning

Similarity/Distance

- For continuous features use Euclidian distance

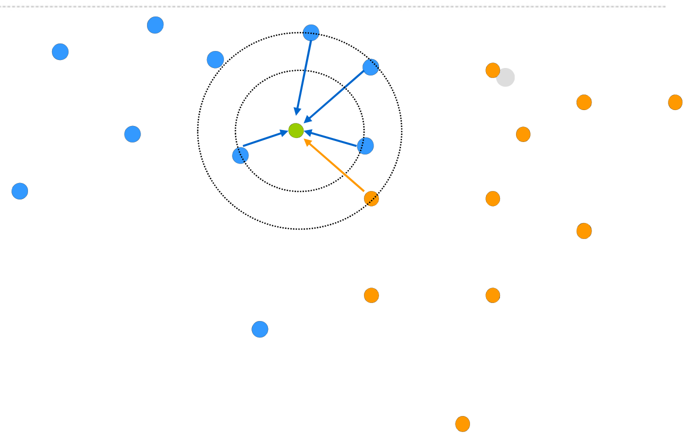
$$Dist(e_1, e_2) = \sqrt{\sum_{i=1}^{n} (f_{1i} - f_{2i})^2}$$

$$e_k = \langle f_{k,1}, f_{k,2}, \dots f_{k,n} \rangle$$

- For discrete features, assume distance between two values is 0 if they are the same and 1 if they are different (eg., Hamming distance for bit vectors).

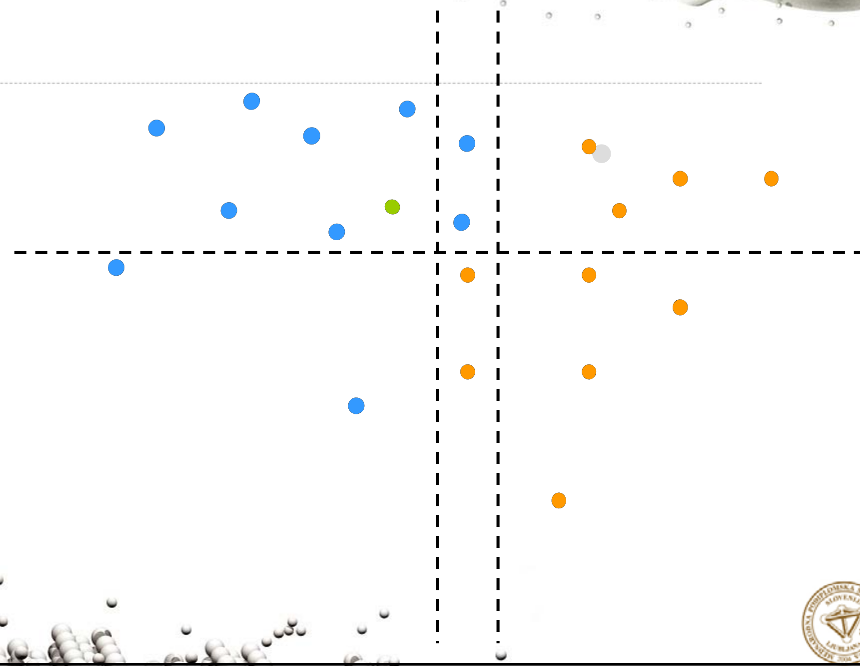
To compensate for difference in units across features, scale all continuous values to the interval [0,1].

Nearest neighbor

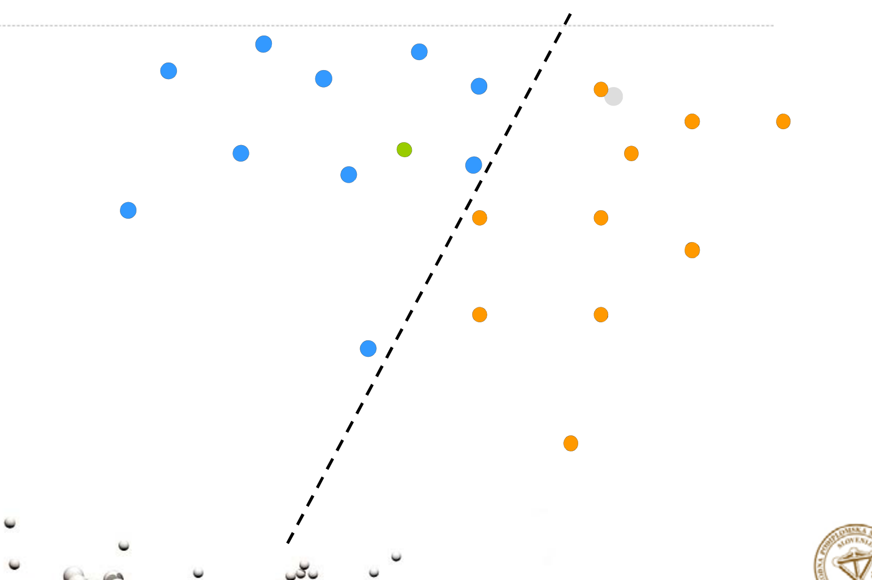


K = 2
K = 5

Decision tree model



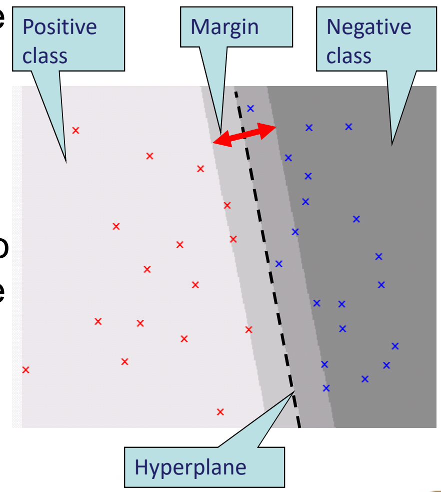
Linear Model



Support Vector Machine

- Learns a hyperplane in higher dimensional space
  - that separates the training data and
  - gives the highest margin
- Implicit mapping of the original feature space into higher dimensional space
  - mapping using so called kernel function (eg., linear, polynomial, …)

Regarded as state-of-the-art in text document classification



SVM Demo

Naïve Bayes

Determine class of example e_k by estimating

$$P(c_i \mid e_k) = \frac{P(c_i)P(e_k \mid c_i)}{P(e_k)} = \arg\max_i P(c_i)P(e_k \mid c_i)$$

- $P(c_i)$ – estimate from the data using frequency: no. of examples with class c_k / no. of all examples
- $P(e_k|c_i)$ – too many possibilities (all combinations of feature values)
  - assume feature independence given the class

$$P(e_k \mid c_i) = \prod_{j=1}^{n} P(f_{kj} \mid c_i)$$

Naïve Bayes on text

$$P(C \mid Doc) = \frac{P(C) \prod_{W \in Doc} P(W \mid C)^{Freq(W, Doc)}}{\sum_i P(C_i) \prod_{W_l \in Doc} P(W_l \mid C_i)^{Freq(W_l, Doc)}}$$

- Document is represented as a set of words W
- For binary classification, each classifier has two distributions: P(W|pos), P(W|neg)
- When having a large collection of binary classifiers (one per category) with unbalanced prior probability, consider only promising categories:
  - calculated P(pos|Doc) is high meaning that the classifier has P(W|pos)>0 for at least some W from the document (otherwise, the prior probability is returned, P(neg) is about 0.90)

Example of Naïve Bayes classifier

	A	B	C	D	E
w1	1	1	1	0	0
w2	0	0	0	0	1
w3	1	0	1	0	0
w4	0	0	0	1	1
w5	1	1	0	0	0

$$P(C \mid Doc) = \frac{P(C) \prod_{W \in Doc} P(W \mid C)^{Freq(W, Doc)}}{\sum_i P(C_i) \prod_{W_l \in Doc} P(W_l \mid C_i)^{Freq(W_l, Doc)}}$$

1.Estimate model parameters from data.

P(pos) = 2/4 = 0.5; P(neg) = 2/4 = 0.5

P(w1|pos) = 2/2 = 1; P(w1|neg) = 0/2 = 0

P(w2|pos) = 0/2 = 0; P(w2|neg) = ½ = 0.5

P(w3|pos) = ½ = 0.5; P(w3|neg)= 0/2 = 0

P(w4|pos) = 0/2 = 0; P(w4|neg) = 2/2 = 1

P(w5|pos) = ½ = 0.5; P(w5|neg) = 0/2 = 0

2. Calculate probability for each class using the model on A.

P(pos|A) = P(pos)* [P(w1|pos)*P(w3|pos) * P(w5|pos)] / sum_c
= 0.5 * [1* 0.5 * 0.5] / 0.125 = 0.125 / 0.125 =1

P(neg|A) = P(neg)* [P(w1|neg)*P(w3|neg) * P(w5|neg)] / sum_c
= 0.5 * [0* 0 * 0] / 0.125 = 0 / 0.125 = 0

3. Classify A returning the most probable class

pos

Generative Probabilistic Models

- Assume a simple (usually unrealistic) probabilistic method by which the data was generated
- Each class value has a different parameterized generative model that characterizes it
- **Training**: Use the data for each category to estimate the parameters of the generative model for that category.
  - **Maximum Likelihood Estimation (MLE):** Set parameters to maximize the probability that the model produced the given training data
  - If M_λ denotes a model with parameter values λ and D_k is the training data for the *k*th class, find model parameters for class *k* (λ_k) that maximize the likelihood of D_k:

$$\lambda_k = \underset{\lambda}{\operatorname{argmax}}\, P(D_k \mid M_\lambda)$$

- **Testing**: Use Bayesian analysis to determine the category model that most likely generated a specific test instance.

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Semi-supervised Learning

www.mps.si

Semi-supervised learning

Similar to supervised learning except that

- we have examples and only some of them are labeled
- we may have a human available for a limited time to provide labels of examples
  - …this corresponds to the situation where all the cartoons in our collection have descriptions, but only a few have label
  - …and occasionally we have a human for a limited time to respond the questions about the cartoons

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Handling the Amount of Data

Big Data

www.mps.si

Information Age - Age of Data Analytics

- Availability of large amounts of data → handling big data
  - millions of documents, sensor readings, astrophysics,…
- Data sources and variety of data → handling different data modalities
  - text understanding, genetics and molecular biology, video streams,…
- Data on different aspects of life → data science
  - fine-grained human behavior, interactions on social media,…

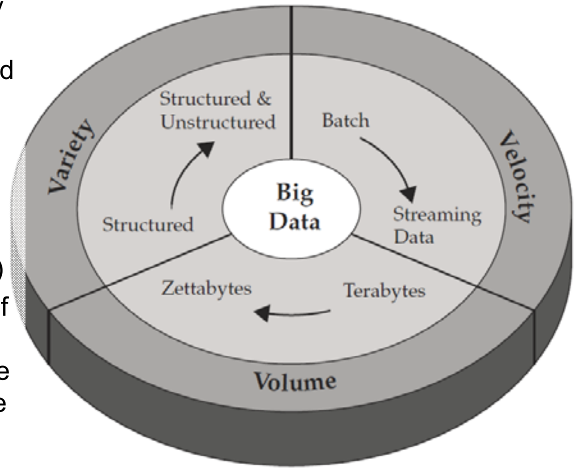
"This is the Information Age — everybody can be informed about ***anything and everything****. There is no secret, therefore there is no sacredness.* ***Life is going to become an open book****. When your computer is more loyal, truthful, informed and excellent than you, you will be challenged. You do not have to compete with anybody. You have to compete with yourself."* *[Bhajan, 2002]*

Small Data, Big Data

- 'Big data' is similar to 'Small data', but bigger
- …but having data bigger consequently requires:
  - different techniques, tools, architectures
- …with an aim to solve new problems and old problems in a better way

Characterization of Big Data: volume, velocity, variety (V3)

- **Volume** – data generated by machines, networks, social media, …. challenging to load and process (how to index, retrieve?)
- **Variety** – many sources and data types with different degree of structure (how to query semi-structured data?)
- **Velocity** – continuous flow of data requires real-time processing influenced by rate of data arrival (how to handle high rate?)



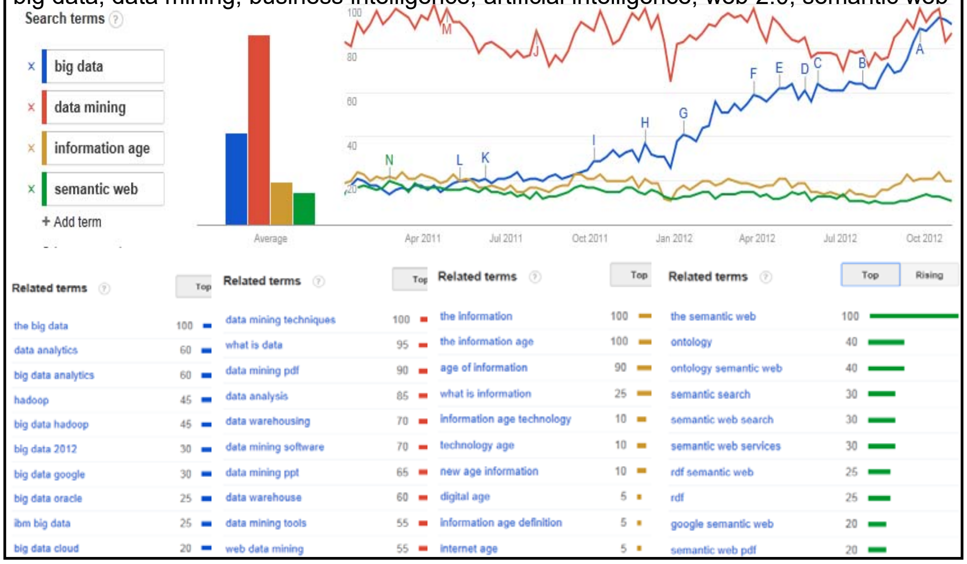
From "Understanding Big Data" by IBM

The extended V3 of Big Data (Vn)

- 1. **Volume** (lots of data = "Tonnabytes")
- 2. **Variety** (complexity, curse of dimensionality)
- 3. **Velocity** (rate of data and information flow)
- 4. **Veracity** (noise and outliers in data, need for verifying the inferred models)
- 5. **Variability** (variance in meaning)
- 6. **Venue** (location)
- 7. **Vocabulary** (semantics)
- 8. **Volatility** (how long is the data valid)

Big Data popularity on the Web

big data, data mining, business intelligence, artificial intelligence, web 2.0, semantic web



What are "atypical" operators on Big-Data

- **Smart sampling** of data
  - …reducing the original data while not loosing the statistical properties of data
- **Finding similar items**
  - …efficient multidimensional indexing
- **Incremental updating** of the models
  - (vs. building models from scratch)
  - …crucial for streaming data
- **Distributed linear algebra**
  - …dealing with large sparse matrices

Sampling on Big-Data

- **Depends on the kind of queries that will be asked**
  - a) sampling x% of all the data points (regardless the source)
  - b) sampling x% of the data sources (eg., users accessing Web site or sensors sending measurements)

  Example: average number of Web pages revisited by the same user
    - Requires sampling all the data for x% of the users
- **Sampling *a/b* fraction of the data (eg., users)**
  - use a hash function to hash the key components of the data stream (eg., username), based on the value of the function decide whether to store the current data or not
  - eg., as the data arrives hash the username to *b* buckets, if the user falls into one of the first *a* buckets store the data

Finding similar items

- Approach as a problem of finding sets with large intersections
  - Jaccard similarity: set_intersection/set_union
- Focus on similarity between the promising pairs of items
  - Eg., usernames with the same hash value, documents of the same length

  Example problem
  - similarity of documents (plagiarism, mirror Web pages, news articles from the same source)
  - Collaborative filtering for movie/book/… recommendation

Data streams

- Data arriving in streams, rapidly so it is not feasible to store all the data
  - Eg., measurements of sensors at different locations – even if one stream is slow there is multitude of them
- What to store depends on the queries that will be asked
  - Standing query (event pattern)
    - trigger an alarm, perform an operation on each arrival of a data point (eg., average the last 100 readings of sensor S), report max. so far
  - Ad-hoc query
    - Store sliding window of the last *n* data points
      - eg., the last 10 values of wind speed
    - Store the last *t* time units readings
      - eg., wind speed during the last hour,
      - eg., the number of unique users on the Web site in the past month – store the complete stream for the last month with the time stamp, so we can remove the old data as new arrives

Clustering

- Grouping data points according to similarity
- Algorithms:
  - Hierarchical
    - a) agglomerative – combining two most similar clusters
    - b) divisive – breaking a cluster

    Cluster similarity calculation
      - Similarity of their centroids or clustorids
      - Sum (or average) similarity of all pairs of points (one from each cluster)
  - Point assignment – assign each point to one of the clusters
    - K-means clustering assigning the points to the most similar of k clusters

Clustering on streams

- BFR algorithm – k-means variant assuming clusters are normally distributed around the centroid
  - Instead of points storing summaries of the clusters + summaries of isolated mini clusters + outliers
- CURE – instead of centroid using a collection of representative points
  - Cluster a small sample of data to choose representative points, move representative points towards centroids, merge clusters with close representatives
  - Assign all other points to one of the clusters based on similarity to representatives
- Clustering on a sliding window – assumes we are interested in clustering of the last *m* points

Why is Big Data *BIG?*

- Mostly due to repeated observations over time and/or space

Examples
  - Web logs with millions of visits per day
  - Supermarket transactions log - thousands of retail stores with tens of thousands of products and millions of customers
  - Satellites regularly sending images

- Big data – "data whose size forces us to look beyond the tried-and-true methods that are prevalent at the time" [A. Jacobs, CACM-2009]

Storing data on more machines

- Most big data have inherent temporal and/or spatial dimension
  - Data with time dimension should be stored and processed at least in a partial temporal ordering
  - Distributed storing of the data should consider the kind of queries that will be asked – if we want different type of queries i.e. over time and over location the data can be replicated to improve efficiency (and provide redundancy over potential hardware failure)
- A cluster of 10 machines is 10 times more likely to require a service than one machine

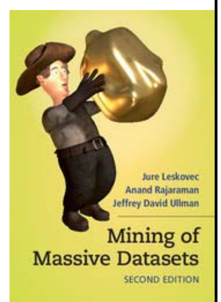
Example:10 years of observations collected at 15s intervals from 1000 sensor sites can be stored on 10 machines:
- All observations for each year on one machine (eg., to return average value for the last year of all sensors)
- All observations for 100 sensors on one machine (eg., to make analysis for one sensor over 10 years)

Analytical operators on Big Data

- On the top of the previous operations we perform usual data mining/machine learning/statistics operators:
  - **Supervised** learning (classification, regression, …)
  - **Non-supervised** learning (clustering, different types of decompositions, …)
  - …
- …we are just more careful which **algorithms** we choose (typically linear or sub-linear versions)

…guide to Big-Data algorithms

- An excellent overview of the “Big Data” algorithms is the book “**Leskovec, Rajaraman, Ullman: Mining of Massive Datasets**”
  - Downloadable from: http://www.mmds.org/
  - Associated MOOC (from Oct 2014): https://www.coursera.org/course/mmds



Homework reading/video

- **Analyzing Text and Social Network Data with Probabilistic Models**, Padhraic Smyth, Center for Machine Learning and Intelligent Systems, University of California http://videolectures.net/ecmlpkdd2012_smyth_probabilistic_models/ (70 min)
- **Semisupervised Learning Approaches,** Tom Mitchell, Machine Learning Department, School of Computer Science, Carnegie Mellon University http://videolectures.net/mlas06_mitchell_sla/ (60 min)
- **Dealing with structured and unstructured data at Facebook,** Lars Backstrom, Facebook http://videolectures.net/eswc2011_backstrom_facebook/ (54 min)
- **Large Scale Learning at Twitter,** Aleksander Kołcz, Twitter, Inc. http://videolectures.net/eswc2012_kolcz_twitter/ (50 min)
- **Using Machine Learning Powers for Good,** Rayid Ghani, University of Chicago http://videolectures.net/lsoldm2013_ghani_learning_powers/ (56 min)
- **Sparsity analsysis of term weighting schemes and application to text classification,** Janez Brank, Artificial Intelligence Laboratory, Jožef Stefan Institute http://videolectures.net/slsfs05_brank_satws/ (30 min)
- **Never Ending Language Learning**, Tom Mitchell, Machine Learning Department, School of Computer Science, Carnegie Mellon University http://videolectures.net/akbcwekex2012_mitchell_language_learning/ (55 min)
- **Automatic Discovery of Patterns in News Content**, Nello Cristianini, Department of Engineering Mathematics, University of Bristol http://videolectures.net/workshops2012_cristianini_news_content/ (40 min)

More available at http://capybara.ijs.si/janez/teaching/pef.html