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]

Računalniška analiza besedil v izobraževanju

predavateljica:
prof. dr. Dunja Mladenić

Institut "Jožef Stefan" in
Mednarodna podiplomska šola Jožefa Stefana,
Slovenija

Način izvedbe

- Predavanja
  - spoznavanje **temeljnih pojmov** in konceptov
  - uporaba prijemov na ilustrativnih primerih, **praktično delo v skupinah**
- Vaje in priprava na samostojno delo v obliki seminarjev
  - študenti se pri pripravi **seminarskih nalog** poglobijo v izbrane teme in jih predstavijo vsem sodelujočim
  - študenti v **projektnih nalogah** v manjših skupinah raziskujejo in aplicirajo pridobljena temeljna znanja na konkretne vsebine

Urnik

8.10.2024	Uvod v analizo besedil
15.10.2024	Principi strojnega učenja na besedilih; Podelitev seminarske naloge
22.10.2024	Predstavitev dokumentov
29.10.2024	Tehnike analize besedil
5.11.2024	Praktične vaje; Podelitev skupinskih projektov (dr. Janez Brank)
12.11.2024	Zahtevnejše metode analize besedil; Predstavitve seminarskih nalog
19.11.2024	Zahtevnejše metode analize besedil; Predstavitve seminarskih nalog
26.11.2024	Praktične vaje; Konzultacije skupinskih projektov (dr. Janez Brank)
10.12. 2024	Zahtevnejše metode analize besedil
7.1.2025	Predstavitev skupinskih projektov (dr. Janez Brank)

Cilji/kompetence predmeta

- zmožnost **povezovanja predhodno pridobljenega znanja** iz različnih področij;
- zmožnost **sodelovalnega reševanja** pedagoških problemov v različnih kontekstih;
- razvoj **znanja in razumevanja** na področju uporabe IKT v izobraževanju;
- zmožnost **raziskovanja in prenašanja** spoznanj v prakso;
- zmožnost **predstavitve strokovno-raziskovalnega dela** v strokovni publicistiki in na strokovnih predstavitvah;
- **poznavanje, razumevanje in apliciranje** zahtevnejših vsebin s področja računalništva;
- **obravnavanje problemov** s področja računalništva v pedagoškem kontekstu z zahtevnejšimi hevristikami in strategijami;
- **obvladovanje zahtevnejših pristopov** za zajem, obdelavo, shranjevanje in predstavitev podatkov v pedagoškem kontekstu.

To be successful you must first have an aim, a goal, a destination… aligned with your life.

[Singh Avenali & Mladenić]

Študijske obveznosti

- Obvezna vsaj 80 % prisotnost (po seznamu prisotnih)
- Sprotna oddaja in predstavitev projektnih/seminarskih nalog kot pogoj za pristop k izpitu
- Uspešno opravljen izpit

Način preverjanja znanja in kriteriji za ocenjevanje

- Predstavitev seminarskega dela:
  - predstavitev seminarske naloge v podanem terminu (pred skupino sošolcev)
  - predstavitev skupinskega projekta v podanem terminu (pred skupino sošolcev)

  http://capybara.ijs.si/janez/teaching/pef.html
- Opravljanje izpita
  - pogoj za pristop k izpitu je udeležba na predavanjih in vajah ter uspešno opravljene seminarske naloge in skupinski projekt
- Preverjanje znanja
  - seminarske naloge z zagovorom in predstavitvijo (30%); skupinski projekt s predstavitvijo (30%), izpit (40%)

Vsebina

Uvod
- Umetna inteligenca
- Odkrivanje zakonitosti v
  - spletnih podatkih
  - tekstovnih podatkih
- Raziskovalni pristop
- Primeri nalog

I Principi strojnega učenja
- Strojno učenje
- Modeliranje podatkov
- Nadzorovano učenje
- Primeri algoritmov

II Strojno učenje na besedilih
- Predstavitev tekstovnih podatkov
- Razvrščanje v skupine

III Tehnike analize besedil
- Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija v taksonomije
- Gradnja vsebinskih ontologij

IV Zahtevnejše metode
- Primeri nalog
  - analiza spletnih uporabnikov
  - spletno oglaševanje
- Vizualizacija besedil
- Izdelava povzetkov
- Prekojezično povezovanje besedil
- Primeri nalog

Trendi razvoja umetne inteligence

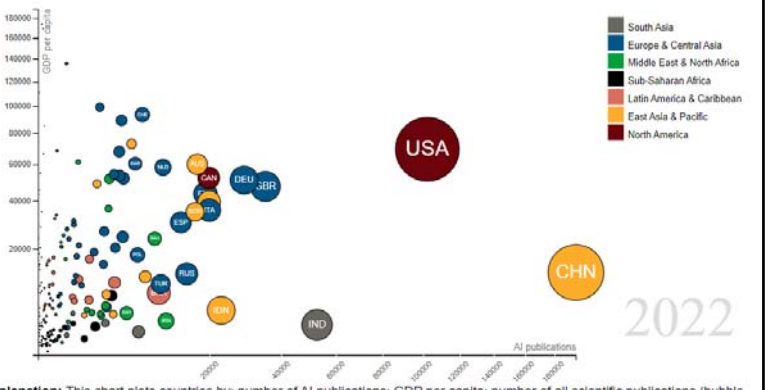
(povzeto po Marko Grobelnik)

- Od sredine 1990tih sledimo učinkovitost AIja po področjih, ki so značilna za človekove sposobnosti
- Okoli 2015 so AI rešitve začele presegati človekove sposobnosti
- Ključen preboj je bil leta 2022, ko je AI presegel človeka v razumevanju besedila/jezika

Source:
Kiela et al. (2021) Dynabench: Rethinking Benchmarking in NLP
OurWorldInData.org/artificial-intelligence • CC BY

Akademske publikacije

- Kitajska prehitela ZDA po številu akademskih publikacij
- Indija se hitro približuje
- Evropa na nivoju ZDA



Explanation: This chart plots countries by: number of AI publications; GDP per capita; number of all scientific publications (bubble size); region (colour); and time (animation). To increase comparability, the "Publications per capita" checkbox divides a country's AI publications and total scientific publications by its total population.

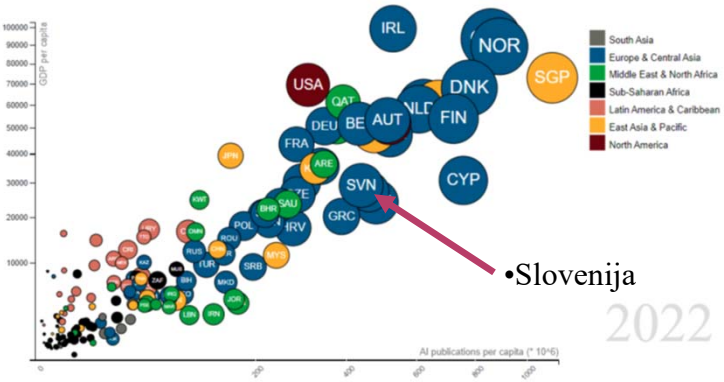
Note: The number of total publications in a country per year is given by the size of its bubble. The 'cumulative' option displays aggregate results since 2000. Data downloads provide a snapshot in time. Due to a lag in reporting, figures for the latest quarter may appear slightly lower than they actually are. This is automatically corrected in subsequent updates. Data downloads provide a snapshot in time. Caution is advised when comparing different versions of the data, as the AI-related concepts identified by the machine learning algorithm may evolve in time. Please see methodological note for more information.

Source of data: OpenAlex
Please cite as: OECD.AI (2023), visualisations powered by JSI using data from OpenAlex, accessed on 4/9/2023, www.oecd.ai

•https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=ai-publications-vs-gdp-per-capita-by-country-region-in-time-openalex

Akademske publikacije "per capita"

- Slovenija med boljšimi v EU
- Singapur, Norveška in Švica med vodilnimi



Explanation: This chart plots countries by: number of AI publications; GDP per capita; number of all scientific publications (bubble size); region (colour); and time (animation). To increase comparability, the "Publications per capita" checkbox divides a country's AI publications and total scientific publications by its total population.

Note: The number of total publications in a country per year is given by the size of its bubble. The 'cumulative' option displays aggregate results since 2000. Data downloads provide a snapshot in time. Due to a lag in reporting, figures for the latest quarter may appear slightly lower than they actually are. This is automatically corrected in subsequent updates. Data downloads provide a snapshot in time. Caution is advised when comparing different versions of the data, as the AI-related concepts identified by the machine learning algorithm may evolve in time. Please see methodological note for more information.

Source of data: OpenAlex
Please cite as: OECD.AI (2023), visualisations powered by JSI using data from OpenAlex, accessed on 4/9/2023, www.oecd.ai

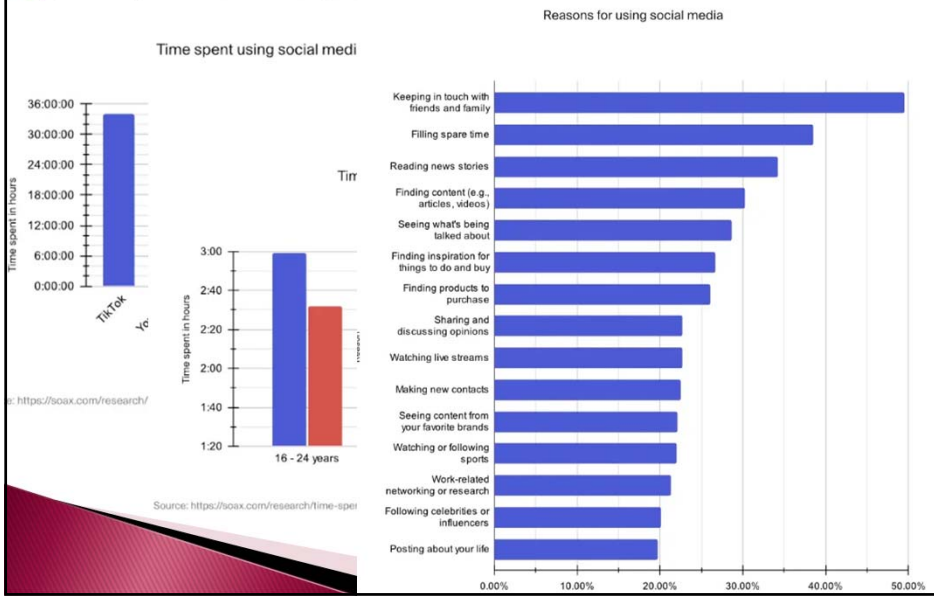
•https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=ai-publications-vs-gdp-per-capita-by-country-region-in-time-openalex

Koliko se uporablja splet in družabni mediji?

Infografike

- uporaba spleta:
  - uporabnik je v povprečju 60 ur na mesec na družabnih medijih, ženske več kot moški
  - https://www.statista.com/chart/18983/time-spent-on-social-media/
  - https://soax.com/research/time-spent-on-social-media

Where People Spend the Most & Least Time on Social Media

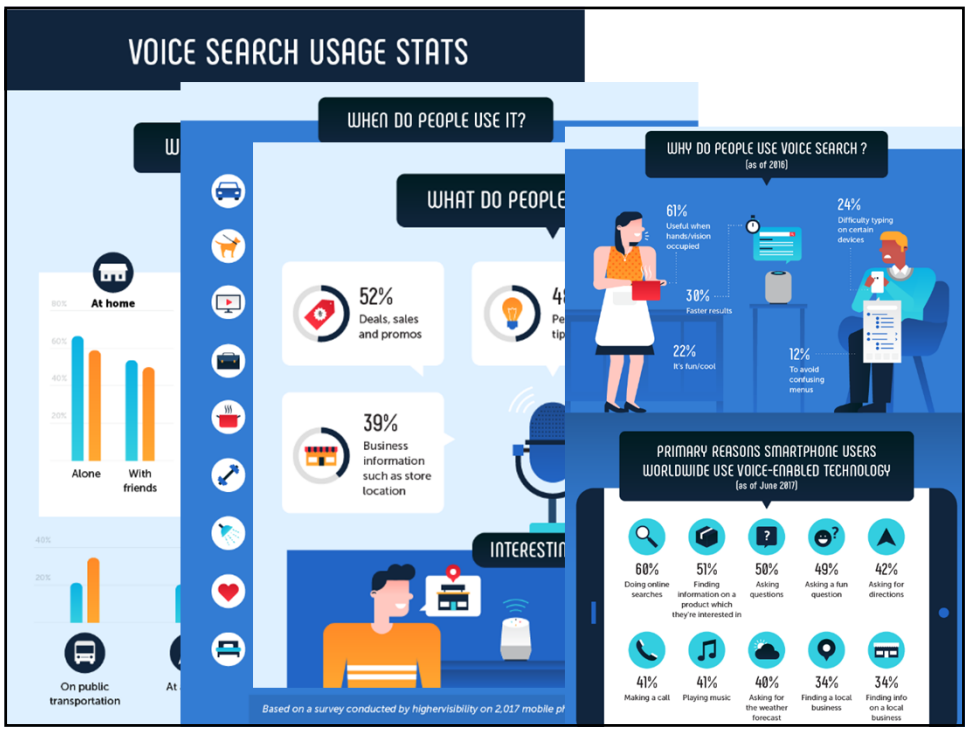


Koliko se uporablja splet in družabni mediji?

Infografike

- uporaba spleta:
  - uporabnik je v povprečju 60 ur na mesec na družabnih medijih, ženske več kot moški
  - https://www.statista.com/chart/18983/time-spent-on-social-media/
  - https://soax.com/research/time-spent-on-social-media
- uporaba glasovnega iskanja:
  - 50% glasovno iskanje med vožnjo avtomobila, 60% za spletno iskanje

https://www.socialmediatoday.com/news/106-fascinating-voice-search-facts-infographic/527108/

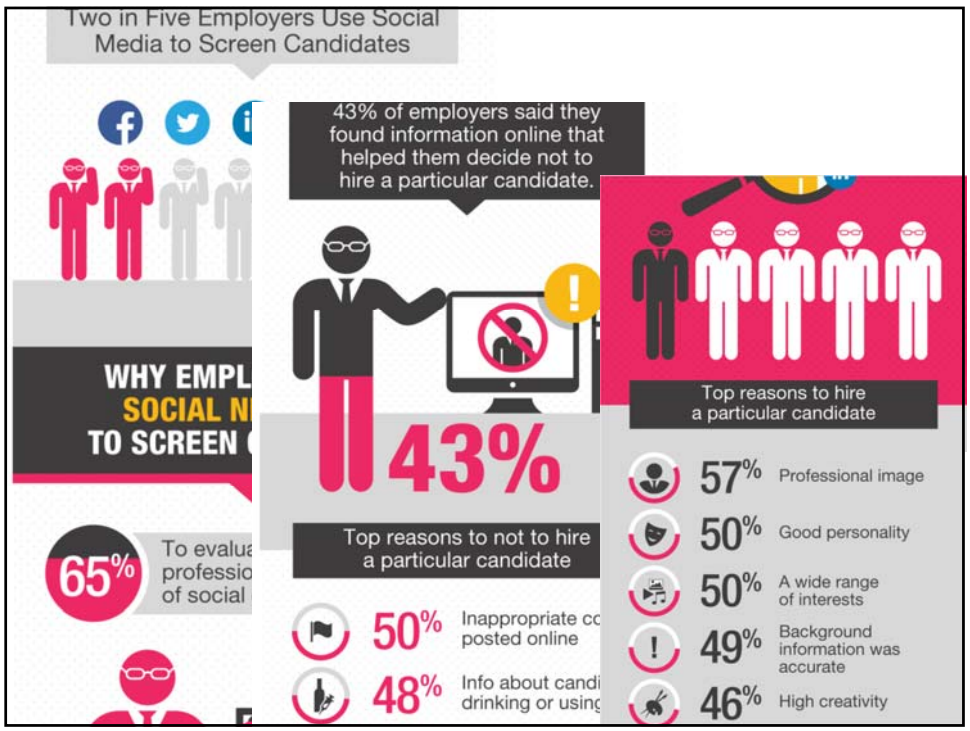


Koliko se uporablja splet in družabni mediji?

Infografike

- uporaba spleta:
  - uporabnik je v povprečju 60 ur na mesec na družabnih medijih, ženske več kot moški
  - https://www.statista.com/chart/18983/time-spent-on-social-media/
  - https://soax.com/research/time-spent-on-social-media
- uporaba glasovnega iskanja:
  - 50% glasovno iskanje med vožnjo avtomobila, 60% za spletno iskanje

  https://www.socialmediatoday.com/news/106-fascinating-voice-search-facts-infographic/527108/
- zaposlovanje - ocenjevanje kandidatov:
  - 2 od 5 delodajalcev uporablja družabna omrežja pri ocenjevanju kandidatov (43% zato ne ponudi zaposlitev)
  - https://www.go-gulf.com/blog/social-media-pre-employment-screening/

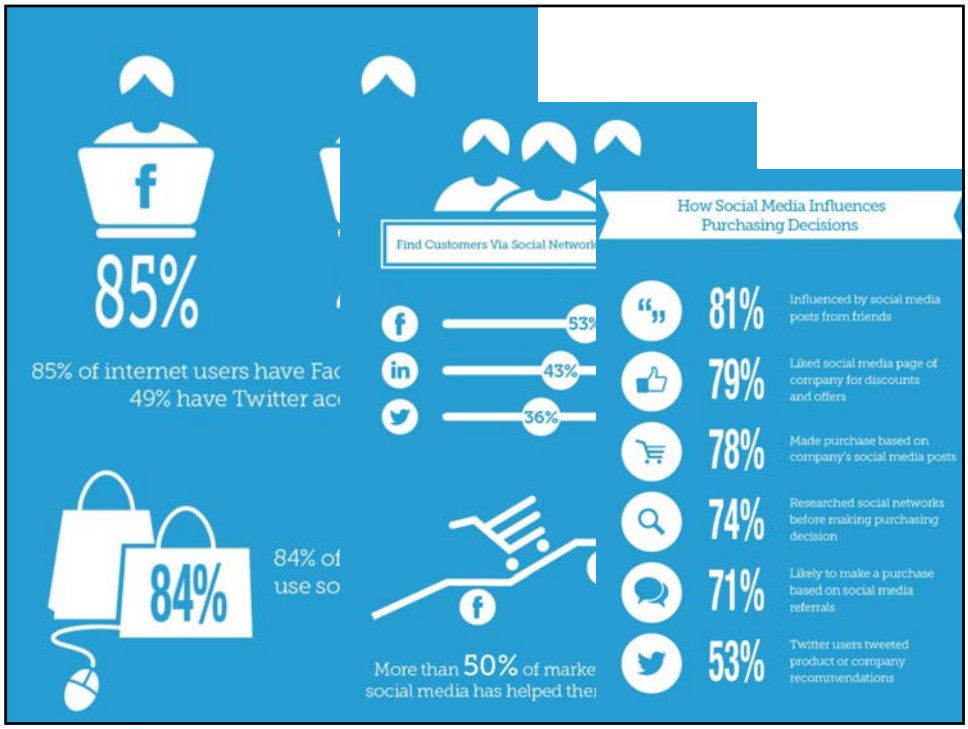


Koliko se uporablja splet in družabni mediji?

Infografike

- uporaba spleta:
  - uporabnik je v povprečju 60 ur na mesec na družabnih medijih, ženske več kot moški
  - https://www.statista.com/chart/18983/time-spent-on-social-media/
  - https://soax.com/research/time-spent-on-social-media
- uporaba glasovnega iskanja:
  - 50% glasovno iskanje med vožnjo avtomobila, 60% za spletno iskanje

  https://www.socialmediatoday.com/news/106-fascinating-voice-search-facts-infographic/527108/
- zaposlovanje - ocenjevanje kandidatov:
  - 2 od 5 delodajalcev uporablja družabna omrežja pri ocenjevanju kandidatov (43% zato ne ponudi zaposlitev)
  - https://www.go-gulf.com/blog/social-media-pre-employment-screening/
- družabni mediji:
  - 84% spletnih kupcev uporablja družabne medije
  - https://www.go-gulf.com/blog/social-media-influence-businesses/



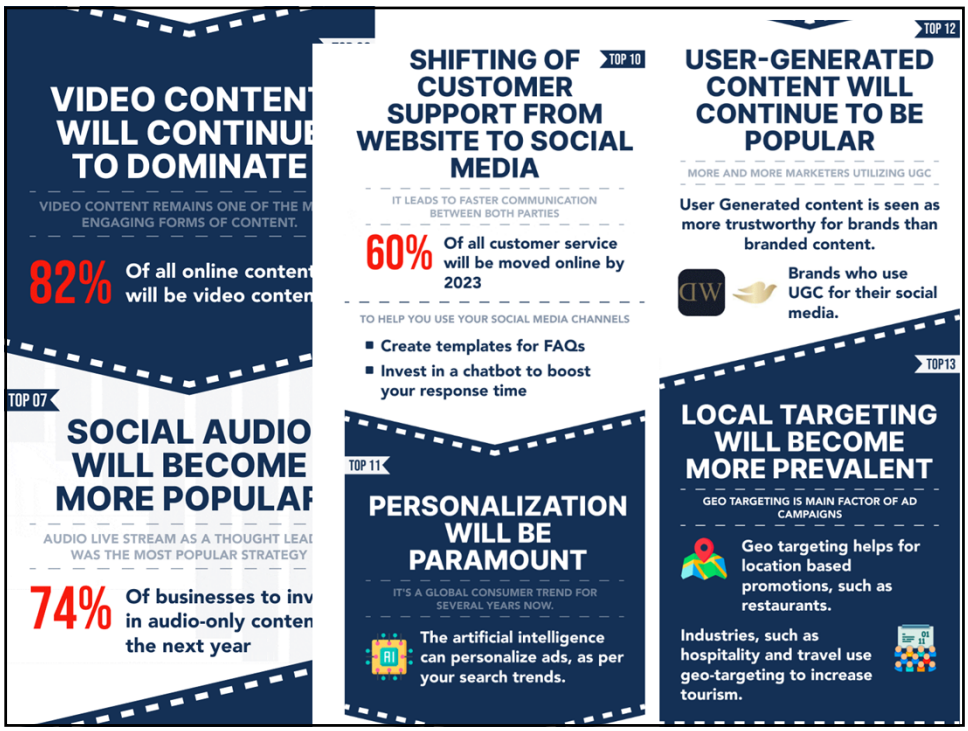
Koliko se uporablja splet in družabni mediji?

Infografike

- uporaba spleta:
  - uporabnik je v povprečju 60 ur na mesec na družabnih medijih, ženske več kot moški
  - https://www.statista.com/chart/18983/time-spent-on-social-media/
  - https://soax.com/research/time-spent-on-social-media
- uporaba glasovnega iskanja:
  - 50% glasovno iskanje med vožnjo avtomobila, 60% za spletno iskanje

  https://www.socialmediatoday.com/news/106-fascinating-voice-search-facts-infographic/527108/
- zaposlovanje - ocenjevanje kandidatov:
  - 2 od 5 delodajalcev uporablja družabna omrežja pri ocenjevanju kandidatov (43% zato ne ponudi zaposlitev)
  - https://www.go-gulf.com/blog/social-media-pre-employment-screening/
- družabni mediji:
  - 84% spletnih kupcev uporablja družabne medije
  - https://www.go-gulf.com/blog/social-media-influence-businesses/
  - 82% spletnih vsebin bodo videji

  https://www.go-globe.com/social-media-trends-in-2022-infographic/



Uvod

- Odkrivanje zakonitosti v spletnih podatkih
- Odkrivanje zakonitosti v tekstovnih podatkih
- Pristop raziskovalnim problemom
- Primeri nalog

Odkrivanje zakonitosti v spletnih podatkih (Web Mining)

- Kaj:
  - iz informacij na spletu pridobiti znanje o spletu!
- Kako:
  - uporabimo metode odkrivanja zakonitosti v podatkih
- Zakaj:
  - cilj je razumeti kompleksne, dinamične podatke v povezavi s spletom za boljše razumevanje, učinkovitejše poslovanje,...

Poleg analize besedil, najbolj tipičen problem je analiza in profiliranje strank na osnovi dnevnika na spletnih strežnikih

Analiza spletnih podatkov

Sestavljena iz treh osnovnih pogledov na splet:

- analiza vsebine na spletu (anlg. Web content mining)
- analiza strukture spleta (anlg. Web structure mining)
- analiza interakcij uporabnikovih s spletom (anlg. Web usage mining)

Potrebno je upoštevati privatnost

- pri podatkih o uporanikovi interakciji s spletom
- pri združevanju podatkov iz različnih virov, ki so bili zbranih za drugačne namene

Uvod

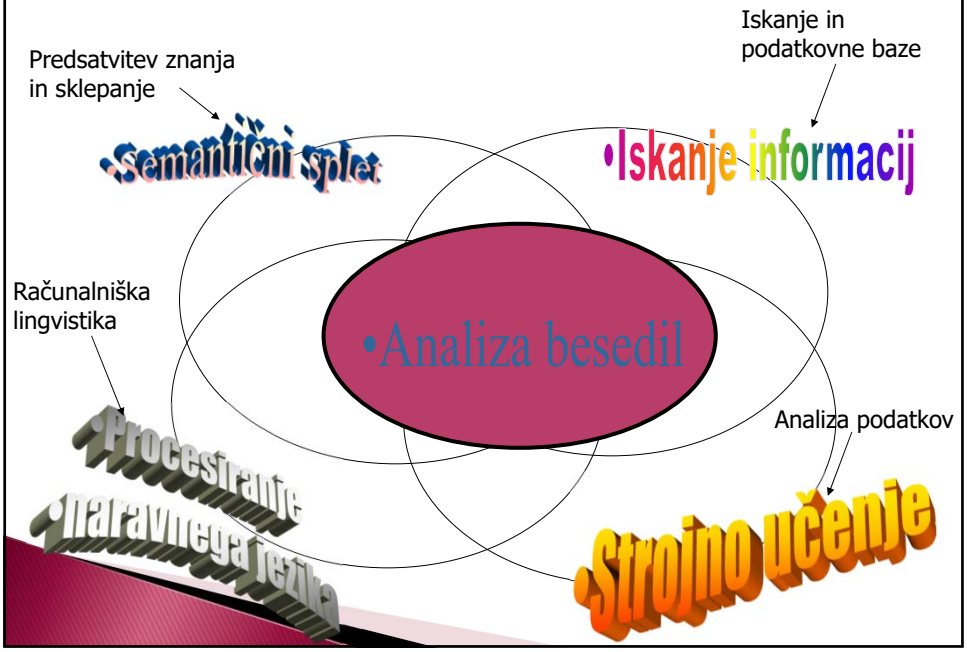
- Odkrivanje zakonitosti v spletnih podatkih
- Odkrivanje zakonitosti v tekstovnih podatkih
- Pristop raziskovalnim problemom
- Primeri nalog

Odkrivanje zakonitosti v tekstovnih podatkih (Text Mining)

Odkrivanje zakonitosti v tekstovnih podatkih lahko definiramo kot

- proces identifikacije veljavnih, novih, potencialno uporabnih in razumljivih vzorcev v tekstovnih podatkih
- iskanje pomena/vsebine tekstovnih podatkov (semanitčne in abstraktne informacije)

Analiza tekstovnih podatkov (besedil)



Uvod

- Odkrivanje zakonitosti v spletnih podatkih
- Odkrivanje zakonitosti v tekstovnih podatkih
- Pristop raziskovalnim problemom
- Primeri nalog

Ustvarjalnost v raziskovanju

Sledimo splošnemu ustvarjalnemu procesu:

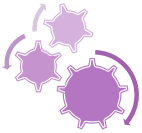
- namen začne proces
- zavest prinese fokus in razširi vpogled
- inteligenca zagotovi pogoje
- energija zagotovi praktično izvedbo

Uporabljeno na raziskovalnem procesu:

- ustvarjalni navdih
- vpogled v podrobnosti in možne posledice
- formulacija in analiza
- implementacija

Pristop raziskovalnim problemom

- Ideja/namen – intuicija
  - Pogledamo idejo v širšem kontekstu
- Strategija – intelektualna analiza
  - Analiza praktičnih aspektov
    - domena, razpoložljivi podatki, obstoječe metode
  - Identifikacija potrebnih korakov
    - vključno z viri (znanje, oprema, čas, ljudje,…)
- Implementacija – praktična akcija
  - Razvoj pristopa
  - Evalvacija in izboljšave
  - Refleksija na proces in izkušnje



Pristop razisk emom

izboljšati učinkovitost in dobro počutje uporabnika

- Ideja/namen – intuicija
  - Pogledamo idejo v širšem kontekstu

mobilna naprava daje uporabniku prilagojene nasvete za aktivnosti

- Strategija – intelektualna analiz
  - Analiza praktičnih aspektov
    - domena, razpoložljivi podatki,
  - Identifikacija potrebnih korakov
    - vključno z viri (znanje, oprema, čas, ljudj

ciljamo vodilne ljudi, uporabimo demografske podatke, znanja in veščine, dnevne aktivnosti

baza znanja o aktivnostih, povratna informacija,…

- Implementacija – praktična akcija
  - Razvoj pristopa
  - Evalvacija in izboljšave
  - Refleksija na proces in izkušnje

nenadzorovano učenje profila uporabnika, semantično označene aktivnosti, aktivno učenje nasvetov

vprašalnik za validacijo, vizualizacija uporabniških profilov

Delo v skupinah – pristop raziskovalnim problemom

- Razdelimo se v skupine
- Vsaka skupina:
  - identificira en zanimiv problem, ki vključuje analizo besedila v izobraževanju
- Skupaj
  - Izberemo en ali dva problema
- Vsaka skupina:
  - nakaže strategijo za reševanje izbranega problema
  - predlaga možno smer implementacije
- Skupaj
  - ocenimo zanimivost, izvedljivost in učinkovitost

Pristop raziskovalnim problemom

Opredelimo glavne gradnike pri analizi podatkov

- Naloga, ki jo rešujemo
- Predstavitev podatkov
- Tehnike analize podatkov

Naloga, ki jo rešujemo

- Iskanje informacij (besedila, slike, video, zvok)
- Modeliranje podatkov - (ne)nadzorovano učenje
- Sumarizacija
- Strojno prevajanje
- Vizualizacija
- …

Predstavitev podatkov

- Besedilo lahko predstavimo na različnih nivojih podrobnosti
  - črke, besede, … pravila v logiki
- Jezik v kateremu je napisano besedilo
  - ene jezik, več jezikov neodvisno, preko-jezično
- Besedilo v kombinaciji z drugimi podatki
  - Večmodalna predstavitev podatkov

Tehnike analize podatkov

- Ročna analiza, strojno učenje, logično sklepanje
- Kompromis med prostorom (za shranjevanje podatkov) in zakasnitvami (hitrost obdelave)
- Upoštevamo zahtevano kvaliteto rešitve, potrebne vire, standarde, …

Uvod

- Odkrivanje zakonitosti v spletnih podatkih
- Odkrivanje zakonitosti v tekstovnih podatkih
- Pristop raziskovalnim problemom
- Primeri nalog

Primeri nalog analize besedil

- Filtriranje tekstovnih informacij
- Pomoč pri preiskovanju svetovnega spleta
- Analiza in izdelava profilov uporabnika
- Identifikacija jezika v besedilu
- Ugotavljanje avtorstva dokumentov
- Ugotavljanje kopiranja dokumentov

Profiliranje uporabnikov

- Uporabnika, ki prebira besedila na elektronskih medijih je mogoče opazovati in beležiti njegove akcije
- Na osnovi njegovih akcij lahko naredimo profil obnašanje posameznika ali profil obnašanja populacije uporabnikov
  - Možna izdelava različnih profilov (npr. vsebinski interesi, emotivni profil, pojavitev imenskih entitet)
- Primeri beleženja akcij so:
  - dnevnik spletnega strežnika
  - elektronska knjiga
  - osebni spletni agent (Personal WebWatcher)

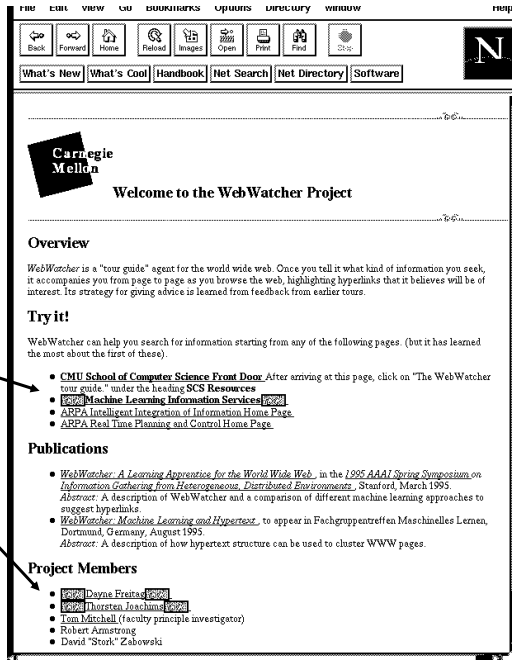
Primer osebnega spletnega agenta Personal WebWatcher

- Uporabnika preiskuje svetovni splet
- Sistem "opazuje" uporabnika, shranuje dokumente, ki jih uporabnik pogleda in gradi profil uporabnikovih interesov

Profil uporabnika sistemu omogoča, da **na zahtevanih dokumentih označi zanimive hiperpovezave,** poišče za uporabnika zanimive spletne strani, primerja uporabnike in med njimi izmenjuje informacije (collaborative filtering), združuje profile iz naslova različnih aktivnosti (splet, disk, e-koledar, novice,e-mail)

Personal WebWatcher na delu

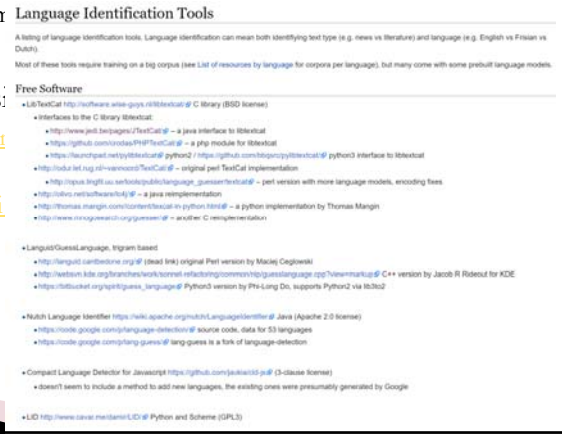
Označi zanimive hiperpovezave



Identifikacija jezika v besedilu

- Problem:
  - za dano besedilo ugotovi v katerem od sistemu znanih jezikov je napisano
- Osnovni postopek za ugotavljanje so preprosti:
  - za vsak jezik izdelamo karakteristično frekvenčno tabelo parov ali trojic črk
  - na osnovi tabele razpoznan

Primer je javno dostopen s
https://www.let.rug.nl/~van
Seznam različnih sistemov
https://aclweb.org/aclwiki


Language Identification Tools

A listing of language identification tools. Language identification can mean both identifying text type (e.g. news vs literature) and language (e.g. English vs Frisian vs Dutch).

Most of these tools require training on a big corpus (see List of resources by language for corpora per language), but many come with some prebuilt language models.

Free Software

- LibTextCat http://software.wise-guys.nl/libtextcat/ C library (BSD license)
  - interfaces to the C library libtextcat:
    - http://www.jweb.be/pages/TextCat/ – a java interface to libtextcat
    - https://github.com/cindas/PHPTextCat – a php module for libtextcat
    - https://launchpad.net/pylibtextcat python2 / https://github.com/bbqsrc/pylibtextcat python3 interface to libtextcat
  - http://odur.let.rug.nl/~vannoord/TextCat/ – original perl TextCat implementation
    - http://opus.tiedl.uu.se/tools/public/language_guesser/textcat – perl version with more language models, encoding fixes
  - http://olivo.net/software/lc4j/ – a java reimplementation
  - http://thomas.mangin.com/content/textcat-in-python.html – a python implementation by Thomas Mangin
  - http://www.mnogosearch.org/guesser/ – another C reimplementation

- Languid/GuessLanguage, trigram based
  - http://languid.cantbedone.org/ (dead link) original Perl version by Maciej Ceglowski
  - http://websvn.kde.org/branches/work/sonnet-refactoring/common/nlp/guesslanguage.cpp?view=markup C++ version by Jacob R Rideout for KDE
  - https://bitbucket.org/spirit/guess_language Python3 version by Phi-Long Do, supports Python2 via lib3to2

- Nutch Language Identifier https://wiki.apache.org/nutch/LanguageIdentifier Java (Apache 2.0 license)
  - https://code.google.com/p/language-detection/ source code, data for 53 languages
  - https://code.google.com/p/lang-guess/ lang-guess is a fork of language-detection

- Compact Language Detector for Javascript https://github.com/jaukia/cld-js (3-clause license)
  - doesn't seem to include a method to add new languages, the existing ones were presumably generated by Google

- LID http://www.cavar.me/damir/LID/ Python and Scheme (GPL3)

Delo v skupinah – poenostavljena identifikacija jezika besedil na osnovi besed

- Razdelimo se v skupine
- Vsaka skupina dobi:
  - primer besedila v neznanem jeziku
  - besedila v treh znanih jezikih
- Glede na skupne besede določi jezik besedila
  - predstavi rezultate
- Skupaj
  - pojasnimo razlike in podobnosti med jeziki

Primer – enostavno, na osnovi besed

- **English**
  - You do not have to compete with anybody. You have to compete with yourself.
- **Italiano**
  - Il leader è responsabile dell'evoluzione dei propri collaboratori come professionisti e come esseri umani.
- **Slovensko**
  - To zahteva seminar in osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.
- **Neznano**
  - Come to "Leadership Potential" seminar for yourself.

Primer – enostavno, na osnovi besed

- **English**
  - You do not have to compete with anybody. You have to compete with yourself.
- **Italiano**
  - Il leader è responsabile dell'evoluzione dei propri collaboratori come professionisti e come esseri umani.
- **Slovensko**
  - To zahteva seminar in osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.
- **Neznano**
  - Come to "Leadership Potential" seminar for yourself.

•Come to "Leadership Potential" **seminar** for yourself. (1, **1**, **1**)
•1. yourself
•2. come
•3. seminar

Primer – na osnovi besed

- **English**
  - In information age, 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. Remember this.
- **Italiano**
  - La vostra capacità di leadership può essere valutata osservando le capacità, la creatività, la qualità della vita, la dedizione e l'impatto dei vostri collaboratori. Il leader è responsabile dell'evoluzione dei propri collaboratori come professionisti e come esseri umani.
- **Slovensko**
  - Dandanes so pritiski okolja postali praktično nevzdržni. Dela imamo vedno več, časa vedno manj. Konkurenca na trgu je vedno bolj neizprosna, zakonodaja težko obvladljiva. Osnova za uspeh v takih okoliščinah zahteva zavestnega vodjo. To zahteva seminar in osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.

Neznano

- Remember to book yourself a flight to come in our leadership seminar.
- Leadership is a difficult skill to master, you cannot just look in a book.
- Fly to come in leadership seminar.

Primer – na osnovi besed

- **English**
  - **In** information age, 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**. Remember this.
- **Italiano**
  - La vostra capacità di leadership può essere valutata osservando le capacità, la creatività, la qualità della vita, la dedizione e l'impatto dei vostri collaboratori. Il leader è responsabile dell'evoluzione dei propri collaboratori come professionisti e come esseri umani.
- **Slovensko**
  - Dandanes so pritiski okolja postali praktično nevzdržni. Dela imamo vedno več, časa vedno manj. Konkurenca na trgu je vedno bolj neizprosna, zakonodaja težko obvladljiva. Osnova za uspeh v takih okoliščinah zahteva zavestnega vodjo. To zahteva osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.

Neznano

- Remember to book yourself a flight to come in our leadership seminar.
- Leadership is a difficult skill to master, you cannot just look in a book.

•**Remember to book yourself** a flight **to come in** our **leadership** seminar. (5, **2**, **2**)
•**Leadership** is a difficult skill **to** master, **you** cannot just look **in** a **book**. (5, **1**, **2**)
•Fly **to come in leadership seminar**. (2, **2**, **3**)

Ugotavljanje avtorstva

- **Problem:**
  - danemu dokumentu moramo prirediti najverjetnejšega avtorja, če imamo predhodno dano bazo dokumentov kjer so avtorji že znani
- **Rešitev:**
  - temelji na tem, da posamezni avtor uporablja značilno frekvenčno sliko besed in fraz

Ugotavljanje kopiranja dokumentov

- **Problem:**
  - za dan dokument moramo ugotoviti ali je bil z neko verjetnostjo izdelan deloma ali v celoti na osnovi enega ali več dokumentov iz baze dokumentov
- **Rešitev:**
  - postopek uporablja zahtevne indeksne metode nad deli besedila različnih velikosti, ki jih uporabi za primerjavo z deli besedila danega dokumenta

Uporabnost in razumljivost na spletu

- **Seeing Vs. Noticing: Helping users find and use information quickly,** Prasad Kantamneni, Yahoo! http://videolectures.net/eswc2011_kantamneni_helping/ (54min)

Delo v skupinah – identifikacija jezika besedil

- Razdelimo se v skupine
- Vsaka skupina dobi besedilo v enem jeziku
  - za jezik izdela frekvenčno tabelo trojic črk
  - predstavi rezultate
- Skupaj
  - pojasnimo razlike med jeziki
  - identificiramo jezik novega besedila z uporabo zgrajenih frekvenčnih tabel

Besedila - na osnovi trojic črk

- **English**
  - In information age, 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.** Remember this.

You have to compete with yourself.

you	hav	ave	com	omp	mpe	pet	ete	wit	ith	our	urs	rse	sel	elf
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1

come professionisti e come esseri umani.

Il leader **è responsabile dell'evoluzione dei propri collaboratori come professionisti e** come esseri umani.

- **Slovensko**
  - Dandanes so pritiski okolja postali praktično nevzdržni. Dela imamo vedno več, časa vedno manj. Konkurenca na trgu je vedno bolj neizprosna, zakonodaja težko obvladljiva. Osnova za uspeh v takih okoliščinah zahteva zavestnega vodjo. **To zahteva seminar in osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.**

To zahteva seminar in osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.

Primer – na osnovi trojic črk

- **English**
  - In information age, 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. Remember this.
- **Italiano**
  - La vostra capacità di leadership può essere valutata osservando le capacità, la creatività, la qualità della vita, la dedizione e l'impatto dei vostri collaboratori. Il leader è responsabile dell'evoluzione dei propri collaboratori come professionisti e come esseri umani.
- **Slovensko**
  - Dandanes so pritiski okolja postali praktično nevzdržni. Dela imamo vedno več, časa vedno manj. Konkurenca na trgu je vedno bolj neizprosna, zakonodaja težko obvladljiva. Osnova za uspeh v takih okoliščinah zahteva zavestnega vodjo. To zahteva seminar in osebo, ki bo s svojo modrostjo in znanjem popeljala skupino in podjetje na pot uspeha.

Neznano

- Come to "Leadership Potential" seminar for yourself

•Come to "Leadership Potential" seminar. (7, 6, 7)
•1. com (computer, compete – come), you, our, urs, rse, sel, elf (you, yourself – yourself)
•2. com, ome, lea, ead, ade, der (come – come, leader – leadership)
•3. pot (pot – potential), sem, emi, min, nar (seminar – seminar)

Seminarske

Računalniška analiza besedil v izobraževanju

- Folije s predavanj (PDF) (17. 10. 2023).

http://capybara.ijs.si/janez/teaching/pef.html

Seminarske naloge

Predstavitev literature

Naloga: preberite enega od naslednjih člankov ali si oglejte enega od naslednjih posnetkov predavanj; o članku oz. predavanju pripravite 10-minutno predstavitev.

Seznam člankov in posnetkov:

Kdo	Članek oz. predavanje
	M. Karlovčec, B. Lužar, D. Mladenić. Core-periphery dynamics in collaboration networks: the case study of Slovenia. *Scientometrics*, 2016. DOI: 10.1007/s11192-016-2154-4.
	Povezava med zavestjo in umetno inteligenco, Marko Grobelnik, IJS (50 min)
	Observing odor-related information in academic domain, Inna Novalija (10 min)
	Observing Water-Related Events for Evidence-Based Decision-Making, Alenka Guček (14 min)
	Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information, Pasquale Lisena (24 min)
	Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification, Taja Kuzman (15 min)
	SLOnet — Slovenian Commonsense Description, Adrian Mladenić Grobelnik (15 min)
	Emotion Recognition in Text using Graph Similarity Criteria, Nadezhda Komarova (15 min)

Skupinski projekt: Poskusi na zbirki besedil

Naloga: s pomočjo orodij iz paketa TextGarden izvedite na zbirki besedil naslednje korake:

[**Novo:** pripravil sem na novo prevedeno verzijo TextGardna. Tisti, ki ste imeli z dosedanjo verzijo težave, poskusite z novo: TextGarden2021.zip. V datoteki .zip je tudi Microsoftov `vcredist_x86.exe`, ki ga **poženite najprej**, da vam inštalira morebitne manjkajoče DLLje. Če boste imeli še vedno težave, mi pišite.]

[Če bodo z zgornjo verzijo težave, poskusite še z naslednjima starejšima verzijama TextGardna: 1, 2.]

- S programom **Txt2Bow** pripravite predstavitev v formatu bag-of-words.
- S programom **BowKMeans** pripravite razvrstitev dokumentov v skupine po metodi *k*-means za dva različna *k* in analizirajte rezultate.
- S programom **BowTrainBinSVM** pripravite klasifikatorje za dve redki in dve pogosti kategoriji. Te klasifikatorje s programom **BowClassify** uporabite na besedilih iz naslednje zbirke (če ste učili klasifikatorje na zbirki *n*, poženite klasifikatorje na besedilih iz zbirke *n* + 1). Izračunajte precision, recall, F_1, klasifikacijsko točnost in narišite krivuljo precision/recall.
  - `BowClassify` izračuna napoved le za en dokument naenkrat. Zato si boste morali napisati program ali skripto, ki datoteko z vašo testno množico bere po vrsticah, iz vsake vrstice izloči dokument, ki je v njej (torej poreže zaporedno število dokumenta in oznake kategorij z začetka vrstice), ga izpiše v samostojno pomožno datoteko, pokliče `BowClassify` in nato prebere njegovo napoved ter si jo nekam shrani, da jo bo kasneje lahko primerjal s pravo pripadnostjo dokumenta kategoriji in na podlagi tega računal precision, recall itd.
  - **Opomba:** `BowClassify` se sesuje, če je bil klasifikator naučen na kategoriji, ki ni imela nobenih negativnih učnih dokumentov; to se pri nekaterih naših zbirkah besedil lahko zgodi npr. s kategorijo GCAT, zato je v svojih poskusih raje ne uporabljajte.
  - **Kako narišemo krivuljo precision/recall?** Testne dokumente uredimo padajoče po vrednosti, ki jo napove klasifikator (s tem je mišljena napoved kot realno število – npr. če je BowClassify pri nekem dokumentu izpisal `_1: -0.926 'GPOL'`", je njegova napoved tuka} -0.926), nato pa se vprašamo: „kakšen precision in recall dobimo, če razglasimo prvih *k* dokumentov za pozitivne, ostale pa za negativne?" To naredimo za vsak *k* od 0 do števila dokumentov in dobljene pare (precision, recall) narišemo na grafu, kjer postavimo precision na *y*-os, recall pa na *x*-os. Takšna krivulja je zanimiva, ker nam pove, kaksne tradeoffe med precisionom in recallom nam ta model omogoča, če smo pripravljeni spremeniti prag, nad katerim napovedujemo pozitivni razred.
- Prejšnji korak izvedite za dve vrsti jeder: linearna (parameter `-t:linear`) in polinomska s stopnjo $d = 3$ (`-t:polynomial -ker_p:3`).
- Za vsako od uporabljenih štirih kategorij poiščite na internetu po eno besedilo, ki ga klasifikator za tisto kategorijo napove kot pozitiven primer.

Uporabo orodij iz paketa TextGarden si bomo ogledali na vajah **7. novembra 2023** in **21. novembra 2023**.

Pripravite poročilo (~10 strani) in 15-minutno predstavitev. (Pri predstavitvi naj nastopijo vsi člani skupine!)

Rok za oddajo: poročila in predstavitve (PowerPoint ali PDF) pošljite na janez.brank@ijs.si do **2. januarja 2024**. Predstavitve bodo potekale na predavanjih **9. januarja 2024**.

Povzetek dosedanje snovi

Uvod: kaj in kdaj

- Kaj je odkrivanje zakonitosti v spletnih podatkih, katere podprobleme zajema?
- Kaj je odkrivanje zakonitosti v tekstovnih podatkih, katera področja združuje?
- Proces reševanja raziskovalnih problemov
- Primeri nalog: filtriranje informacij, preiskovanje spleta, profiliranje uporabnikov, identifikacija jezika besedila, avtorstvo in kopiranje dokumentov

Prvi del: principi strojnega učenja

- Strojno učenje
- Modeliranje podatkov
- Nadzorovano učenje
- Primeri algoritmov

Kaj je strojno učenje?

- Herbert Simon: "Učenje je vsak proces, s katerim sistem na podlagi izkušenj izboljša svoje delovanje."
- Boljše opravljanje naloge T, glede na določeno mero uspeha P, glede na izkušnje E:
  - T: Prepoznavanje ročno napisanih besed
  - P: Odstotek pravilno razvrščenih besed
  - E: Podatkovna zbirka slik ročno napisanih besed, ki jih je označil človek

  - T: Vožnja po štiripasovnih avtocestah z uporabo senzorjev vida
  - P: Povprečna prevožena razdalja pred napako, ki jo oceni človek
  - E: Zaporedje slik in ukazov za krmiljenje, posnetih med opazovanjem človeškega voznika

  - T: Razvrščanje e-poštnih sporočil med neželeno pošto ali legitimna sporočila
  - P: odstotek pravilno razvrščenih e-poštnih sporočil
  - E: Podatkovna baza elektronskih sporočil, nekatera z oznakami, ki jih je določil človek

Postopek za strojno učenje

- Tipično postopek strojnega učenja vključuje:
  - **Vhodne podatke** v neki obliki (n.pr., vektorji številk)
  - **Vhodni parametri** (nastavitev)
  - **Modeliranje podatkov** (z uporabo nekega postopka presiskovanja prostora možnih modelov/rešitev)
  - **Izhod – model** v obliki *y=f(x)*
  - Uporaba modela za **klasifikacijo** novih podatkov
- …ključni elementi:
  - **Jezik za opis modela** – določa zahtevnost modela
    - …pogosto uporabljani so linearni modeli (SVM, Perceptron, Bayes, ...): $y = a\,x1 + b\,x2 + c\,x3 + \ldots$
  - **Preiskovalni algoritem** – določa kvaliteto rezultatov
    - …najboljpogosto je to algoritem za lokalno optimizacijo

Kakšne probleme rešujemo?

- **analiza podatkov in odkrivanje zakonitosti v podatkih** (podatkovne baze, besedila, večpredstavni podatki) ‘ iščemo zakonitosti in vzorce v podatkih
- Rezultat je model, ki ga lahko razumemo kot povzetek podatkov
- Model lahko uporabljamo za:
  - **Razlago** obstoječih pojavov v podatkih
  - **Napovedovanje** prihodnih situacij

Prvi del: principi strojnega učenja

- Strojno učenje
- Modeliranje podatkov
- Nadzorovano učenje
- Primeri algoritmov

Modeliranje podatkov

- Predstavitev podatkov
  - Atributna predstavitev podatkov – vektor spremenljivk
  - Vsak primer iz podatkov predstavimo kot vektor, vsak atribut ima eno izmed možnih vrednosti tega atributa
  - Lahko generiramo nove atribute ali obdržimo le nekatere izmed atributov

transformacija ali kombinacija

izbira podmnožice atributov

Primer – opis risanke

Mojster Miha

vozila = da
ljudje = da
živali = ne

vektor = [1, 1, 0]

Atributi:

- vozila[da, ne]
- ljudje [da, ne]
- živali [da, ne]

Osnovni pristopi

Kdaj uporabiti kateri pristop?

- **Nadzorovano učenje** (klasifikacija, uvrščanje)
  - …imamo podane opise risank in oznake ali je risanka zanimiva za mlajše otroke ali ne; cilj je poiskati pravila s katerimi lahko napovemo zanimivost poljubne riskanke za manjše otroke
- **Pol-nadzorovano učenje**
  - …imamo podane opise risank in oznake zanimivosti **vendar samo za nekatere riskanke**; cilj je poiskati pravila s katerimi lahko napovemo zanimivost poljubne riskanke za manjše otroke in ob tem uporabiti tudi neoznačene podatke
- **Nenadzorovano učenje** (razvrščanje)
  - …imamo podane opise risankč cilj je poiskati skupin epodobnih risank

Prvi del: principi strojnega učenja

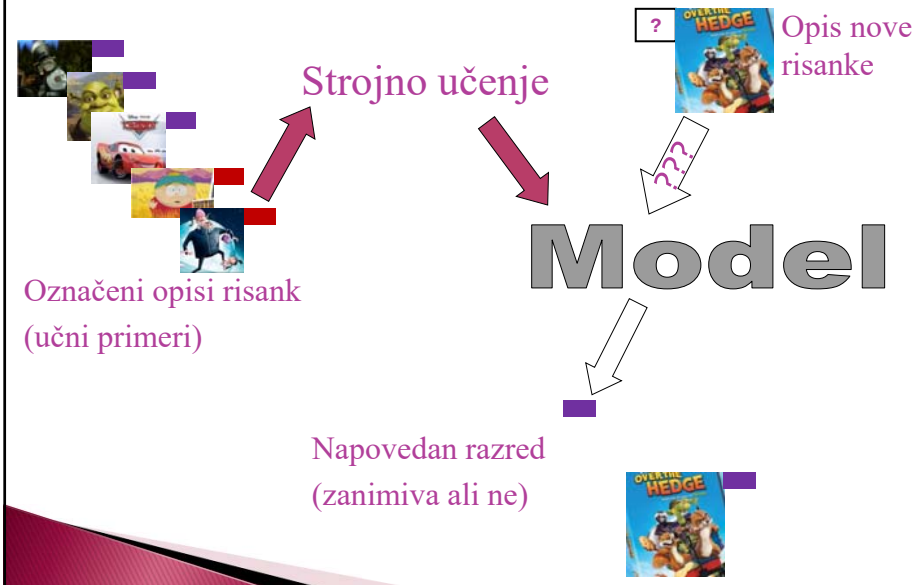
- Strojno učenje
- Modeliranje podatkov
- Nadzorovano učenje
- Primeri algoritmov

Nadzorovano učenje

Uvrstimo primer v množico predefiniranih razredov:

- Medicinska diagnostika
  - …pacijentu priredimo diagnozo
- Kreditne vloge klientov
  - …ocenimo primernost kreditne vloge
- Odkrivanje prevar v e-poslovanju
  - …napoved ali je nek poslovni dogodek prevara
- Finančne investicije
  - …nasvet ali naj kupimo, prodamo ali počakamo z delnicam na borzi
- Filtriranje nezaželjenih e-sporočil
  - …napoved ali je e-sporočilo nezaželjeno (spam)
- Predlaganje zanimivih člankov iz spletnih novic
  - …oceimo koliko bo članek zanimiv za uporabnika na osnovi model auporabnikovih interesov

Napoved zanimivosti risanke



Nadzorovano učenje

Podana: množica označenih primerov predstavljenih z vektorjem atributov

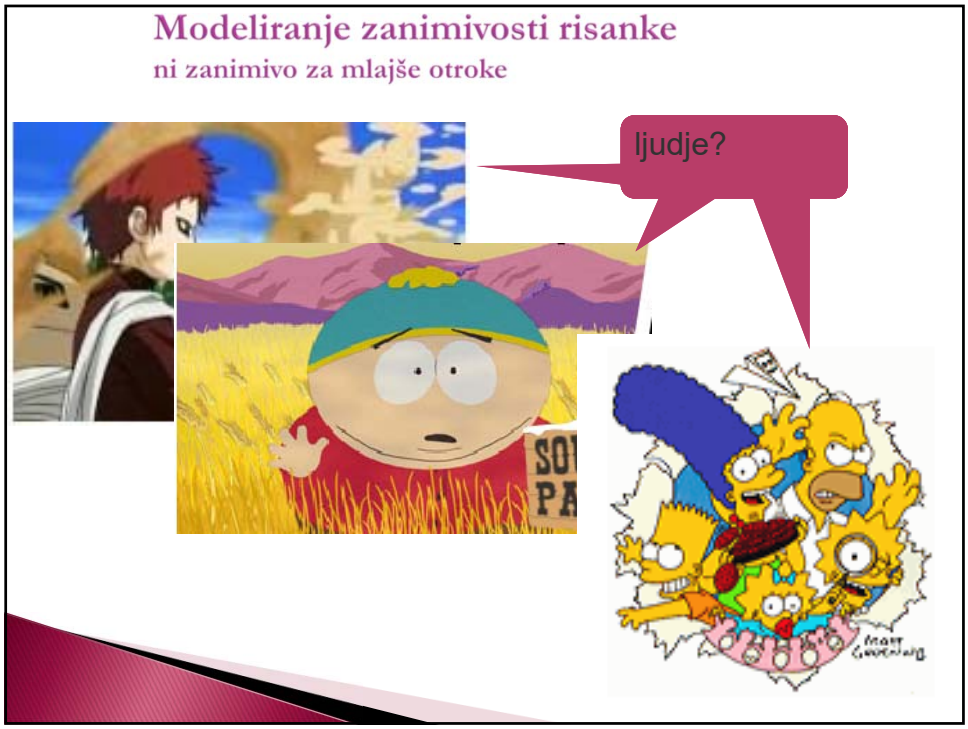
Cilj: zgraditi model ciljne funkcije, ki bo dodelil pravilno oznako neoznačenim primerom

- Vrednosti atributov:
  - diskretne (n.pr., barva_oči ∈ {modra, zelena, rjava, siva})
  - zvezne (n.pr., starost ∈ [0..200])
  - urejene (n.pr., velikost ∈ {majhen, srednji, velik})
- Vrednosti ciljne funkcije – oznake:
  - diskretne (klasifikacija) ali zvezne (regresija)
  - izključujoče (n.pr., medicinska diagnoza) ali ne (n.pr., isto besedilo govori lahko o več vsebinskih področji)
  - Medsebojno v relacijah (vsebinske taksonomije dokumentov, n.pr., DMoz)

Ciljna funkcija je lahko:

- predstavljena na različne načine (spravljeni primeri, simbolični model, numerično, grafični model,…)
- modelirana z uporabo različnih algoritmov


kratka?
Modeliranje zanimivosti risanke zanimivo za
mlajše otroke
dolga & živali?
vozila?
živali ali vozila?
dolga & živali?


Modeliranje zanimivosti risanke
ni zanimivo za mlajše otroke
ljudje?

Modeliranje zanimivosti risanke - vektorji atributov

Naslov	Junaki	Trajanje
Bob the builder	vozila, ljudje	10 min
Pixar-Locomotion	vozila	5 min
Ice age	živali	90 min
Over the hedge	živali	60 min
Cars	vozila	90 min
Anima	ljudje	90 min
South Park	ljudje	30 min
Simpson	ljudje	20 min

Ciljna funkcija

Večja izrazna moč predstavitve funkcije omogoča boljšo aproksimacijo dejanskega dogajanje

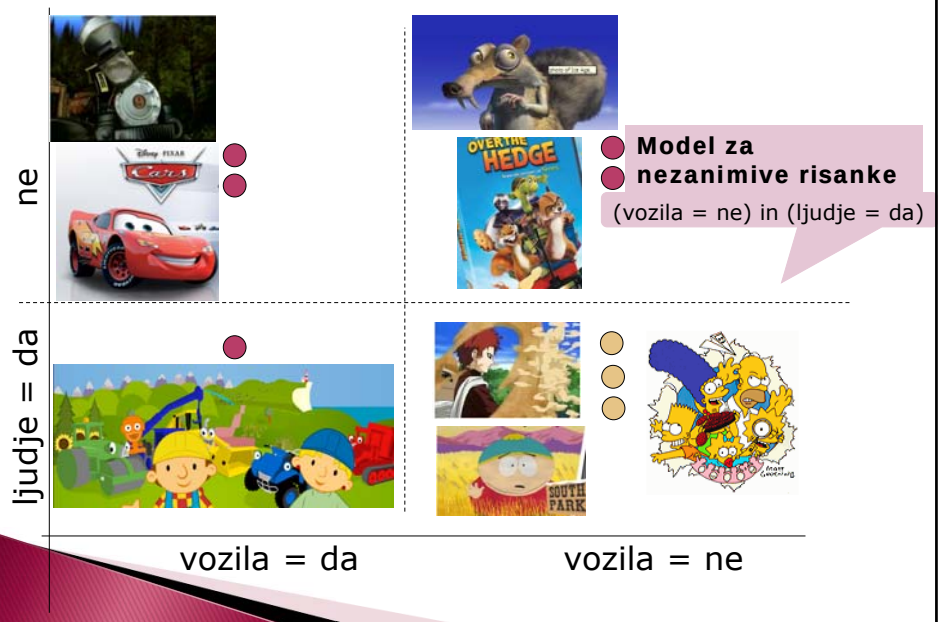
- vendar se je težje naučimo, potrebujemo več primerov za natančno aproksimacijo funkcije
- želimo čim bolj enostavno predstavitev, ki je še dovolj močna za problem, ki ga rešujemo

Primer modeliranja zanimivosti risanke

Vrednosti ciljne funkcije: diskretne oznake (klasifikacija), izključujejo se

Risanka zanimiva za mlajše otroke: da | ne

Primer vizualizacije podatkov



Posplošitev podatkov

- Model mora zadosti posplošiti podatke, da omogoči klasifikacijo novih primerov (takšnih, ki se še niso pojavili v učnih podatkih)
- Tabela s primeri na vpogled, ne omogoča posplošitev
  - nov primer, ki ni bil v učnih podatkih ne bomo znali klasificirati

Occamova britev:

- Želimo čim bolj enostaven (splošen) model, ki še vedno omogoča korekten opis podatkov

Prvi del: principi strojnega učenja

- Strojno učenje
- Modeliranje podatkov
- Nadzorovano učenje
- Primeri algoritmov

Algoritmi za učenje klasifikacijskih modelov

Shranjevanje podatkov
- Metoda najbližjih sosedov

Simbolični
- Odločitvena drevesa
- Odločitvena pravila

Numerični
- Perceptron
- Winnow
- Metode podpornih vektorjev
- Linearna regresija

Grafični modeli
- Naivni Bayesov klasifikator
- Bayesove verjetnostne mreže

Globoko učenje
- Nevronske mreže
- Transformerji

Metoda najbližjih sosedov

- Shranimo nespremenjene vse učne primere
  - Enostavno, zahteva učinkovito poizvedovanje
- Klasifikacija **s primerjavo** novega primera s shranjenimi učnimi primeri in oceno oznake primera na osnovi **oznak *k* najbližjih** učnih primerov (*k*-NN)
  - Občutljiv na izbrano metriko pri računanju razdalje med novim primerom in učnimi primeri (običajno se upotablja evklidska razdalja, pri besedilih kosinusna razdalja)

Mera podobnosti

- Zvezni atributi
  - normaliziramo na interval [0,1]

Evklidska razdalja

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

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

- Diskretni atributi
  - razdalja med različnimi vrednostmi = 1
  - Razdalja med enakimi vrednostmi = 0

Metoda najbližjih sosedov - primer

Model

aslov	Junaki	Trajanje
Bob th uilder	vozila, ljudje, …	kratko
Pixar-L omotion	vozila, …, …	kratko
Ice age	…, …, živali	dolgo
Over the edge	…, …, živali	zmerno
Cars	vozila, …, …	dolgo
Anima	…, ljudje, …	dolgo
South Park	…, ljudje, …	zmerno
Simpson	…, ljudje, …	kratko

Atributi

vozila (da, ne)	ljudje (da, ne)	živali (da, ne)	Trajanje [min] (kratko, zmerno, dolgo)

$$L_1(e_1,e_2)=\|e_1-e_2\|=\left(\sum_{i=1}^{|F|}(f_{1i}-f_{2i})\right);f_i-f_j=1;f_i\neq f_j$$

Metoda najbližjih sosedov - primer

Naslov	Junaki	Trajanje	Dist.
Bob the builder	~~vozila~~, ljudje, …	kratko	3
Pixar-Locomotion	vozila, …, …	kratko	4
Ice age	…, …, živali	dolgo	3
Over the hedge	…, …, živali	zmerno	2
Cars	vozila, …, …	dolgo	4
Anima	…, ljudje, …	dolgo	3
South Park	…, ljudje, …	kratko	3
Simpson	…, ljudje, …	kratko	3

Jungle book	ljudje,	živali	zmerno

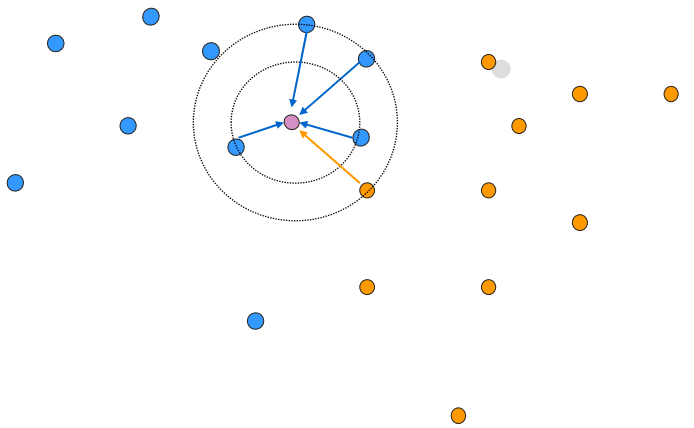
$$L_1(e_1,e_2)=\|e_1-e_2\|=\left(\sum_{i=1}^{|F|}(f_{1i}-f_{2i})\right);f_i-f_j=1;f_i\neq f_j$$

Metoda najbližjih sosedov - primer

Naslov	Junaki	Trajanje	Dist.	Razdalja (v, l, ž, t)
Bob the builder	~~vozila~~, ljudje, …	kratko	3	1+0+1+1=3
Pixar-Locomotion	vozila, …, …	kratko	4	1+1+1+1=4
Ice age	…, …, živali	dolgo	3	1+1+0+1=3
Over the hedge	…, …, živali	zmerno	2	1+1+0+0=2
Cars	vozila, …, …	dolgo	4	1+1+1+1=4
Anima	…, ljudje, …	dolgo	3	1+0+1+1=3
South Park	…, ljudje, …	kratko	3	1+0+1+1=3
Simpson	…, ljudje, …	kratko	3	1+0+1+1=3
Jungle book	ljudje,	živali	zmerno		

$$L_1(e_1,e_2) = \| e_1 - e_2 \| = \left(\sum_{i=1}^{|F|} (f_{1i} - f_{2i}) \right); f_i - f_j = 1; f_i \neq f_j$$

k-najbližjih sosedov



- K = 2
- K = 5

Odločitvena drevesa

- Odločitveno drevo je sestavljeno iz:
  - Notranjih vozlišč – atributov
  - Vej – podmnožice vrednosti atributov
  - Listov – razredi
- Ena pot v drevesu od korena do lista ustreza enemu odločitvenemu pravilu, kjer so pogoji povezani s operatorjem IN
- Predstavlja sklasifikacijsko funkcijo, ki je hkrati simbolični opis podatkov
- uporabimo ga za klasifikacijo primerov

Informacijski prispevek

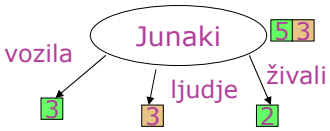
- Pri gradnji drevesa, izbiro atributa, ki bo v vozlišču naredi algoritem glede na kvaliteto atributa
- pogosto se uporablja **informacijski prispevek** atributa

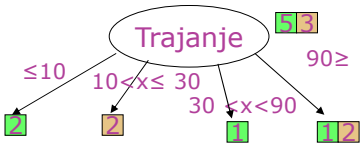
$$InfGain(S,F) = Entropy(S) - \sum_{f \in Values(F)} \frac{|S_f|}{|S|} Entropy(S_f)$$

$$Entropy(S) = -\sum_i p_i \log p_i$$

Odločitvena drevesa – primer gradnje

Naslov	Junaki	Trajanje
Bob the builder	vozila	10 min
Pixar-Locomotion	vozila	5 min
Ice age	živali	90 min
Over the hedge	živali	60 min
Cars	vozila	90 min
Anima	ljudje	90 min
South Park	ljudje	30 min
Simpson	ljudje	20 min

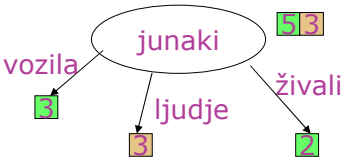




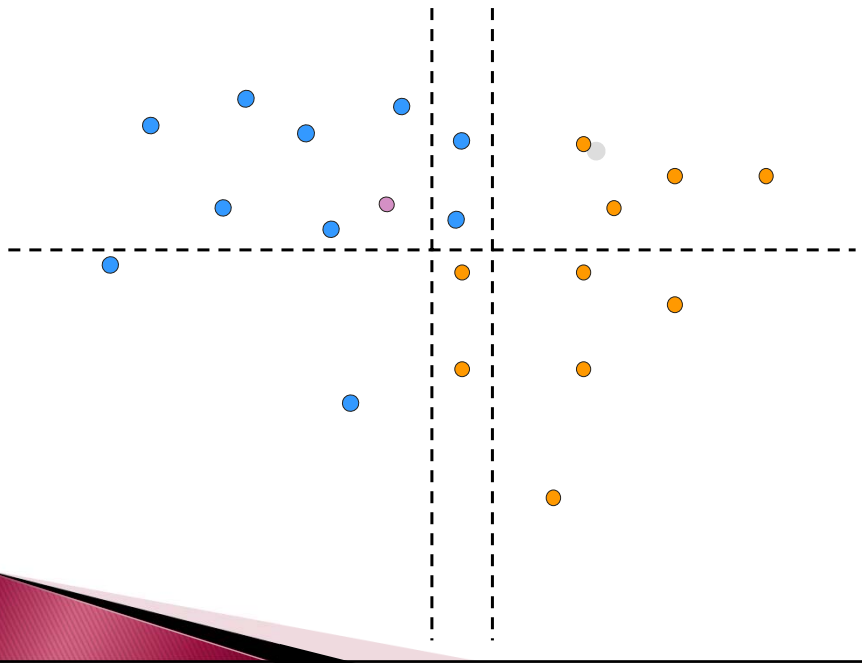
Odločitvena drevesa – primer uporabe

Naslov	Junaki	Trajanje
Bob the builder	vozila	10 min
Pixar-Locomotion	vozila	5 min
Ice age	živali	90 min
Over the hedge	živali	60 min
Cars	vozila	90 min
Anima	ljudje	90 min
South Park	ljudje	30 min
Simpson	ljudje	20 min

Jungle book	živali	60 min



Odločitvena drevesa



Odločitvena pravila

Zgradimo pravila z dodajanjem pogojev:

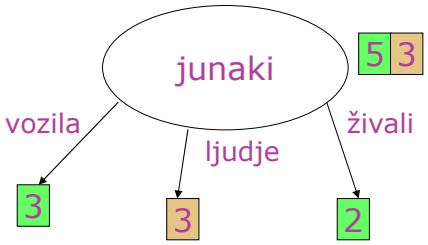
- Λ – omejuje pravilo (manj primerov ustreza pravilu)
- V – posplošuje pravilo (več primerov ustreza pravilu)

- Maksimiziramo kvaliteto vsakega pravila (n.pr., pravilu ustrezajo primeri iz istega razreda) in hkrati poskušamo z množico vseh pravila zajeti vse primere

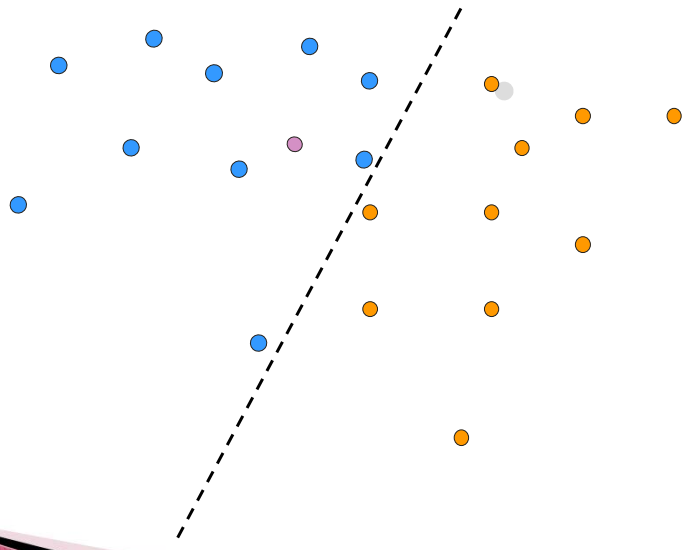
Odločitvena pravila

Prevedba odločitvenega drevesa v pravila

- **Če** (junaki = vozila) **potem** zanimiva
- **Če** (junaki = ljudje) **potem** nezanimiva
- **Če** (junaki = živali) **potem** zanimiva

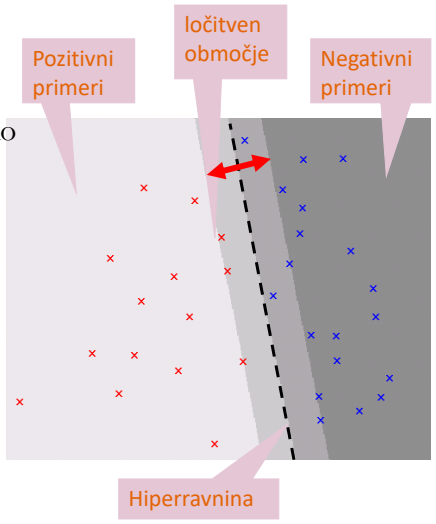


Linearni model - Perceptron



Metoda podpornih vektorjev

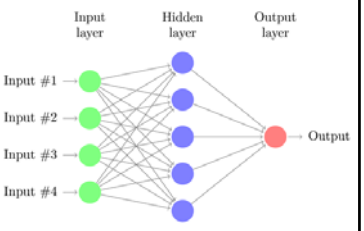
- Nauči se hiperravnino v visokodimenzionalnem prostoru
  - loči učne primere glede na oznako
  - je najbolj oddaljena od najbližjih primerov
- Uporabi implicitne transformacije atributnega prostora v kompleksnejši atributni prostor
  - jedrene funkcije

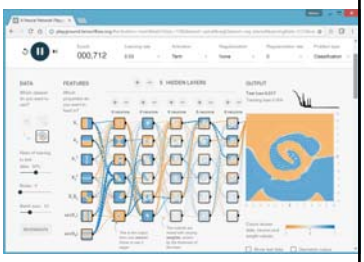


SVM Demo

Nevronske mreže

- Nevronske mreže so trenutno najpopularnejši algoritem v AIju
  - Osnovni algoritem je že star (1950), ki je pa doživel preporod po 2010, ko je procesna moč računalnikov zelo narasla
  - Nevronske mreže so sinonim za **globoko učenje**, ki je v skoraj vsaki moderni AI rešitvi
- Nevronske mreže so sestavljene iz množice preprostih gradnikov (nevroni), ki jih povezujejo povezave (sinapse)
  - …podobno našemu razumevanju delovanja možganov
- Demonstracija Googlovega paketa ***TensorFlow***:
  - http://playground.tensorflow.org





Naivni Bayesov klasifikator

Določi vrednost razreda za primer e_k s pomočjo ocene pogojne verjetnosti

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

- S pomočjo učne množice aproksimira
  - apriorne verjetnostni razredov
    P(c_i) = (št. primerov razreda c_i / št. vseh primerov)
  - pogojne verjetnosti vrednosti atributov pri danem razredu P($e_k \mid c_i$)

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

- Predpostavlja pogojne neodvisnosti vrednosti atributov pri danem razredu

Naivni Bayesov klasifikator na besedilih

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

- Dokument predstavimo kot množico besed W
  - Risanko "Ice age" predstavimo z množico besed: animals, squirrel, ice,…
- Za vsako besedo ocenimo: P(W | pos), P(W | neg)

Napovedovanje pri volitvah v ZDA

- **Using Machine Learning Powers for Good,** Rayid Ghani, University of Chicago

http://videolectures.net/lsoldm2013_ghani_learning_powers/ (56 min)

Povzetek dosedanje snovi

Principi strojnega učenja

- Kaj je strojno učenje?
- Kdaj uporabljamo nadzorovano učenje?
- Primeri algoritmov za strojno učenje

Drugi del: strojno učenje na besedilih

- Predstavitev tekstovnih podatkov
- Razvrščanje v skupine

Predstavitev besedila – vektor besed

- Tekstovne podatke razdelimo na enote, na primer besede ali fraze
  - izbrišemo ločila
  - velike črke nadomestimo z majhnimi
  - lahko tudi izbrišemo znane pogoste besede
    - Angleščina: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ALSO, ...
    - Slovenščina: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ...
- Vsaki besedi priredimo eno mesto v vektorju
  - vanj zapišemo utež besede, recimo število pojavitev v besedilu

Primer predstavitve besedila

A: ti veš

B: kot da ti veš

C: jaz vem da ti veš

D: jaz vem

	ti	veš	kot	da	jaz	vem
A	1	1	0	0	0	0
B	1	1	1	1	0	0
C	1	1	0	1	1	1
D	0	0	0	0	1	1

Delo v skupinah – predstavitev besedil in ugotavljanje avtorstva

- Razdelimo se v skupine
- Vsaka skupina dobi nekaj besedil različnih avtorjev
  - za vsako besedilo zapiše predstavitev z vektorjem besed, s tem da ne upoštevamo pogoste besede (poenostavitev: uporabi frekvence besed namesto uteži)
  - za izbrano novo besedilo neznanega avtorja ugotovi najbolj verjetnega avtorja

En stavek neznanega avtorja in 16 znanih

This is the **information age**, everybody can be **informed** about anything and everything. **Imagine** an **open book**. There is no **secret**, therefore there is no **sacredness.**

1. In information age, 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.
2. 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.
3. All information will be available and if all information is available life will become unbearable.
4. Because with that information, it is good thing that information will be available, it will be terrible thing what to do with this information.
5. With all that information there is no system, where individual is also aware that they have to have also self-control and self-discipline and they should have totally their personality into their own self-control.
6. What we are talking about here is developing the ability to see the consequence of the sequence.
7. See the consequence and remember that when everyone wins, you will make a long-term ally. That is invaluable.
8. When you make your commitment, be ready to stand by it. Everybody will want to work with you.
9. Be as steady as the sun. Everybody wants a person who is consistent and whom they can count on.
10. The turning point comes when you understand who you are and accept it. At that moment you begin living and stop surviving because you can no longer hide.
11. To excel means to go beyond, out of the cell, trespassing beyond given limits and boundaries, to rise, to project.
12. Through adventure after adventure, we move away from what is known and familiar to the unknown, so that the unknown may become known.
13. Passing from Industrialization to the Age of Information technology resulted in an incredible change in human consciousness.
14. Today, information is accessible to everyone, the challenge is to consciously choose the right information and verify it through your experience.
15. Leadership comes when you use self-authority, caliber and discipline to guide you to be yourself, expanding your virtues having faced the challenges in the process.
16. Life offers an extraordinary opportunity to excel. To be excellent means to discover who we are and have the courage to manifest it.

Izračun uteži

- Vsaki besedi (značilki) iz vektorja priredimo utež
- Pogosto uporabljana utež je normalizirana frekvenca besede TFIDF:

$$tfidf(w) = tf.\log(\frac{N}{df(w)})$$

- Tf(w) – term frequency (število pojavitev besede w v dokumentu)
- Df(w) – document frequency (število dokumentov, ki vsebujejo besedo w)
- N – število vseh dokumentov
- Tfidf(w) – normalizirana utež besede v dokumentu

Beseda ima večjo utež,
če se večkrat pojavi v dokumentu

Beseda ima večjo utež, če je pojavlja v manj dokumentih

Primer izračuna uteži

A: ti veš

B: kot da ti veš

C: jaz vem da ti veš

D: jaz vem

	ti	veš	kot	da	jaz	vem
A						
B						
C						
D						

	ti	veš	kot	da	jaz	vem
A						
B						
C						
D						
DF						
Log(N/DF)						

Primer izračuna uteži

A: ti veš
B: kot da ti veš
C: jaz vem da ti veš
D: jaz vem

	ti	veš	kot	da	jaz	vem
A	1	1	0	0	0	0
B	1	1	1	1	0	0
C	1	1	0	1	1	1
D	0	0	0	0	1	1

	ti	veš	kot	da	jaz	vem
A	0.42	0.42	0	0	0	0
B	0.42	0.42	2	1	0	0
C	0.42	0.42	0	1	1	1
D	0	0	0	0	1	1
DF	3	3	1	2	2	2
Log(N/DF)	0.42	0.42	2	1	1	1

Krnjenje in lematizacija besed

- Različne oblike iste besede so lahko problematične, posebej pri naravnih jezikih z visoko pregibnostjo, kot je Slovenščina
  - smej, smejal, smejala, smejale, smejali, smejalo, smejati, smejejo, smejeta, smejete, smejeva, smeješ, smejemo, smejiš, smeje, smejoč, smejta, smejte, smejva
- Besedo nadomestimo z njenim krnom (smej) ali lemo (smejati)
  - Običajno z uporabo hevrističnih pravil – približek

Drugi del: strojno učenje na besedilih

- Predstavitev tekstovnih podatkov
- Razvrščanje v skupine

Razvrščanje v skupine

- Razvrščanje v skupine je proces iskanja razdelitve podatkov
- Podatki znotraj skupine naj bi bili čimbolj podbni
- Podatki med skupinami naj bi bili čim bolj različni
- Pri razvrščanju je ključna mera podobnosti/razdalje
  - Kosinusna razdalja
  - Jaccardova podobnost
  - razdalje Minkowskega (norma k)

$$L_k(d_n, d_m) = \| d_n - d_m \|_k = \left(\sum_{i=1}^{|W|} (w_{in} - w_{im})^k \right)^{1/k}$$

  - razdalja Manhattan (k=1), evklidska razdalja (k=2)

Podobnost med dokumenti

- Vsak dokument predstavimo z vektorjem uteži besed iz slovarja

D = <x>

- Podobnost ocenimo glede na število in utež skupnih besed
  - izračunamo kot kosinus kota med vektorjema:

$$Similarity\ (D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_{1j}^2} \sqrt{\sum_k x_{2k}^2}}$$

Delo v skupinah – razvščanje v skupine

- Razdelimo se v skupine
- Vsaka skupina dobi nekaj besedil različnih vendar neznanih avtorjev
  - razdeli besedila v skupine glede na avtorstvo (določi število skupin)
  - odgovor predstavi in utemelji

Besedila I

- This is the information age, everybody can be informed about anything and everything. There is no secret, therefore there is no sacredness.
- In information age, 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.
- 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.
- All information will be available and if all information is available life will become unbearable.
- Because with that information, it is good thing that information will be available, it will be terrible thing what to do with this information.
- With all that information there is no system, where individual is also aware that they have to have also self-control and self-discipline and they should have totally their personality into their own self-control.
- What we are talking about here is developing the ability to see the consequence of the sequence.
- See the consequence and remember that when everyone wins, you will make a long-term ally. That is invaluable.
- When you make your commitment, be ready to stand by it. Everybody will want to work with you.

Besedila II

- Be as steady as the sun. Everybody wants a person who is consistent and whom they can count on.
- The turning point comes when you understand who you are and accept it. At that moment you begin living and stop surviving because you can no longer hide.
- To excel means to go beyond, out of the cell, trespassing beyond given limits and boundaries, to rise, to project.
- Through adventure after adventure, we move away from what is known and familiar to the unknown, so that the unknown may become known.
- Passing from Industrialization to the Age of Information technology resulted in an incredible change in human consciousness.
- Today, information is accessible to everyone, the challenge is to consciously choose the right information and verify it through your experience.
- Leadership comes when you use self-authority, caliber and discipline to guide you to be yourself, expanding your virtues having faced the challenges in the process.
- Life offers an extraordinary opportunity to excel. To be excellent means to discover who we are and have the courage to manifest it.

Primer predstavitve besedila

A: ti veš
B: kot da ti veš
C: jaz vem da ti veš
D: jaz vem

	ti	veš	kot	da	jaz	vem
A	1	1	0	0	0	0
B	1	1	1	1	0	0
C	1	1	0	1	1	1
D	0	0	0	0	1	1

Primet izračuna podobnosti besedil

Sim(A, B) =
Sim(A, C) =
Sim(A, D) =
Sim(B, C) =
Sim(C, C) =

	ti	veš	kot	da	jaz	vem
A	1	1	0	0	0	0
B	1	1	1	1	0	0
C	1	1	0	1	1	1
D	0	0	0	0	1	1

Primet izračuna podobnosti besedil

$$\text{Sim(A, B)} = \frac{2}{\sqrt{2}\sqrt{4}} = \frac{2}{1.41*2} = 0.71$$

$$\text{Sim(A, C)} = \frac{2}{\sqrt{2}\sqrt{5}} = \frac{2}{1.41*2.24} = 0.63$$

$$\text{Sim(A, D)} = \frac{0}{\sqrt{2}\sqrt{2}} = \frac{0}{1.41*1.41} = 0.00$$

$$\text{Sim(B, C)} = \frac{3}{\sqrt{4}\sqrt{5}} = \frac{3}{2*2.24} = 0.67$$

$$\text{Sim(C, C)} = \frac{5}{\sqrt{5}\sqrt{5}} = \frac{5}{5} = 1.00$$

	ti	veš	kot	da	jaz	vem
A	1	1	0	0	0	0
B	1	1	1	1	0	0
C	1	1	0	1	1	1
D	0	0	0	0	1	1

Metode za razvrščanje

- Hierahične
  - metode združevanja - v vsakem koraku postopka združimo dve ali več skupin
  - metode cepitve - na vsakem koraku izbrano skupino razcepimo na dve ali več skupin
- Nehierarhične
  - potrebno vnaprej podati število skupin iskane razvrstitve
  - optimizacija podane začetne razvrstitve (n.pr., min razdalje primerov znotraj skupine)
- Geometrijske metode
  - Preslikajo večrasežni prostor v dvo- ali trirazsežni prostor (n.pr., metoda glavnih komponent)
- Grafovske metode

Primeri metode za razvrščanje

- Za razvršanje besedil se pogosto uporablja metoda voditeljev (k-means)
- **Podana**:
  - množica dokumentov (n.pr., TFIDF vektorjev),
  - mera podobnosti (n.pr., kosinusna razdalja)
  - ***K*** (število skupin iskane razvrstitve)
- **Za vsako izmed *K*** skupin naključno izberi dokument in ga postavi kot začetno vrednost centroida skupine
- **Dokler ni razvrstitev stabilna ponovi**
  - Vsak dokument priredi skupini, kateri je najboljpodoben (podobnost med dokumentom in centroidom skupine)
  - Za vsako skupino izračunaj nov centroid (povprečje dokumentov skupine)

Primer razvrščanja z metodo voditeljev

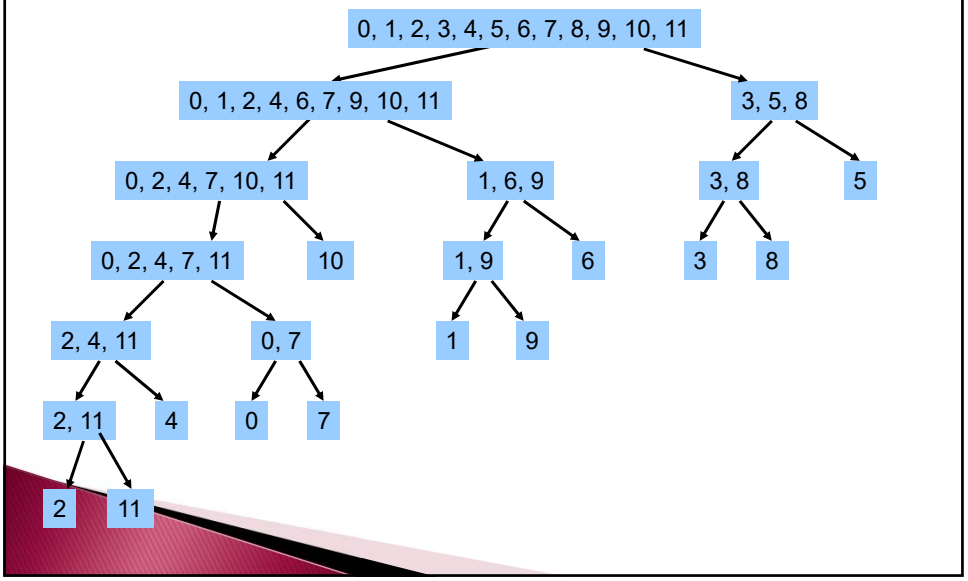
Primeri:

- A: 1,0,1,0,1
- B: 1,0,0,0,1
- C: 1,0,1,0,0
- D: 0,0,0,1,0
- E: 0,1,0,1,0

K=2

1. Naključno izberemo dva primera n.pr. A, D za voditelja skupin I: A, II: D
2. Izračunamo podobnost do ostalih primerov
   BI= 0.82, BII= 0, CI= 0.82, CII= 0, EI= 0, EII= 0.7
3. Razvrstimo primere I: (A,B,C) II: (D,E)
4. Izračunamo centroide skupin
   I: 1,0,0.67,0,0.67 II: 0,0.5,0,1,0
5. Izračunamo podobnost do centroidov
   AI= 0.88, AII= 0, BI= 0.77, BII= 0, CI= 0.77, CII= 0, DI= 0, DII= 0.82, EI= 0, EII= 0.87
6. Razvrstimo primere I: (A,B,C) II: (D,E)
7. Ugotovimo, da je razvrstitev stabilna in končamo

Primer razvrščanja v hierarhijo z metodo voditeljev (bisecting k-means)



Nenadzorovano učenje

- **Data Clustering: 50 Years Beyond K-means,** Anil K. Jain, Department of Computer Science and Engineering, Michigan State University

http://videolectures.net/ecmlpkdd08_jain_dcyb/ (42 min)

- **A theory of similarity functions for learning and clustering,** Avrim Blum, School of Computer Science, Carnegie Mellon University
  http://videolectures.net/sicgt07_blum_atosf/ (56 min)

Delo v skupinah – metoda voditeljev

- izberemo dokumente za voditelje skupin
- vsak dobi kratek dokument
- dokler ni razvrstitev stabilna ponovi:
  - ponovi za vsak dokument
    - priredi dokument najboljpodobni skupini (kosinusna razdalja)
  - za vsako skupino izračunaj centroid (povprečje dokumentov skupine)
- rešitev predstavi in utemelji

Delo v skupinah – uporaba poenostavljene podobnosti za razvrščanje z voditelji

- Naključno izberemo n voditelje (n.pr., 3)
- Vsak voditelj dobi en dokument (n.pr., Q1, Q7, Q11), ostali dobijo vsak svoj dokument (n.pr., Q1-Q13)
- Dokler ni razvrsitev stabilna, ponavljaj
  - Vsak poišče najbolj podobnega voditelja in mu pusti kopijo svojega dokumenta
  - Vsak od treh voditeljeve zgradi centroid svoje skupine

Besedila

- Q1: This is the **information age**, everybody can be **informed** about anything and everything. **Imagine** an **open book**. There is no **secret**, therefore there is no **sacredness**.
- Q7: What we are **talking** about here is **developing** the **ability** to see the **consequence** of the **sequence**.
- Q11: The **turning point** comes when you **understand** who you are and **accept** it. At that **moment** you **begin living** and **stop surviving** because you can no **longer hide**.

Besedila

- Q2: In **information age, 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. **Remember** this.
- Q3: **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**.
- Q4: All **information** will be **available** and if all **information** is **available, life** will **become unbearable**.
- Q5: Because with that **information**, it is **good thing** that **information** will be **available**, it will be **terrible thing** what to do with this **information**.
- Q6: With all that **information** there is no **system**, where **individual** is also **aware** that they have to have also **self-control** and **self-discipline** and they should have **totally** their **personality** into their own **self-control**.
- **Q8: See** the **consequence** and **remember** that when everyone **wins**, you will **make** a **long-term ally**. That is **invaluable**.
- Q9: When you **make** your **commitment**, be **consistent**, be **steady**, be **ready** to **stand** by it. Everybody will **want** to **work** with you.
- Q10: Be as **steady** as the **sun**. Everybody **want** a **person** who is **consistent** and whom they can **count** on.
- Q12: To **excel means** to go beyond, out of the **cell**, **trespassing** beyond **given limits** and **boundaries**, to **rise**, to **project**, **begin** to **accept**.
- Q13: Through **adventure** after **adventure**, we **move** away from **limits**, from what is **known** and **familiar** to the **unknown**, so that the **unknown** may **become known**.

Besedila

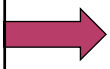
- Q1: This is the **information age**, everybody can be **informed** about anything and everything. **Imagine** an **open book**. There is no **secret**, therefore there is no **sacredness**.
  1. [Q2, Q4, Q5, Q6] life, become, computer, loyal, truthful, excellent, challenged, compete, remember, available, unbearable, good, thing, terrible, system, individual, aware, self-control, self-discipline, totally, personality
  2. [Q13, Q2, Q4, Q5, Q6] adventure, move, limits, known, familiar, unknown
- Q7: What we are **talking** about here is **developing** the **ability** to see the **consequence** of the **sequence**.
  1. [Q8] see, remember, win, make, long-term, ally, invaluable
  2. [Q8, Q9, Q3] commitment, consistent, steady, ready, stand, want, work, truth, simple, straight, smile, say, know, live
  3. [Q10, Q8, Q9, Q3] sun, person, count
- Q11: The **turning point** comes when you **understand** who you are and **accept** it. At that **moment** you **begin living** and **stop surviving** because you can no **longer hide**.
  - [Q12] excel, means, cell, trespassing, given, limits, boundaries, rise, project

Povzetek dosedanje snovi

Drugi del: kako?

- Kako predstavimo tekstovne podatke?
- Kako izračunamo uteži posameznim besedam?
- Kaj je krnjenje in kaj je lematizacija besed? Zakaj se uporablja?
- Kako razvrščamo dokumente v skupine?

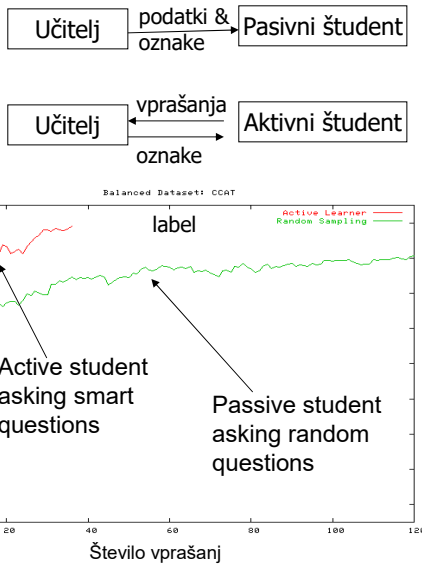
Tretji del: tehnike analize besedil



- Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija v taksonomije
- Gradnja vsebinskih ontologij
- Primeri nalog

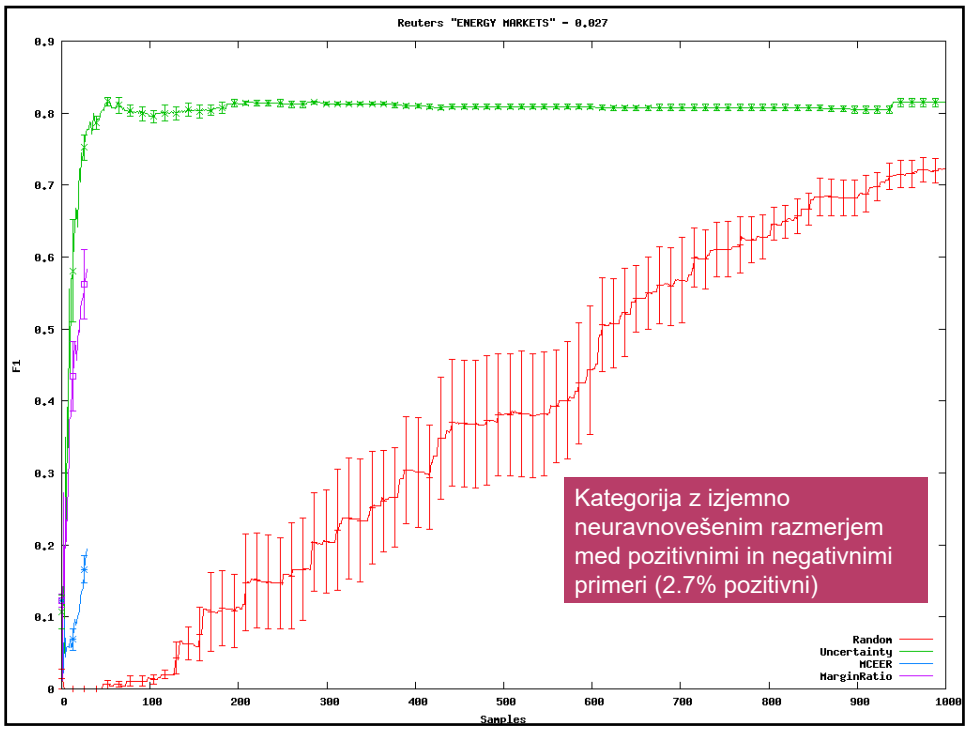
Aktivno učenje

- Uporabljamo, ko naloga vsebuje oznake, vendar imamo malo označenih primerov oz. je označevanje drago
- Uporabnika prosimo za oznake zanimivih primerov
- Zahteva bistveno manj ročnega označevanja za enake rezultate modeliranja kot naključno označevanje



Metode aktivnega učenja

- **Vzorčenje z negotovostjo (Uncertainty sampling)**
  - izberemo primer, ki je najbližje meji med dvema razredoma
- **Maximum margin ratio change**
  - izberemo primer, za katerega predvideamo največji vpliv na velikost mejnega pasu (margine size)
- **Monte Carlo Estimation of Error Reduction**
  - izberemo primer, ki najbolj podpira neš trenutni model
- **Naključno izberemo primer**

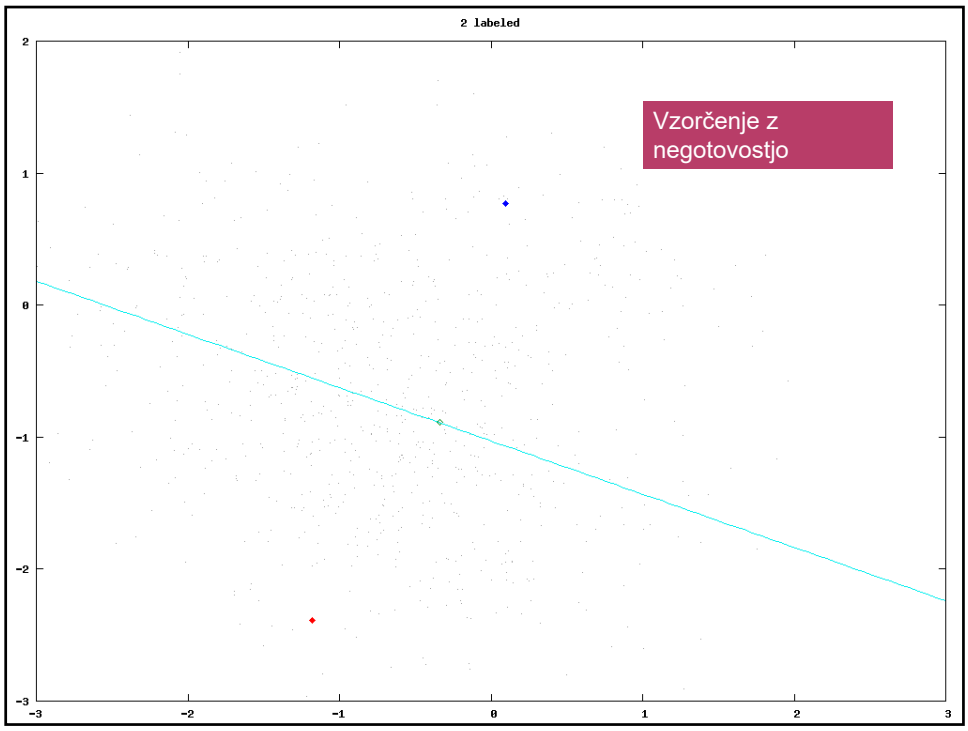


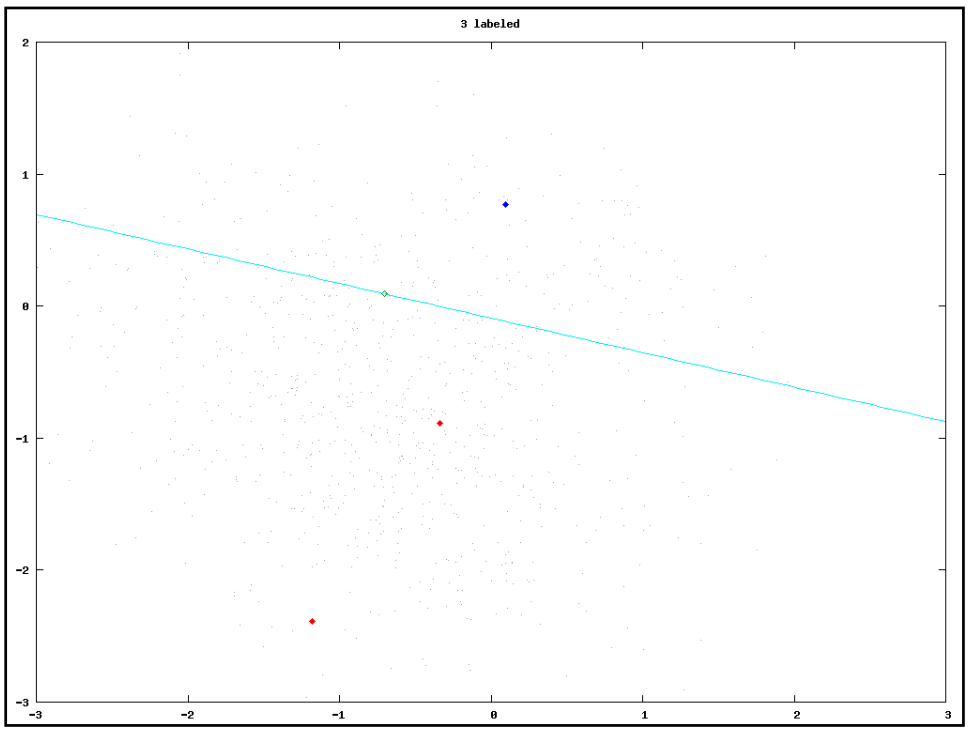
Prikaz delovanja metode aktivnega učenja na primeru

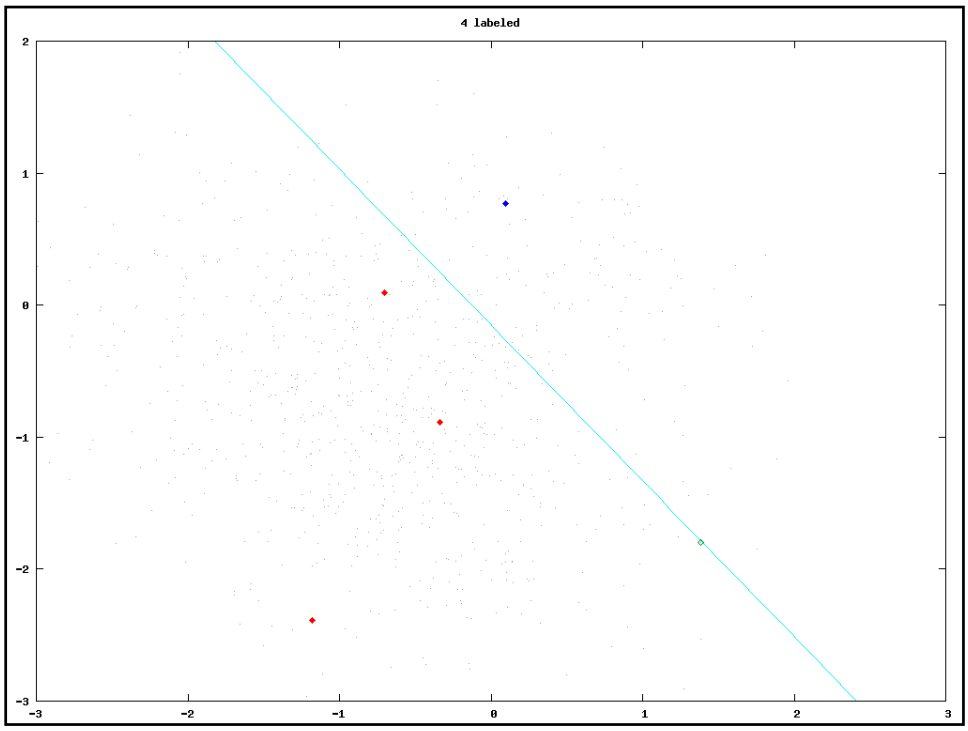
- Potrebujemo vsaj po en označen primer iz vsakega razreda (rdeči in modri)
- Sistem izmed vseh preostalih primerov izbere primer, ki naj ga uporabnik označi (zeleni krogi)
- Pomovno zgradi model z upoštevanjem nanovo označenega primera (premica)

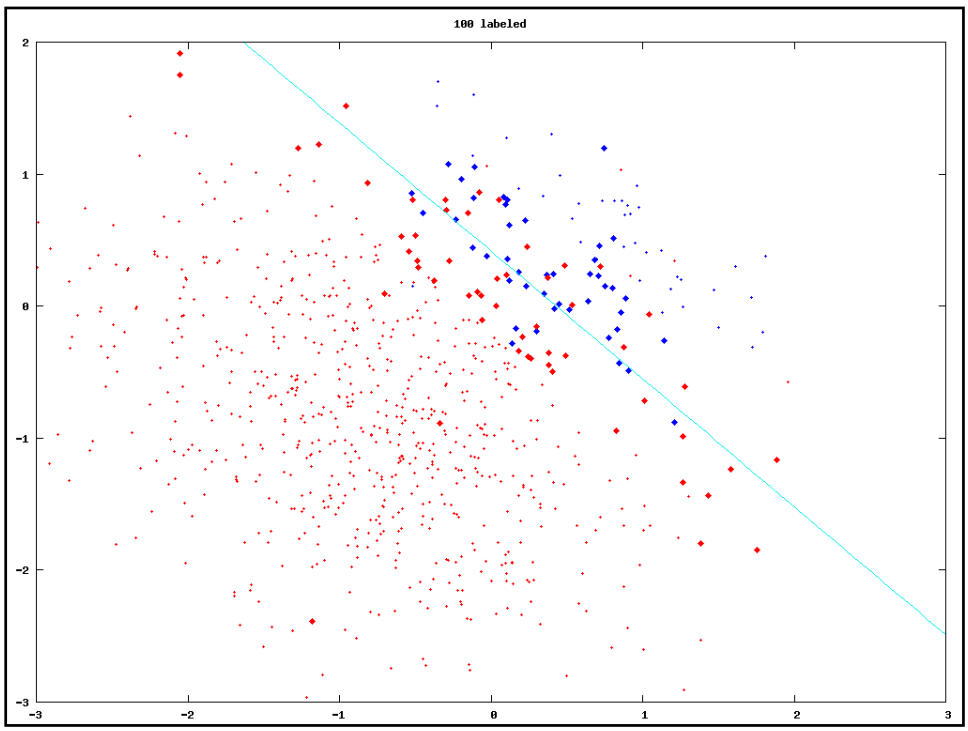
Za prikaz uporabimo linearni SVM model in primerjamo dve metodi izbire primerov za označevanje

- Naključno izberemo neoznačene primere
- Izberemo naoznačen primer z metodo vzorčenje z negotovostjo (primer, ki je najbližje ločitveni hiperravnini)


2 labeled
Vzorčenje z negotovostjo


3 labeled


4 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


100 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Polnadzorovano učenje

- **Semisupervised Learning Approaches,** Tom Mitchell, Machine Learning Department, School of Computer Science, Carnegie Mellon University http://videolectures.net/mlas06_mitchell_sla/ (60 min)

Tretji del: tehnike analize besedil

- Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija v taksonomije
- Gradnja vsebinskih ontologij
- Primeri nalog

Kategorizacija dokumentov

- Problem:
  - Dano imamo množico vsebinskih kategorij. Cilj je, da novemu besedilu (dokumentu) priredimo eno ali več kategorij
- Vsebinske kategorije
  - nestrukturirane (npr. earn, corn, money...)
  - strukturirane (npr. science, computer science, data mining,...education, higher education,...)
- Problem je podoben prirejanju ključnih besed dokumentom

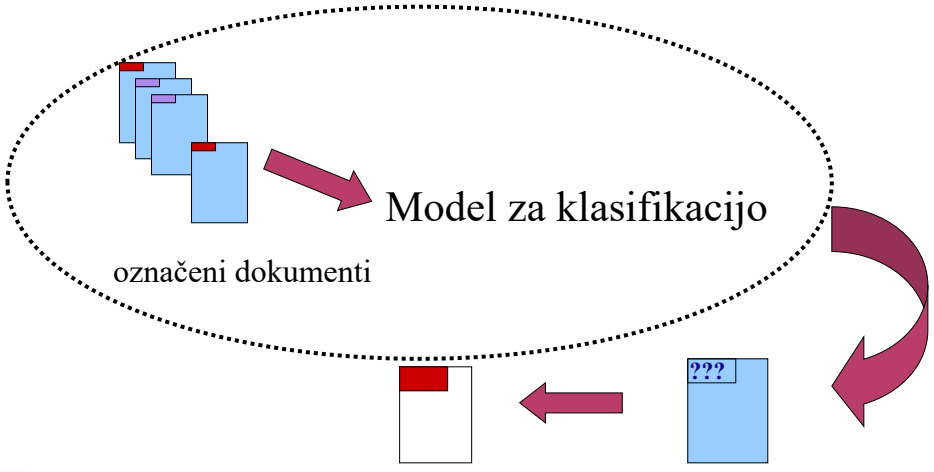
Kako deluje kategorizacija?

- Najprej moramo imeti dokumente, ki so označeni (ročno klasificirani v kategorije)
  - Npr. Yahoo! Spletna taksonomija ali Reuters-21578 novice ali Reuters-2000 novice
- … nato jih (običajno) predelamo v predstavitev primerno za učenje (vreča besed)
- … nato zgradimo klasifikatorje
- … in na koncu jih uporabljamo za klasifikacijo neoznačenih dokumentov

Eksperiment z ročno klasifikacijo spletnih strani

- Eksperiment: desetim ljudem so dali nalogo, da ročno razvrstijo 15 strani, ki jih vrne iskalnik za 5 različnih tematik
- Ugotovitve:
  - Glavna ugotovitev je, da ni velike podobnosti med razvrstitvami različnih ljudi
  - Ljudje se tudi razlikujejo v številu skupin v katere so razvrstili dokumente

Kategorizacija dokumentov



Metode za kategorizacijo dokumentov

- Najpogosteje uporabljane metode so:
  - Support Vector Machines
  - Logistic Regression
  - Perceptron
  - Naive Bayes
  - Winnow
  - Nearest Neighbour
  - ....

Opis postopka - Perceptron

Vhod:

- Množica dokumentov $\boldsymbol{D}$ podana z vektorji števil (n.pr. TFIDF)
- Vsak dokument ima oznako +1 (pozitiven/zanimiv) ali -1 (negativen/nezanimiv)

Rezultat:

- Linearni model $\boldsymbol{w_i}$ (utež za vsako besedo iz nabora besed)

Postopek:

- **Začetne vrednosti vseh uteži** $\boldsymbol{w_i}$ postavimo na 0
- **Ponovimo** $\boldsymbol{N}$ krat
  - **za vsak dokument** $\boldsymbol{d}$ iz $\boldsymbol{D}$
    - // Klasificiramo dokument $\boldsymbol{d}$ z uporabo trenutnega modela $\boldsymbol{w_i}$
    - **Če je** sum($\boldsymbol{d_i}$*$\boldsymbol{w_i}$) >= 0 **potem** priredimo dokumentu pozitiven razred
    - **Sicer** priredimo dokumentu negativen razred
    - **Če smo se zmotili pri klasifikaciji potem**
      - // popravimo uteži vseh besed, ki se pojavijo v dokumentu $\boldsymbol{d}$
      - $\boldsymbol{w_{t+1}} = \boldsymbol{w_t}$+sign(true-class) * Beta // (vhodni parameter Beta>0)
      - // sign(positive) = 1 in sign(negative) =-1

Example of Perceptron

	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

if sum(d_i *w_i) >= 0 **then**
Classify(d) = pos
else Classify(d) = neg

if Classify(d) wrong **then**
// adjust weights for words occurring in d
$w_{t+1} = w_t$ +sign(true-class) * Beta
// where sign(pos) = 1; sign(neg) = -1

1. Initialization: W = [0, 0, 0, 0, 0]; Beta = 0.8

2. Iteration 1

B*W = 0; Classify(B) = pos; C*W = 0; Classify(C) = pos; D*W = 0; Classify(D) = pos (wrong!)

W = [0, 0, 0, - 0.8, 0]

E*W = -0.8; Classify(E) = neg

3. Iteration 2

B*W = 0; Classify(B) = pos; C*W = 0; Classify(C) = pos; D*W = -0.8; Classify(D) = neg

E*W = -0.8; Classify(E) = neg

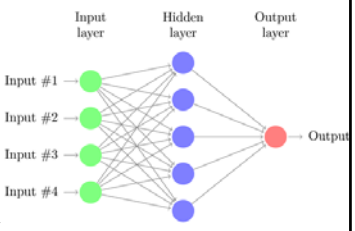
Classify(A)

W = [0, 0, 0, - 0.8, 0] * A = [1, 0, 1, 0, 1]

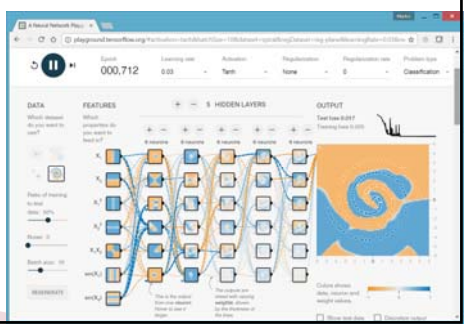
Classify(A) = **pos**

Example learning algorithm: Neural Networks

- Neural Networks
  - appeared first in early 90ties, but really revolutionized AI after 2010
  - today Neural networks are a synonym for "***Deep Learning***" solving several previously unsolved problems



- Neural Networks are a composition of many very simple building blocks (analytical functions)
  - once we connect them in a large connected network, they can jointly solve state-of-the-art AI problems
  - consisting from neurons, connected with synapses to simulate architecture of the brain



- Demonstration of Google's ***TensorFlow*** package:
  - http://playground.tensorflow.org

Mere uspešnosti – Ocena kvalitete modela

$$Precision(M) = TP / (TP + FP)$$

$$Recall(M) = TP / (TP + FN)$$

$$Accuracy(M) = \sum_i TP_i / (TP_i + FN_i)$$

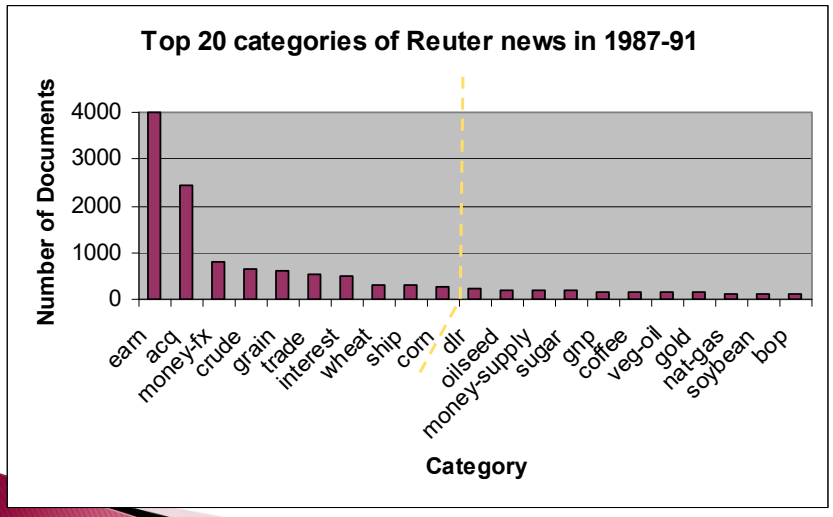
$$F_\beta(M) = \frac{(1+\beta^2) Precision(M) \times Recall(M)}{\beta^2 Precision(M) + Recall(M)}$$

- natančnost (precision), priklic (recall)
- klasifikacijska točnost (accuracy)
- točka ravnotežja (Break-even point: precision=recall)
- F-mera (kombinira natančnost in priklic)

Primer kategorizacije - Reuters novice

- Uredniki so vsakemu dokumentu priredili eno ali več vsebinskih kategorij
- Javno dostopna množica Reuters novic iz leta 1987
  - 120 kategorij: earn, *acquire, corn, rice, jobs, oilseeds, gold, coffee, housing, income,...*
- Pogosto uporabljana za raziskave metod
- Od leta 2000 je za raziskovalne namene na voljo tudi večja množica z 830,000 Reuters novic

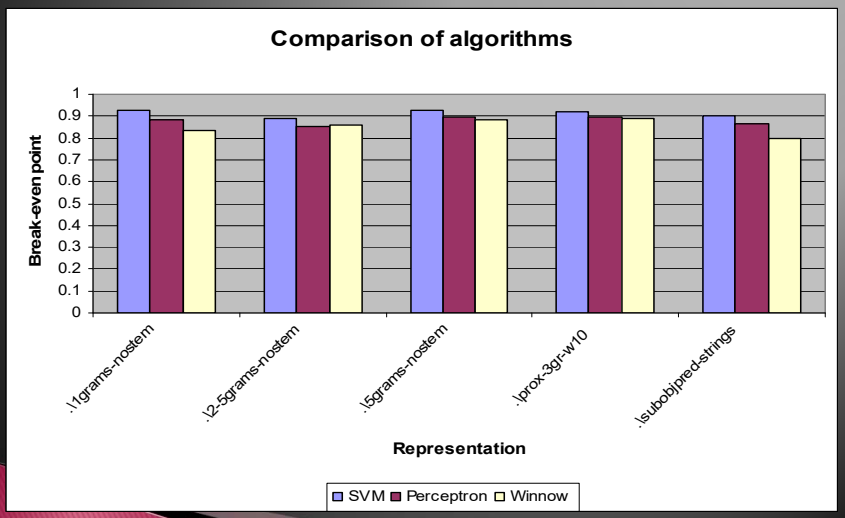
Distribucija dokumentov po kategorijah (Reuters-21578)



Primer Perceptron modela za Reuters kategorijo "Acquisition"

Značilka	Utež
STAKE	11.5
MERGER	9.5
TAKEOVER	9
ACQUIRE	9
ACQUIRED	8
COMPLETES	7.5
OWNERSHIP	7.5
SALE	7.5
OWNERSHIP	7.5
BUYOUT	7
ACQUISITION	6.5
UNDISCLOSED	6.5
BUYS	6.5
ASSETS	6
BID	6
BP	6
DIVISION	5.5
…	

Primerjava postopkov (SVM, Perceptron & Winnow) na novicah Reuters-21578 - različne značilke



Semantični splet

- **Populating the Semantic Web by Macro-Reading Internet Text,** Tom Mitchell, Machine Learning Department, School of Computer Science, Carnegie Mellon University http://videolectures.net/iswc09_mitchell_ptsw/ (56min)

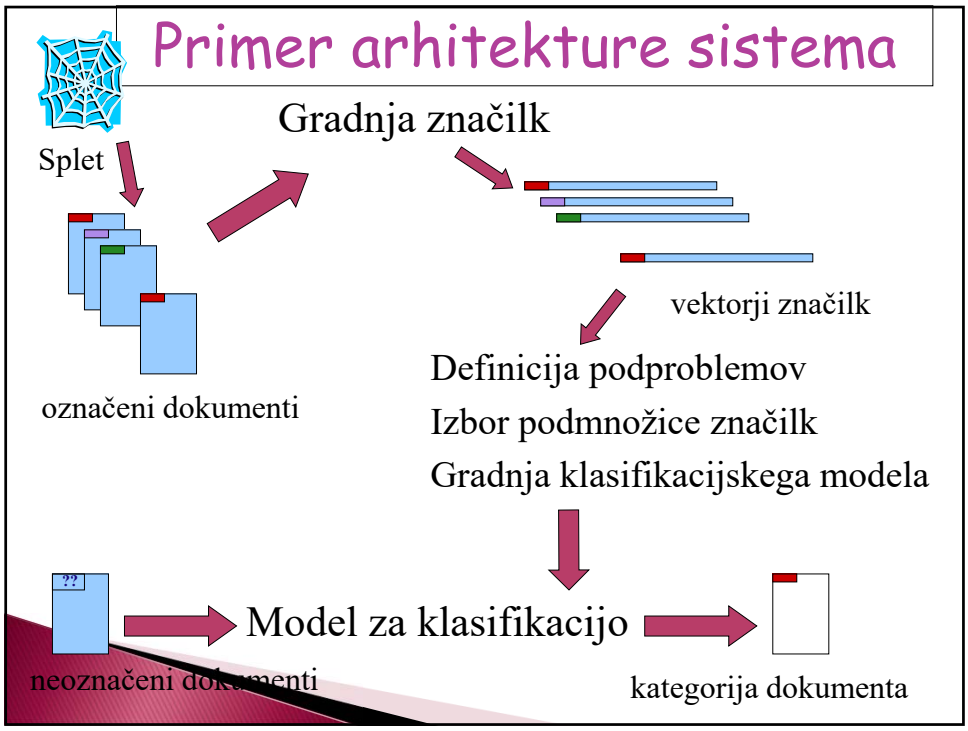
Tretji del: tehnike analize besedil

- Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija v taksonomije
- Gradnja vsebinskih ontologij
- Primeri nalog

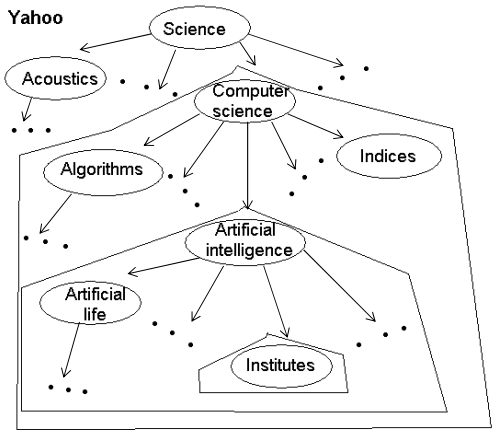
Kategorizaciji v taksonomije

- Vzeli smo obstoječo zbirko ročno kategoriziranih besedil v strukturirane vsebinske kategorije *Yahoo!*
- S Text Mining metodami smo zgradili model, ki posnema ročno delo urednikov
- Z uporabo modela znamo novemu, še nevidenemu dokumentu prirediti vsebinske kategorije in ustrezne ključne besede.

Primer arhitekture sistema



Definicija podproblemov



- za vsako vsebinsko kategorijo zgradimo ločen klasifikator, ki napove verjetnost, da nek nov dokument pripada kategoriji

Primer rezultatov kategorizacije

`Reference:Libraries:Library and Information Science:Information Retrieval' 0.9999
`Reference:Libraries' 0.9987
`Reference:Libraries:Indices' 0.9523
`Reference:Libraries:Library and Information Science:Institutes' 0.9388
`Reference:Libraries:Library and Information Science' 0.9327

Predicted category:
Rank = 1, Probability = 0.9999
Assigned keywords:
Reference, Libraries, Library and Information Science, Information Retrieval, Indices, Institutes
(Recall = 1, Precision = 0.67 (4 out of 6))

Experimental setting: Odds ratio, vector size=1, pruning=(1.0,3), probability=0.90

Klasifikacija besedil

- **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)
- **Semi-supervised Learning for Text Classification,** Anastasia Krithara, Xerox Research Centre Europe, Xerox http://videolectures.net/mlss07_krithara_ssl/ (14 min)
- **Cross Language Text Classification via Multi-view Subspace Learning,** Yuhong Guo, Department of Computer and Information Sciences, Temple University http://videolectures.net/nipsworkshops2012_guo_subspace_learning/ (12 min)

Povzetek dosedanje snovi

Učenje na tekstovnih podatkih

- Polnadzorovano učenje
  - Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija dokumentov v taksonomije

Tretji del: Tehnike analize besedil

- Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija v taksonomije
- Gradnja vsebinskih ontologij
- Primeri nalog

Gradnja vsebinskih ontologij

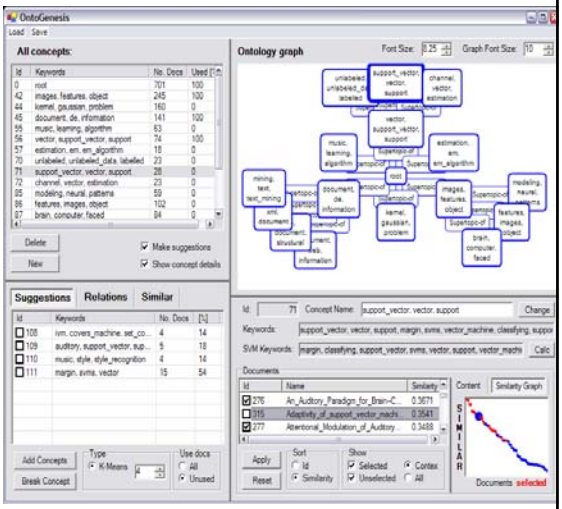
- Pod vsebinsko ontologijo si predstavljamo povezano množico pojmov, ki opisujejo vsebine
- Primeri so:
  - Yahoo! taksonomija spletnih strani
  - MEDLINE taksonomija medicinskih člankov
- Pogledali bomo primer sistema za gradnjo vsebinskih ontologij - OntoGen

Osnovne lastnosti sistema OntoGen

- Pol-avtomatski pristop
  - Metode analize tekstovnih podatkov uporabimo za svetovanje uporabniku in izdelavo različnih vpogledov v problemsko področje
  - Uporabnik lahko vpliva na rezultate preko parametrov uporabljenih metod
  - Končna odločitev je v rokah uporabnika
- Podatkovno-voden pristop
  - Večina pomoči, ki jo sistem nudi uporabniku je na osnovi podatkov, ki jih uporabnik predloži
  - Primere na osnovi katerih sistem deluje opišemo z množico značilk, n.pr. vektorji besed

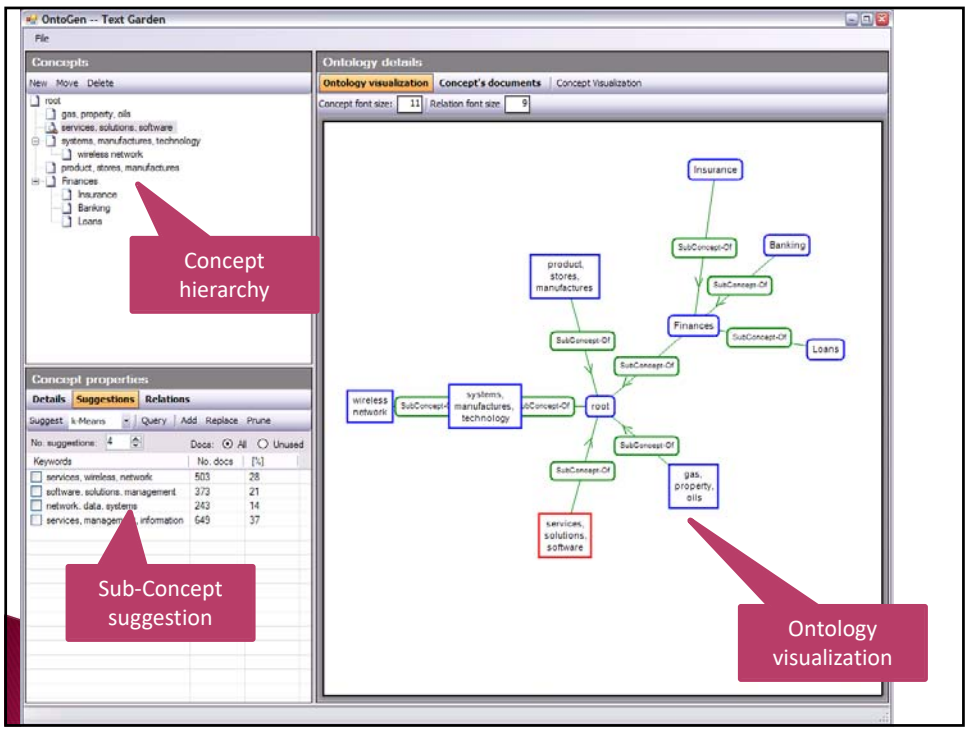
OntoGen

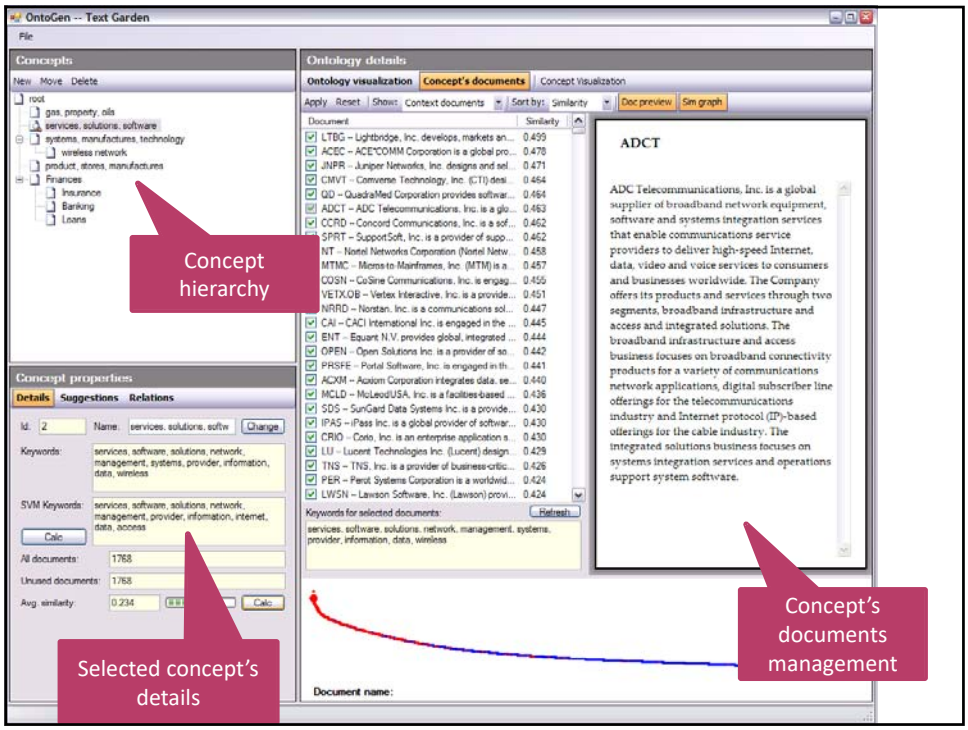
- Sistem je namenjen gradnji vsebinskih ontologij
- Metode razvrščanja so uporabljene za svetovanje vsebinskih skupin
- Izločanje ključnih besed pomaga uporabniku pri poimenovanju konceptov
- Interaktiven uporabniški vmesnik



OntoGen – dodatne izobljšave

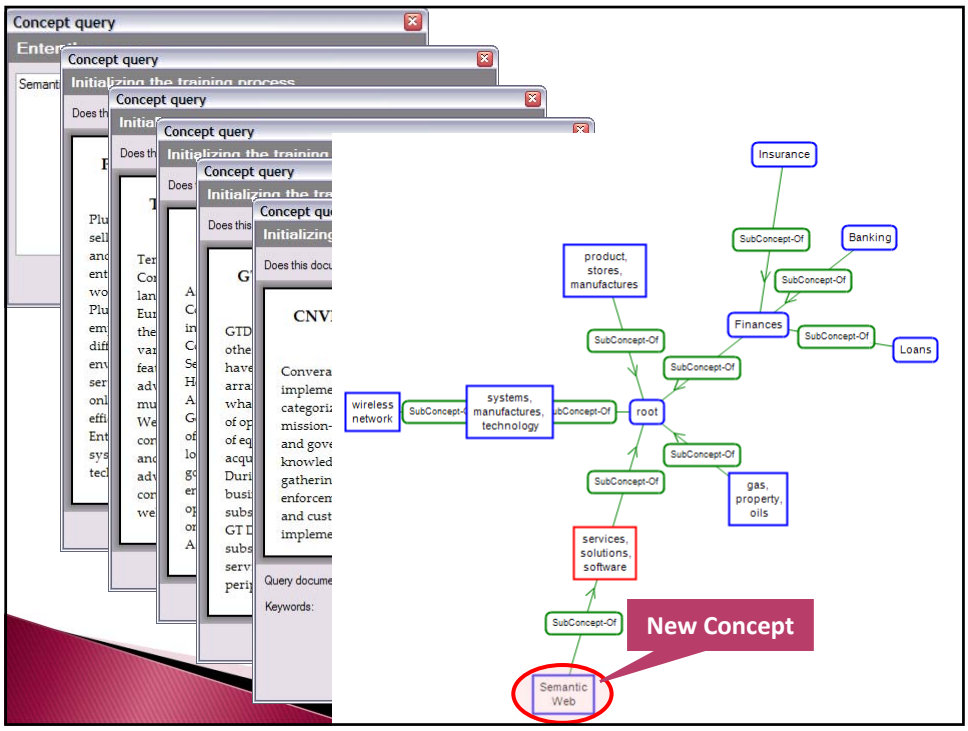
- Izboljšave uporabniškega vmesnika
  - Na osnovi povratne informacije končnih uporabnikov
  - Na osnovi opažanj avtorjev sistema
- Razširitev z novimi funkcionalnostmi:
  - Aktivno učenja
    - Učenje konceptov na osnovi uporabnikovih vprašanj in kategorizacije previdno izbranih posameznih dokumentov
  - Simultane ontologije
    - Optimizacija mere podobnosti glede na podane kategorije dokumentov
  - Vizualizacija primerov, ki pripadajo konceptom
    - Integracija Document Atlas sistema za vizualizacijo množice dokumentov
  - Naseljevanje ontologij (ontology population)
    - Interaktivna klasifikacija novih primerov v obstoječo ontologijo





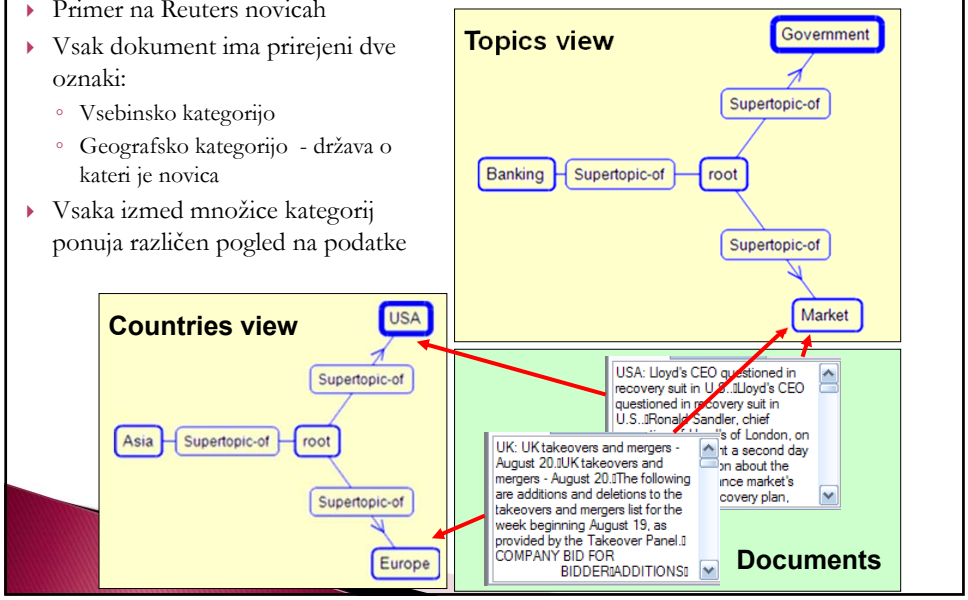
Aktivno učenja

- Metoda na osnovi vzorčenja z negotovostjo
  - razdalje primera od hiper-ravnine, ki jo zgradi metoda podpornih vektorjev (SVM)
- Prve označene dokumente dobimo z uporabo povpraševanj uporabnika in metodo najbližjih sosedov
- Na vsakem koraku izberemo primer, ki je najbližje hiper-ravnini in vprašamo uporabnika za oznako

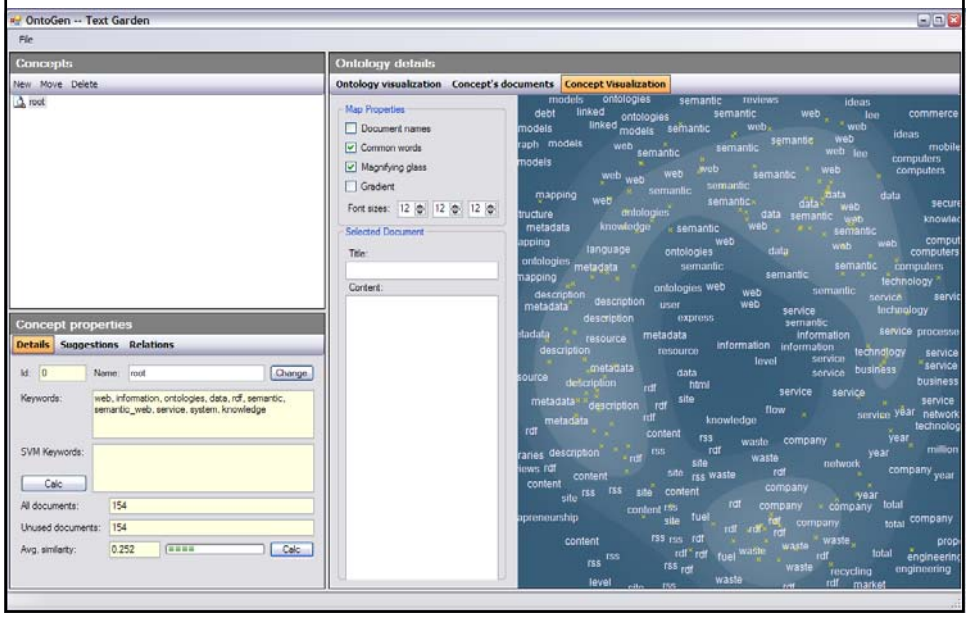


Simultane ontologije

- Primer na Reuters novicah
- Vsak dokument ima prirejeni dve oznaki:
  - Vsebinsko kategorijo
  - Geografsko kategorijo - država o kateri je novica
- Vsaka izmed množice kategorij ponuja različen pogled na podatke

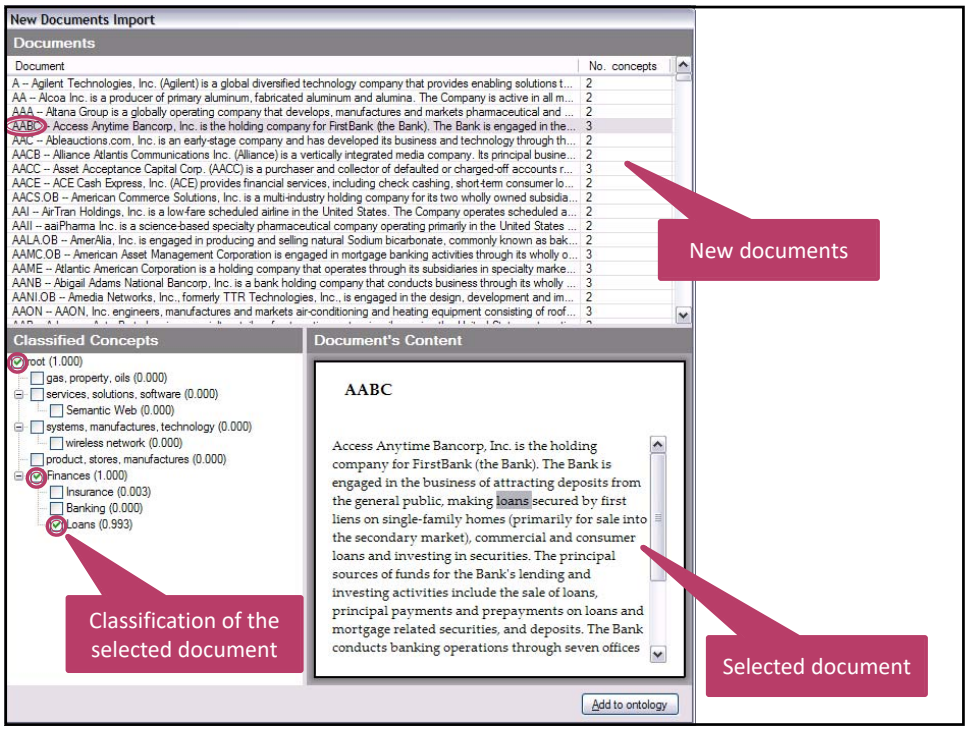


Vizualizacija primerov konceptov



Naseljevanje ontologij

- Uporabimo linearno metodo podpornih vektorjev, ki klasificira en razred naspram vsem ostalim
- Interaktiven uporabniški vmesnik, kjer lahko uporabnik določi končno klasifikacijo



Semantična informacija besedila

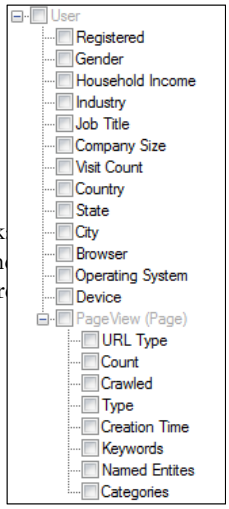
- **High-coverage extraction of semantic assertions from text,** Dunja Mladenić, Artificial Intelligence Laboratory, Jožef Stefan Institute
  http://videolectures.net/sikdd2011_mladenic_assertions/ (12 min)

Tretji del: Tehnike analize besedil

- Aktivno učenje
- Osnovna kategorizacija dokumentov
- Kategorizacija v taksonomije
- Gradnja vsebinskih ontologij
- Primeri nalog

Aplikacija: Analiza spletnih uporabnikov

- Značilke
  - Pridobljene iz opisa uporabnika
  - Predstavitev z vektorjem
  - Normalizacija vrednosti posamezne značilke
- Učna množica
  - En obisk = en vektor
  - En uporabnik = centroid vseh obiskov uporabnik
  - Uporabniki iz ciljnega segmenta so pozitivni prim
  - Naključni vzorec uporabnikov iz negativnega razr
- Postopek klasifikacije
  - Metods podpornih vektorjev
  - Primerno za visoko dimenzionalne podatke
  - Linearno jedro



Evalvacija

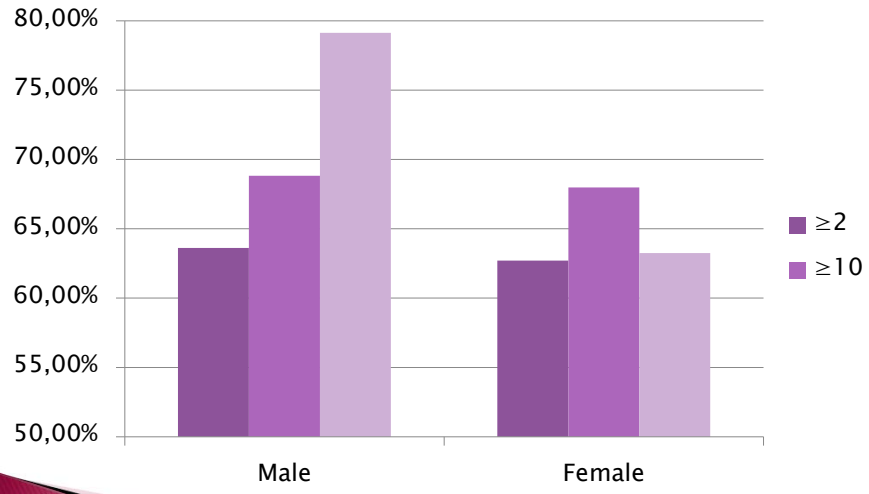
- Realni podatki ene od največjih medijskih hiš
  - 5 millionov uporabnikov dnevno, 1 million registriranih uporabnikov
- Testirali smo napovedovanje ene od treh demografskih dimenzij:
  - spol, starost, prihodek
- Tri skupine uporabnikov, odvisno od števila dosedanjih obiskov spletne strani: ≥2, ≥10, ≥50
- Evalvacija: Break Even Point (BEP), 10-fold prečno preverjanje

Category	Size
Male	250,000
Female	250,000

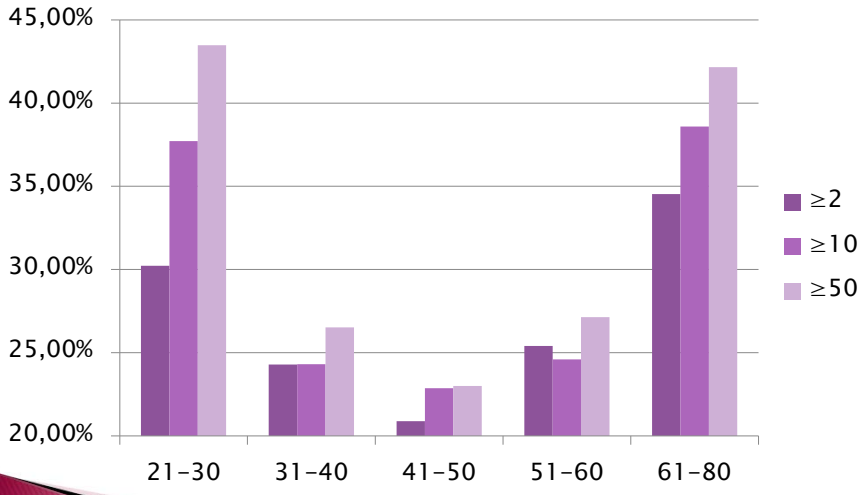
Category	Size
21-30	100,000
31-40	100,000
41-50	100,000
51-60	100,000
61-80	100,000

Category	Size
0-24k	50,000
25k-49k	50,000
50k-74k	50,000
75k-99k	50,000
100k-149k	50,000
150k-254k	50,000

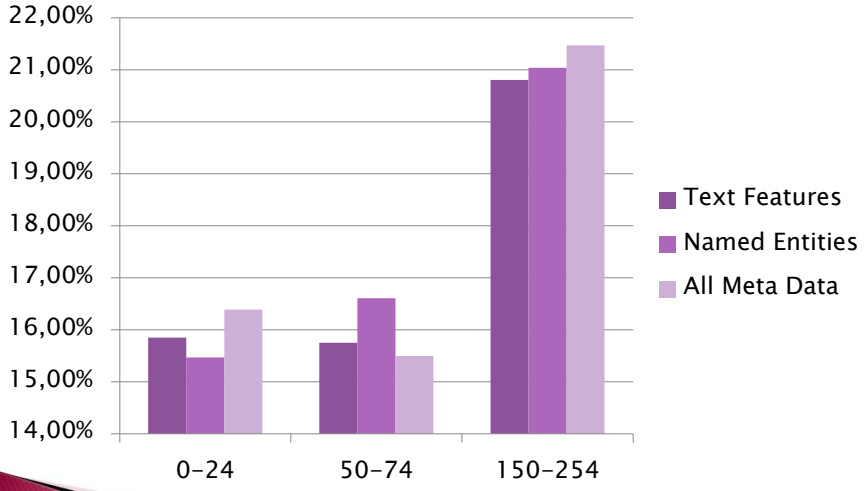
Spol



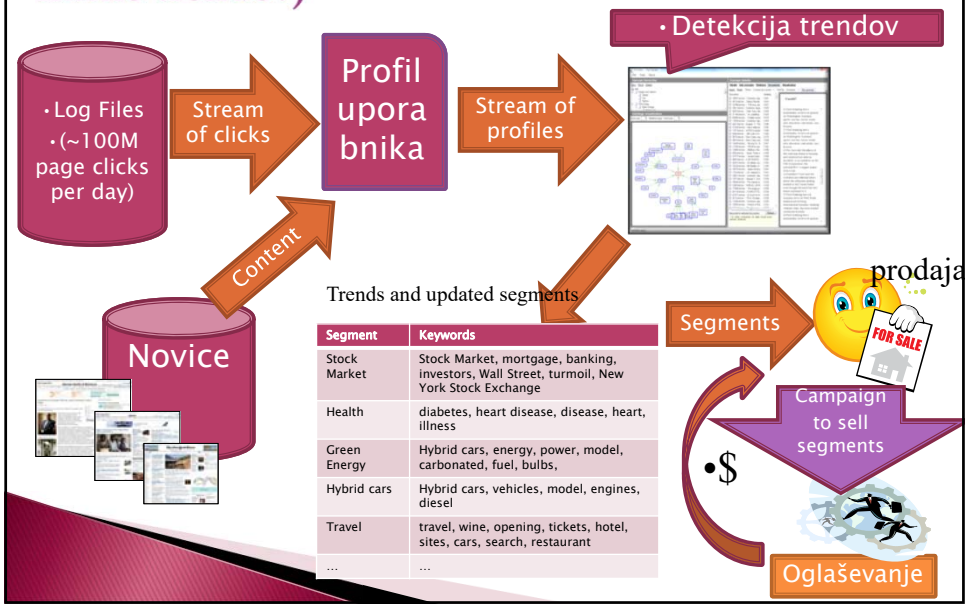
Starost



Prihodki (≥10 obiskov)



Aplikacija: spletno oglaševanje (detekcija mirko trendov)



Dnevna količina podatkov

- 50Gb of uncompressed log files
- 50-100M clicks
- 4-6M unique users
- 7000 unique pages with more then 100 hits

Analiza interakcij uporabnikov

- **User Modeling Combining Access Logs, Page Content and Semantics,** Blaž Fortuna, Artificial Intelligence Laboratory, Jožef Stefan Institute http://videolectures.net/www2011_fortuna_umc/ (15 min)
- **Deconvolution of Networks into Communities,** Jure Leskovec, Computer Science Department, Stanford University http://videolectures.net/kdd2013_leskovec_online_communities/ (23 min)

Povzetek dosedanje snovi

Učenje na tekstovnih podatkih

- Gradnja vsebinskih ontologij
  - Uporaba nenadzorovanega in polnadzorovanega učenja
- Analiza spletnih uporabnikov
- Spletno oglaševanje

Četrti del: Zahtevnejše metode

- Vizualizacija tekstovnih podatkov
- Iskanje na spletu
- Vizualizacija novic
- Izdelava povzetkov
- Prekojezično povezovanje dokumentov
- Analiza socijalnih omrežji
- ....

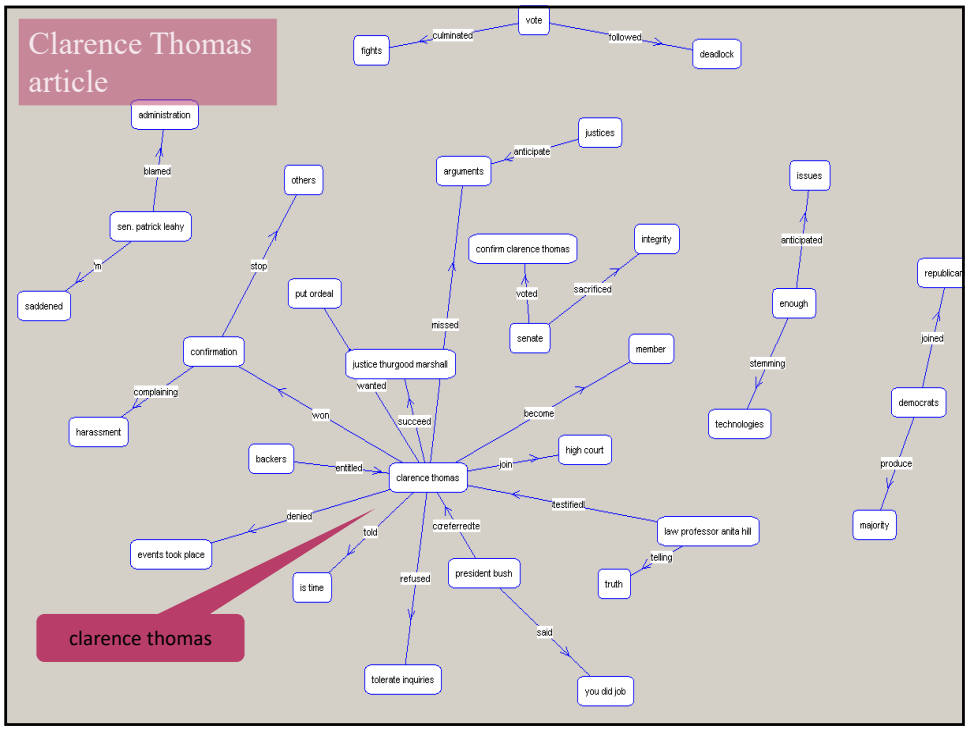
Vizualizacija besedil

Odvisno od podatkov in potreb uporabimo različne metode za vizualno predstavitev:

- Posameznega dokumenta
- Množice dokumentov
- Množice medsebojno povezanih dokumentov (n.pr. spletne strani)

Vizualizacija dokumenta

- Vizualiziramo manjšo količino tekstovnih podatkov – en dokument, običajno nekaj sto besed
- Zaradi manjše količine podatkov težko temeljimo pristope na statistiki podatkov
- Uporabimo analizo naravnega jezika (tip in vloga besed v stavkih)



Vizualizacija množice dokumentov

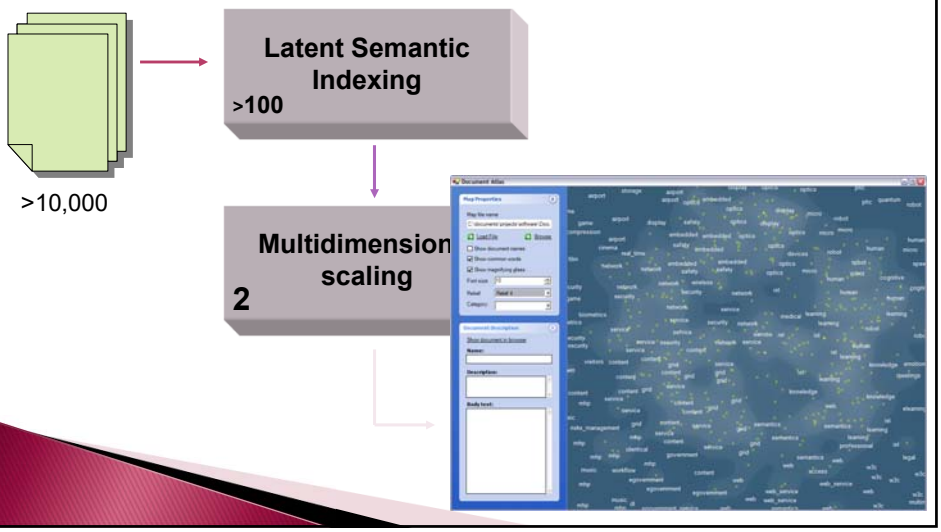
Imamo veliko množico dokumentov

- Kaj so glavne vsebine dokumentov?
- Kateri dokumenti so si podobni?
- Katere vsebino so sorodne?
- Kako omogočiti uporabniku pregled nad celotnim prostorom dokumentov?

Definicija problema

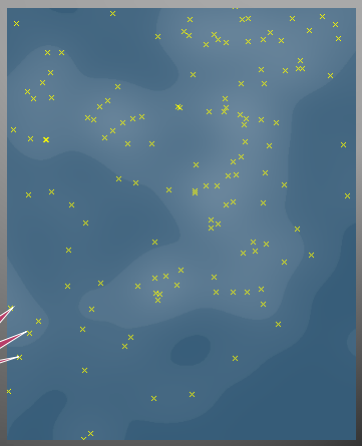
- Imamo množico dokumentov v visoko-dimenzionalnem prostoru (predstavitev z vektorjem besed) – običajno >10,000 dimenzij!
- **Za vizualizacijo potrebujemo prikaz v dveh dimenzijah!**

Postopek



Izgradnja ozadja

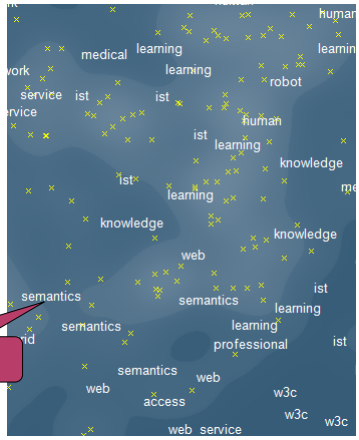
- Gostoto dokumentov uporabimo za gradnjo ozadja – svetlejša barva pomeni višje področje
- Strnjene skupine ponazorimo s konturnim linijam



Dokumenti

Ključne besede

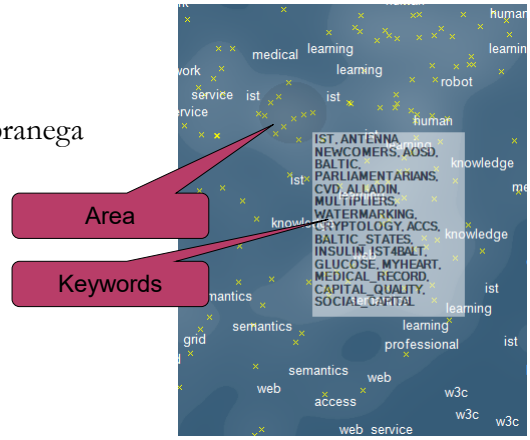
Vsaki točki lahko priredimo množico ključnih besed na osnovi povprečja vektorjev dokumentov iz njene okolice



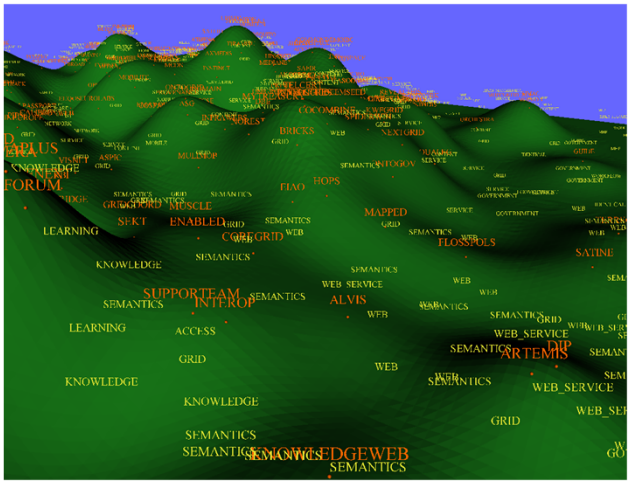
Ključne besede

Ključne besede

- Uporabnik si lahko ogleda podrobnosti
  - ključne besede izbranega podpodročja



Omogoča tudi prikaz v 3D

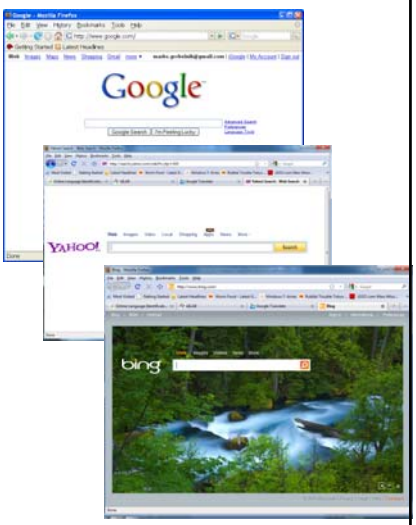


Četrti del: Zahtevnejše metode

- Vizualizacija tekstovnih podatkov
- Iskanje na spletu
- Vizualizacija novic
- Izdelava povzetkov
- Prekojezično povezovanje dokumentov
- Analiza socijalnih omrežji
- ....

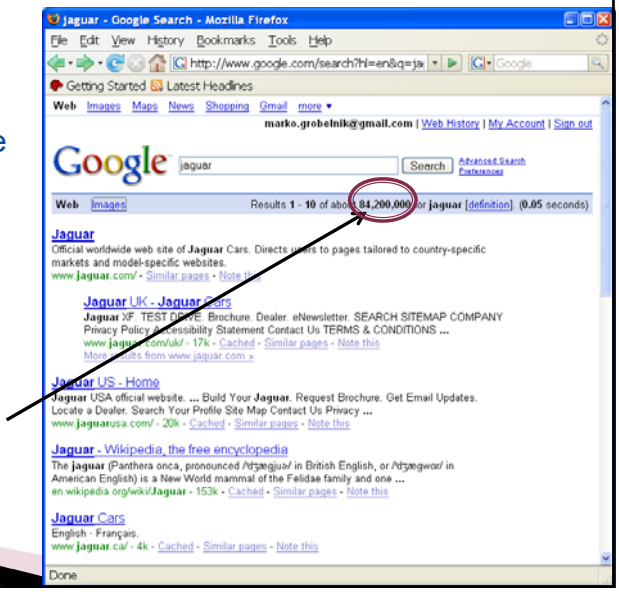
Iskanje na spletu

- iskanje na spletu – eno od nabolj pogostih opravil na spletu, ki vključuje delo z besedili
- …vendar – kako napredna je tehnologija iskanja danes?
  - …uporabna v večini primerov, vendar ne preveč napredna!



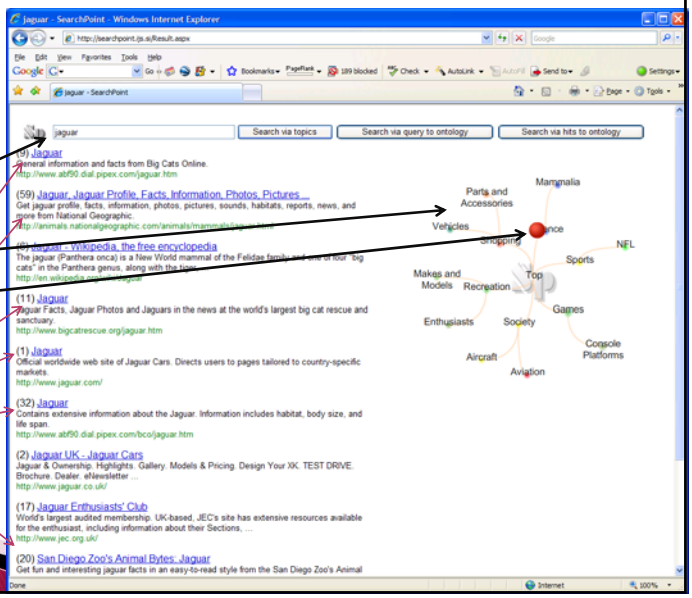
Primer – iskanje "jaguar"

- "**jaguar**" ima več pomenov
- …prva stran, ki jo vrne iskalnik vsebuje le majhen delež relevantnih strani
- …obstaja še veliko odgovorov (84 200 000 v podanem primeru)



Kontekstno občutljivo iskanje

http://searchpoint.ijs.si



Query

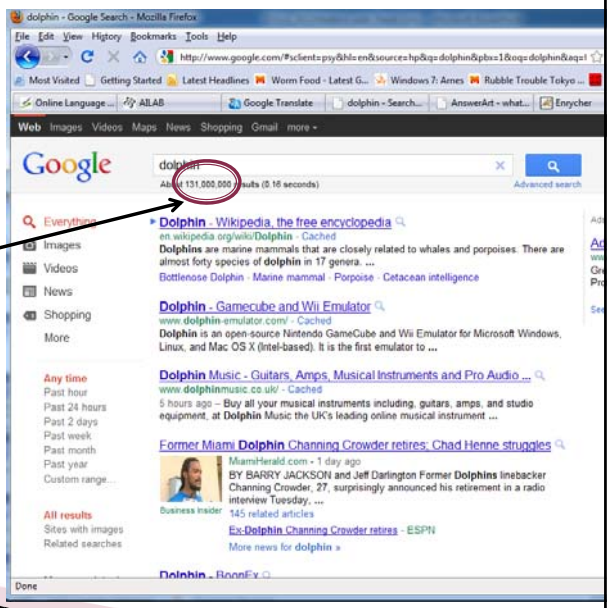
Conceptual map

Search Point

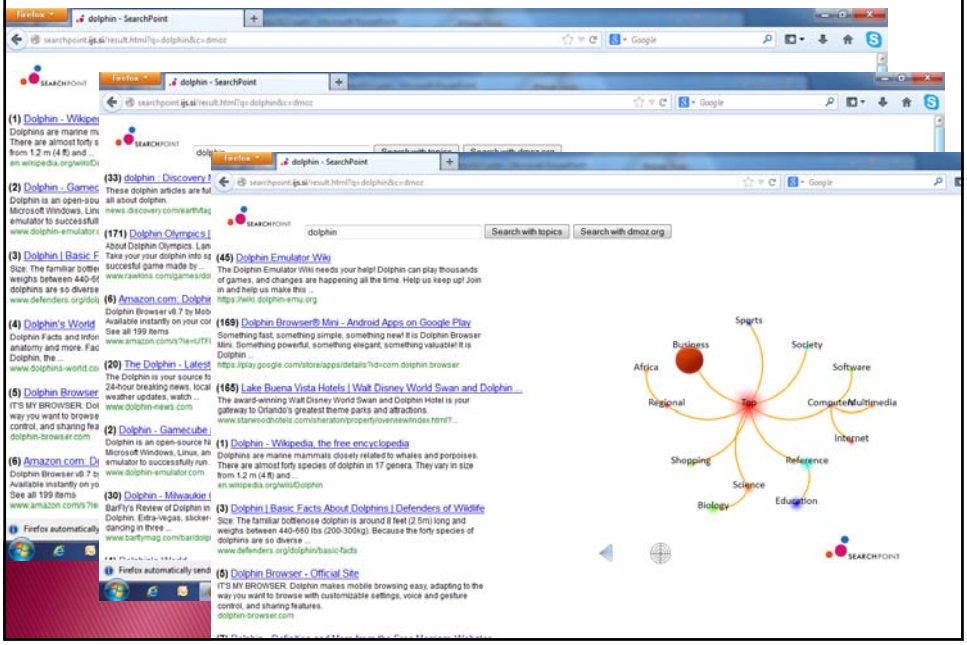
Dynamic contextual ranking based on the search point

Primer – iskanje "dolphin"

- Dvoumno vprašanje
- Poleg prvih nekaj odgovorov, ki jih vidimo, je še 131 000 000



Kontekstno občutljivo iskanje



Učenje na velikih omrežjih

- **Classification and Clustering in Large Complex Networks,** Tina Eliassi-Rad, Department of Computer Science, Rutgers, The State University of New Jersey http://videolectures.net/solomon_eliassi_rad_classification/(56 min)
- **Private traits and attributes are predictable from digital records of human behavior,** Thore Graepel, Microsoft Research, Cambridge, Microsoft Research http://videolectures.net/lsoldm2013_graepel_human_behavior/ (25 min)

Četrti del: Zahtevnejše metode

- Vizualizacija tekstovnih podatkov
- Iskanje na spletu
- Vizualizacija novic
- Izdelava povzetkov
- Prekojezično povezovanje dokumentov
- Analiza socijalnih omrežji
- ....

Vizualizacija novic

- Pomen posamezne novice veliko lažje razumemo, če poznamo kontekst (novice iz daljšega obdobja)
- Novice običajno omenjajo osebna imena ljudi, mest, podjetij,...
- Pri preiskovanju novic si lahko pomagamo z vizualizacijo konteksta omenjenih osebnih imen
  - Potrebno je identificirati osebna imena
  - Novice predstavimo kot vektorje besed

Identifikacija osebnih imen

"Several Countries Say the Bug Is in **Y2K** Reports From **Gartner**"
Baltimore Sun (11/27/99) P. 11C

Although the **Gartner Group** is considered a leading expert on **Y2K** readiness, some countries that received unfavorable ratings say the group's reports are inaccurate and have possibly harmed foreign investment. **South Africa**, for example, says international grain trader **Cargill** named a **Gartner** report as a factor in its decision not to deliver to **South Africa** for two weeks around Jan. 1. Later, **South Africa** received a positive rating from **Gartner**. "**Gartner Group** has a vested interest in stirring up panic," says **Jamaica**'s government **Y2K** coordinator **Luke Jackson**. "They're consultants. That's what they do." **Jackson** says **Gartner** never approached him in compiling the report. Likewise, **Ecuador**'s national **Y2K** coordinator **Jacqueline Herrera** says **Gartner** never called her before releasing a report that showed her country lagging in **Y2K** readiness. "The conclusions of this report are inaccurate," **Herrera** says. Meanwhile, **Gartner**, which maintains the confidentiality of its sources, supports its findings and says its information comes from thousands of its clients and other private companies.

Novico lahko predstavimo samo z osebnimi imeni, ki jih vsebuje

Izločanje informacij

Originalno besedilo novice

Gartner Group [7] – "Gartner Group", "Gartner"
Y2K [4] – "Y2K"
South_Africa [3] - South_Africa
Jacqueline Herrera [2] – "Jacqueline_Herrera", "Herrera"
Cargill [1] – "Cargill"
Jamaica [1] – "Jamaica"
Luke_Jackson [2] – "Luke Jackson", "Jackson"
Ecuador [1] – "Ecuador"

Različne pojavne oblike istih osebnih imen je potrebno identificirati in združiti

Postopek za identifikacijo osebnih imen

- Problem je zahteven, običajno se ga rešuje z ročno napisanimi pravili in seznami v kombinaciji s strojnim učenjem
- Enostavna rešitev sloni na uporabi velikih začetnic in nekaj hevristik za združevanje različnih pojavnih oblik istega osebnega imena

(Janez Drnovšek = Predsednik Drnovšek = Drnovšek)

Primer na množici novic

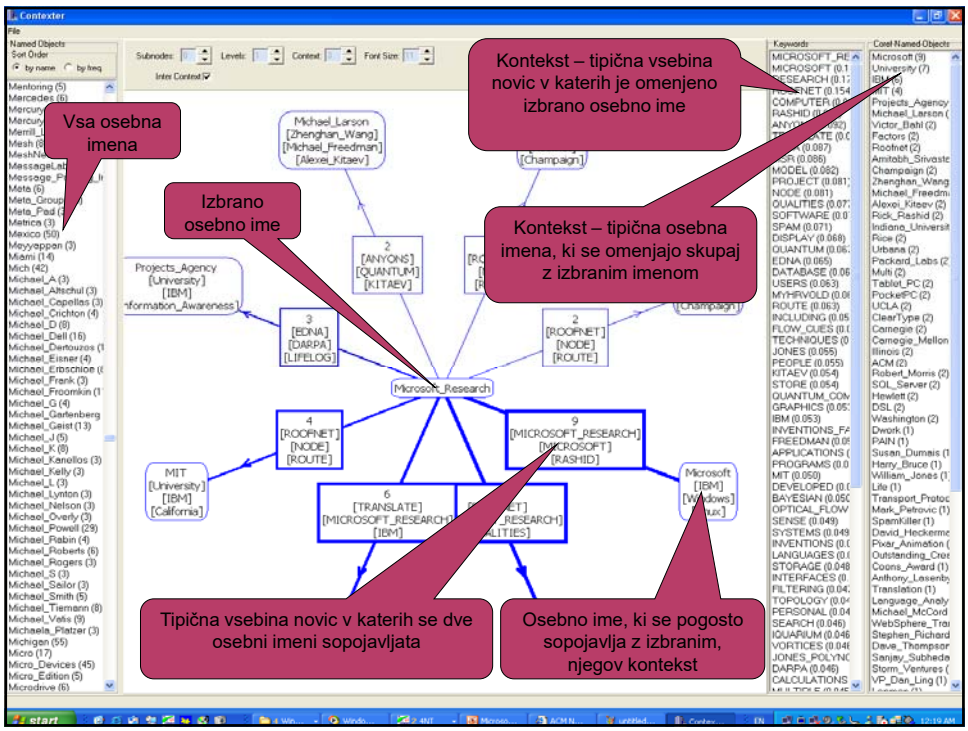
- ACM tehnološke novice
  - Javno dostopne na http://www.acm.org/technews/
  - …11000 news articles from December 1999
- Primer novice objavljene aprila 2004 (http://www.acm.org/technews/articles/2004-6/0409f.html#item1):

"Panelists Call for Lightweight Linux"
IDG News Service (04/07/04); McMillan, Robert

Linux complexity was a hot topic for discussion at this week's ClusterWorld Conference & Expo in San Jose, with Linux users and distributors divided on whether to make the open-source operating system more lithe or more complex. Linux is big in the high-performance computing community, having largely pushed aside proprietary vendors such as Cray. But despite claiming more than half of the spots in the Top500 Supercomputer Sites list, Linux clusters still can be improved on, according to conference panelists. Argonne National Laboratory computer scientist Rusty Lusk called for a simpler Linux that would be more modular and easy to manage architecturally. Currently, there are too many different software library versions, creating "dependency" problems and unnecessary complexity. Simple distributions that are easy to build on would be a tremendous boon to customers, said Henry Hall, president of systems integrator Wild Open Source. A tool that allowed users to track and audit changes made to the basic kernel is also important, he said. But Scyld Linux co-creator Donald Becker predicted Linux development would lead to a more complex operating system, albeit one that is easy to customize through componentization. No mainstream Linux distribution has yet addressed the unique needs of the high-performance computing community head-on, said SUSE's Timothy Beloney. SUSE parent firm Novell recently formed a team to work on the issue, and is especially looking at how to make Linux more customizable.
Click Here to View Full Article

Primer sistema za pregledovanje novic - Contexter

- Prikaže vsa osebna imena iz celotne množice novic
- Uporabnik z izbiro osebnega imena doseže
  - Vizualni prikaz konteksta v katerem se v novicah ime omenja
  - Izpis tipičnih besed iz konteksta



Obravnava strukturiranih in nestrukturiranih podatkov

- **Predicting Positive and Negative Links in Online Social Networks,** Jure Leskovec, Computer Science Department, Stanford University http://videolectures.net/www2010_leskovec_ppn/ (23 min)
- **Dealing with structured and unstructured data at Facebook,** Lars Backstrom, Facebook http://videolectures.net/eswc2011_backstrom_facebook/ (54 min)

Četrti del: Zahtevnejše metode

- Vizualizacija tekstovnih podatkov
- Iskanje na spletu
- Vizualizacija novic
- Izdelava povzetkov
- Prekojezično povezovanje dokumentov
- Analiza socijalnih omrežji
- ....

Metode za izdelavo povzetkov

- Problem: na osnovi podanega besedila izdelaj povzetek, ki odraža vsebino dokumenta
- Metode:
  - **Statistične, neodvisne od jezika besedila** – povzetek je množica stavkov izbranih iz celotnega dokumenta
  - **Uporabljajo lastnosti naravnega jezika –** semantična analiza za ugotovitev pomena besedila in generiranje novega, krjašega besedila, ki povzame pomen

Izdelava povzetkov z metodo izbiranja

- Pristop ima tri faze:
  1. Analiza originalnega besedila
  2. Določanje pomembnih delov besedila
  3. Izdelava primernega povzetka
- Večina metod uporablja linearno obteževanje – vsaka enota besedila (stavek) je ocenjen:

  Ocena(U)=PozicijaVBesedilu(U)+
  BližinaZnačilnihFraz(U)+Statistika(U)+
  DodatnaZastopanost(U)
- …povzetek je sestavljen iz nekaj najbolj ocenjenih stavkov iz podanega besedila

Izdelava povzetkov z upoštevanjem lastnosti naravnega jezika

- Vsaj delno moramo 'razumeti' pomen dokumenta
  - uporabimo lingvistično in semantično strukturo besedila
- Pogledali bomo primer pristopa, ki kombinira analizo naravnega jezika in strojno učenja:
  - Temelji na gradnji semantične mreže dokumenta in identifikaciji njenih pomembnih delov, ki potem tvorijo povzetek
  - Ocenjen je bil na standardni množici dokumentov in njihovih povzetkov ("Document Understanding Conference")
  - Dosega 70% priklic (recall) in 25% preciznost (precision) na izločenih trojicah subjekt-predikat-objekt

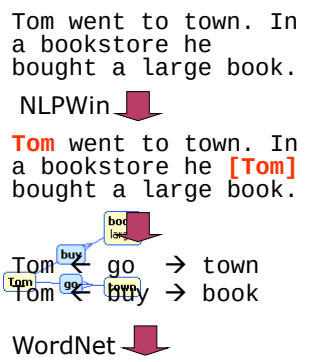
Osnovan ideja

(povzeto po nalogi Leskovec, maj 2004)

- Povzetek
  - trojice osebek – povedek – predmet
- Trojice bomo sestavili v semantično mrežo
  - Točke so osebki in predmeti
  - Povezave so glagoli
  - Tako zajamemo kontekst besedila
- Graf se bomo naučili porezati, da bi v njem ostale le pomembne točke

Postopek sumarizacije

1. Besedilo enega dokumenta razdelimo na stavke
2. Naredimo stavčno analizo
3. Povežemo imenske entitete
   - Ugotovili bi radi, da je Janez Drnovšek = Predsednik Drnovšek = Drnovšek
4. Razrešimo navezovanje zaimkov
   - Zaimke povežemo z imeni
5. Najdemo trojice Osebek – Povedek – Predmet
6. Iz trojice sestavimo graf
7. Trojico vsako opišemo z atributi
8. Z uporabo strojnega učenja trojice klasificiramo v povzetek



Atributi za opis trojice

- Osebek–Povedek–Predmet opišemo z:
  - Stavčna analiza (18(+100) atributov) – lokalne značilnosti:
    - Besedna vrsta, tip imena, globina v drevesu, biti,…
  - Graf (14 atributov) – globalni kontekst:
    - stopnja točke, PageRank, Hubs and Authorities,…
  - Ostalo (5 atributov):
    - Položaj stavka v besedilu, požaj in frekvenca besede
- Malo neničelnih atributov – v povprečju le 72 od 466 različnih atributov

Učenje izdelovanja povzetkov

- Učimo se klasifikacijskega pravila, ki nam pove, ali trojica sodi v povzetek ali ne
- Uporabimo metodo podpornih vektorjev (SVM)
- Različno obtežimo posamezne skupine atributov

Testna množica DUC 2002

- "Document understanding conference"
- Časopisni članki o 30 različnih tematikah
- V vsaki tematiki je 10 člankov
- Za vsak članek imamo
  … ročno narejen povzetek
  … izbranih nekaj stavkov, ki sodijo v povzetek
- Članki so dolgi (6000 znakov, 50 stavkov po 22 besed)

Učenje med tematikami

- Testiramo na enem dokumentu
- Izmed ostalih vzamemo *N* naključnih dokumentov in se na njih naučimo

število učnih dokumentov	Učna množica			Testna množica		
	natančnost	priklic	F1	natančnost	priklic	F1
10	33,48	88,44	48,58	23,05	64,67	33,99
20	31,26	86,45	45,92	24,49	68,69	36,11
50	29,42	82,53	43,37	25,75	72,81	38,04
100	28,64	79,91	42,16	26,25	73,22	38.64

Natančnost = TP / (TP + FP)
Priklic = TP / (TP + FN)
F1 = 2 N P / (N + P)

Kako resne so napake?

Učenje znotraj tematike

- Vzamemo vse dokumente (okrog 5), ki govorijo o isti tematiki
- Na enem testiramo, na ostalih se učimo

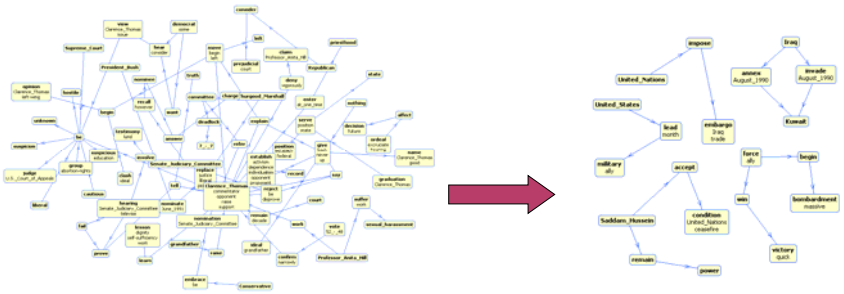
Učna množica			Testna množica		
natančnost	priklic	F1	natančnost	priklic	F1
36,49	90,63	52,03	23,60	60,05	33,89

Ni dovolj učnih primerov

Izgradnja modela

- S strojnim učenjem zgradimo model, ki ugotovi katere trojice subjekt-predikat-objekt sodijo v povzetek
- Uporabljena je metoda podpornih vektorjev (Support Vector Machine - SVM)

Semantična mreža dokumenta | **Povzetek semantične mreže**



Primer avtomatsko narejenega povzetka

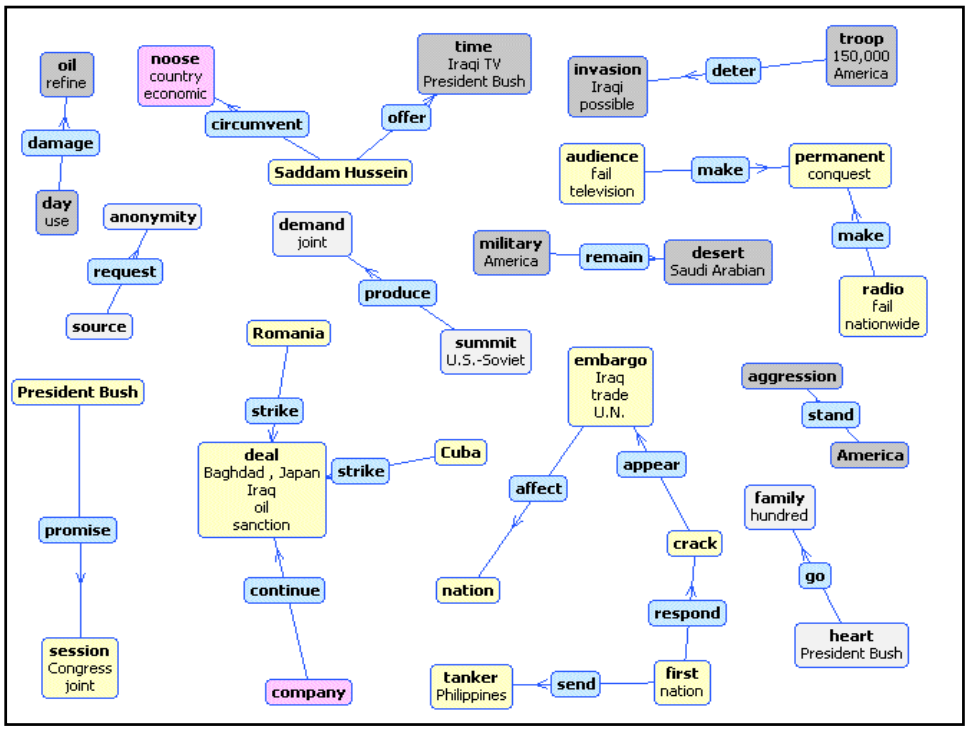
Cracks Appear in U.N. Trade Embargo Against Iraq.

Cracks appeared Tuesday in the U.N. trade embargo against Iraq as Saddam Hussein sought to circumvent the economic noose around his country. ... increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that it is not doing its share to oppose Baghdad, Japan ... to nations most affected by the U.N. embargo on Iraq. President Bush on Tuesday night promised a joint session of Congress and a nationwide radi... Hussein will fail" to make his conquest of Kuwait permanent. ``America must stand up to aggression, and we will," said Bush, who added that the U... Arabian desert indefinitely. ``I cannot predict just how long it will take to convince Iraq to withdraw from Kuwait," Bush said. More than 150,000 U.S. troops have been s... to the Persian Gulf region to deter a possible Iraqi invasion of Saudi Arabia. Bush's aides said the president would follow his address to Congress with a televised message for the Iraqi p...ple, declaring the world is united against their government's invasion of Kuwait. Saddam had offered Bush time on Iraqi TV. The Philippines and Namibia, the first of the developing ...ions to respond to an offer Monday by Saddam of free oil _ in exchange for sending their own tankers to get it _ said no to the Iraqi leader. Saddam's offer was seen as a none-too-st... attempt to bypass the U.N. embargo, in... ...ent survey, Cuba and Romania have st... ...ation. The report, made available... ...ran told The Associated Press that ... ...cial comment from Tehran or Bagh... ...it Sunday to Tehran, the first by a ... ...industry sources in the region, contact... ...e because of damages to refineries ... ...up for the damage Iraq inflicted ... ...o days after the U.S.-Soviet summit t... ...ry specialists from Iraq, where they ... ...edize the 5,800 Soviet citizens in Iraq ... ...hange, and it will not change. America ... ...off the face of the Earth." In other c... ...em women and children leaving their ... ...altimore from Iraq. Evacuees spoke ... ...order," said Thuraya, 19, who wo... ...had told U.S. officials that America... ...ot known how many men the Iraqi n... ...ia. He said there was little indication h... ...onth estimate generally used by gov... ...sed disgruntled lawmakers ``a signific... ...the crisis in the gulf, said Tuesday at t... ...," said Hiroyasu Horio, an official with... ...etary Fund, and $600 million would b... ...lan and Turkey. Japan has already pro... ...-military uses. But critics in the Unit... ...'s constitution bans the use of force ... ...ered developing nations free oil if they ... ...fulfilled its oil requirements, and Namibia said it would not ``sell its sovereignty" for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of free oil as a ``propaganda ploy." Venezuela, an OPEC member, has led a drive among oil-producing nations to boost production to make up for the shortfall caused by the loss of Iraqi and Kuwaiti oil from the world market. Their oil makes up 20 percent of the world's oil reserves. Only Saudi Arabia has higher reserves. But according to the State Department, Cuba, which faces an oil deficit because of reduced Soviet deliveries, has received a shipment of Iraqi petroleum since U.N. sanctions were imposed five weeks ago. And Romania, it said, expects to receive oil indirectly from Iraq. Romania's ambassador to the United States, Virgil Constantinescu, denied that claim Tuesday, calling it ``absolutely false and without foundation.''.

Povzetek, ki ga je napisal **človek**

Cracks appeared in the U.N. trade embargo against Iraq. The State Department reports that **Cuba and Romania have struck oil deals with Iraq** as others attempt to trade with Baghdad in defiance of the sanctions. Iran has agreed to exchange food and medicine for Iraqi oil. **Saddam has offered developing nations free oil if they send their tankers to pick it up.** Thus far, none has accepted.
Japan, accused of responding too slowly to the Gulf crisis, has promised $2 billion in aid to countries hit hardest by the Iraqi trade embargo. **President Bush has promised that Saddam's aggression will not succeed**.

7800 znakov, 1300 besed



Izdelovanje povzetkov

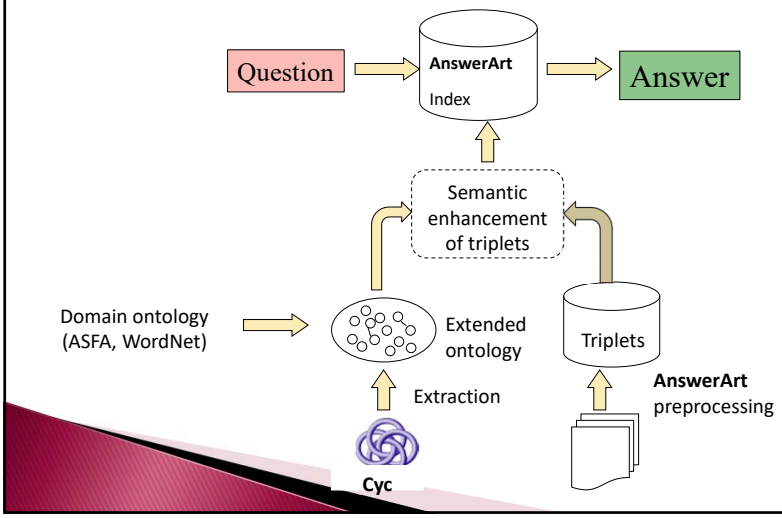
- **A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization,** Lu Wang, Department of Computer Science, Cornell University

http://videolectures.net/acl2013_wang_sentence/ (22 min)

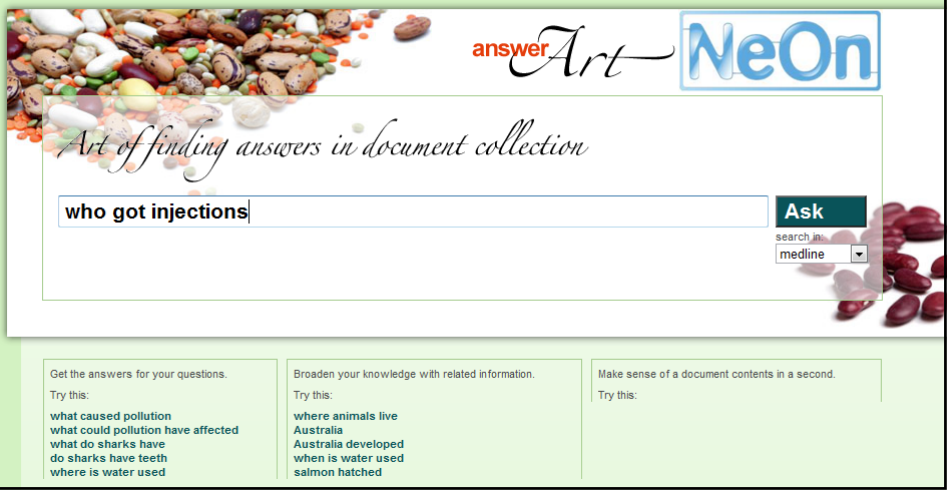
- **Towards Robust Abstractive Multi-Document Summarization,** Jackie Chi Kit Cheung, Department of Computer Science, University of Toronto

http://videolectures.net/acl2013_cheung_summarization/ (24 min)

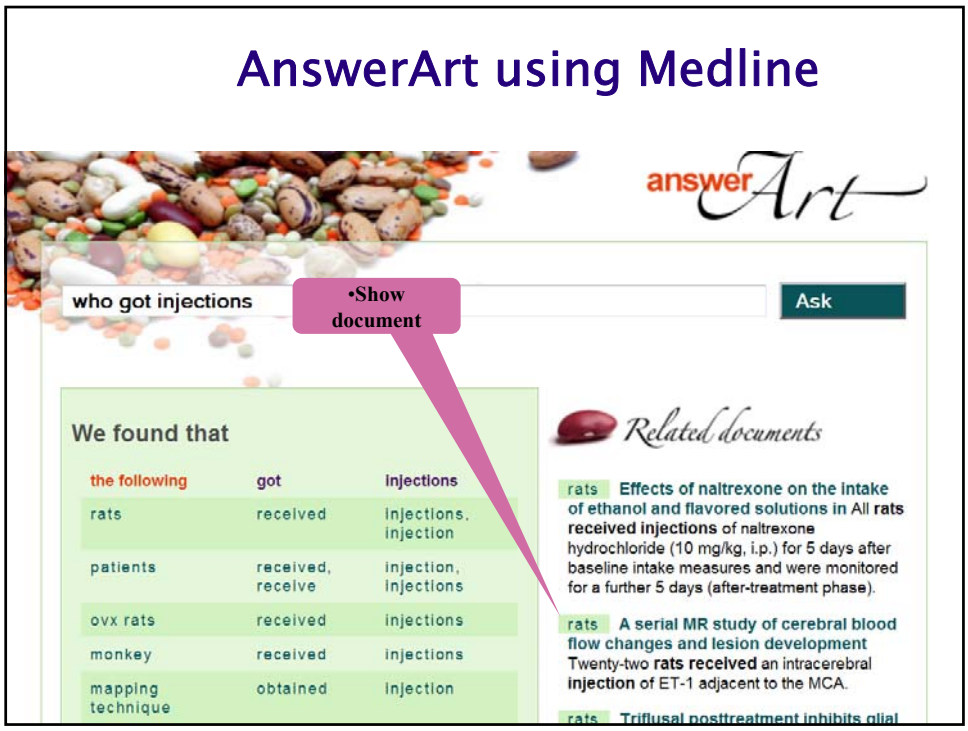
AnswerArt – Arhitektura sistema

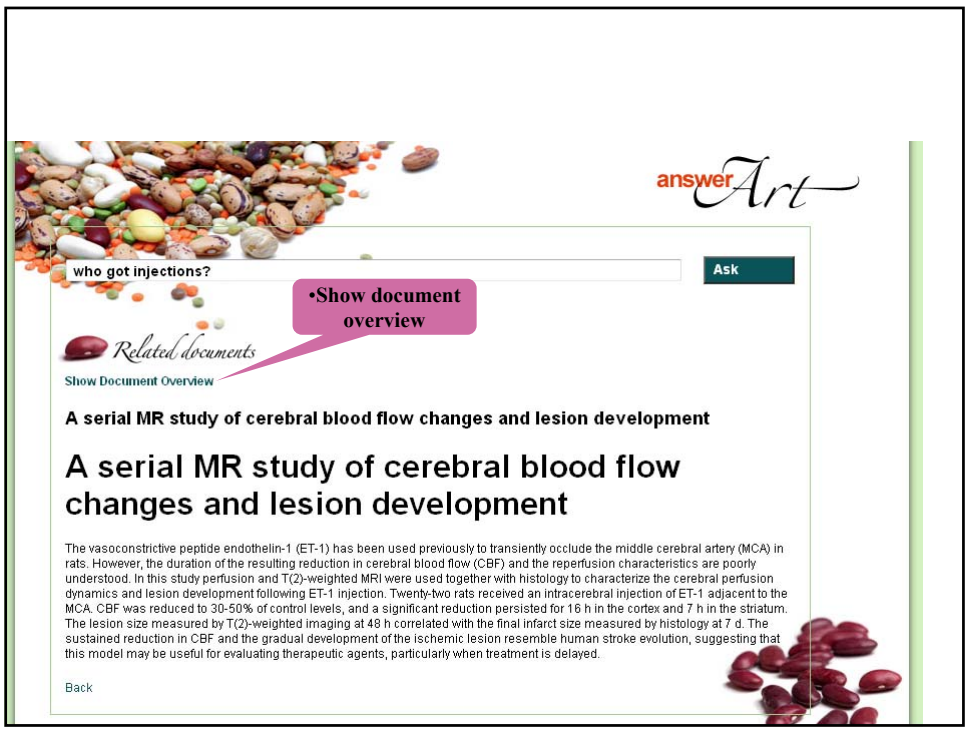


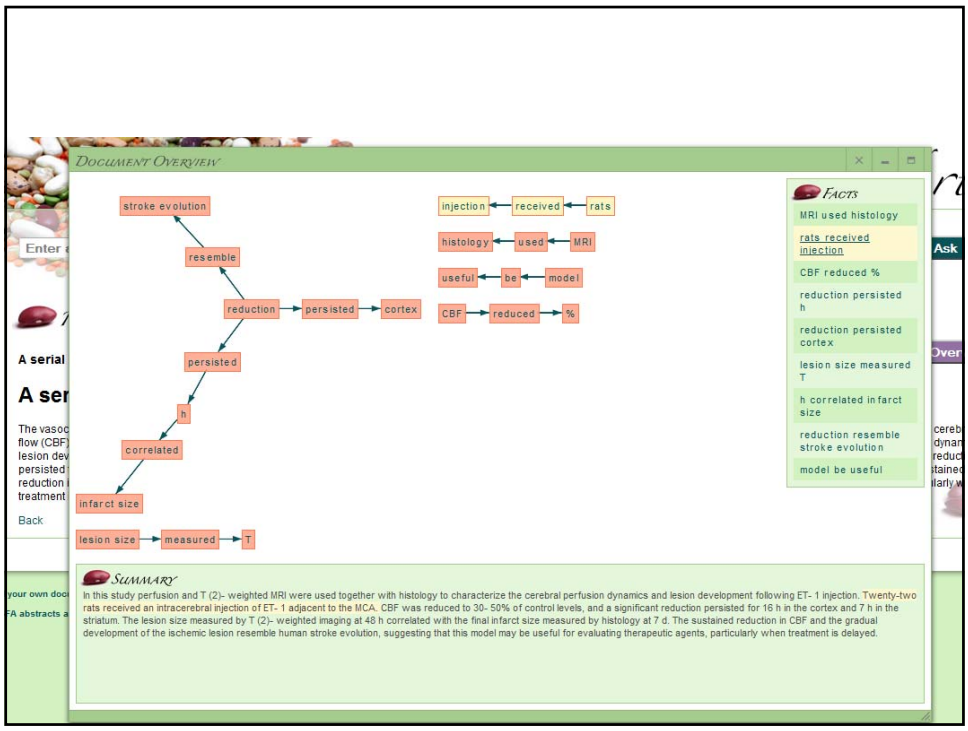
AnswerArt using Medline

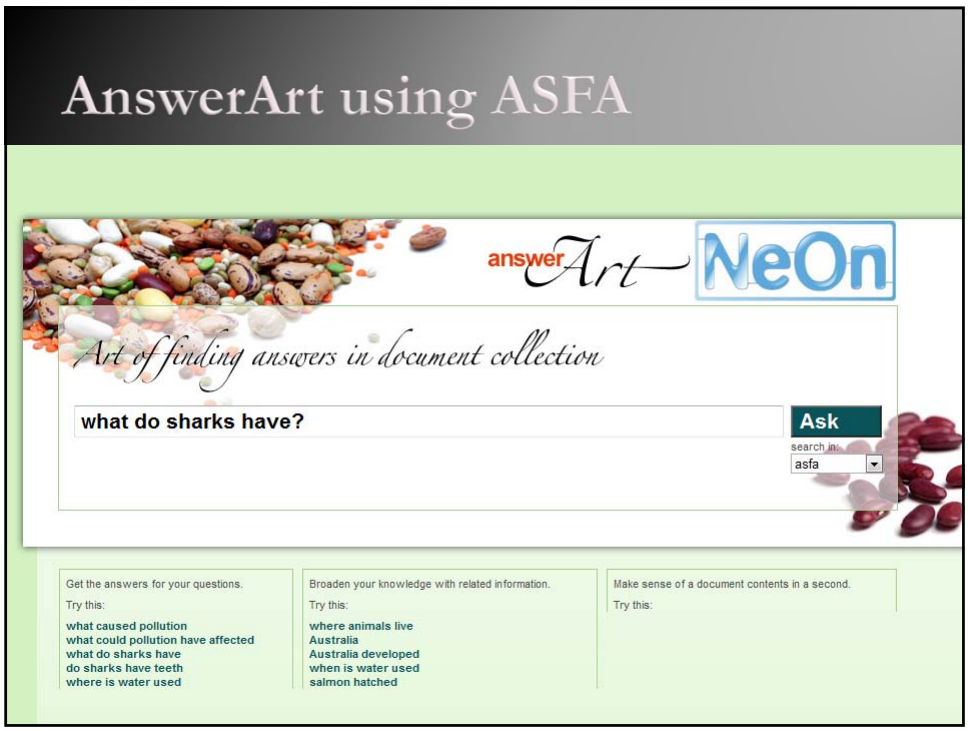


AnswerArt using Medline

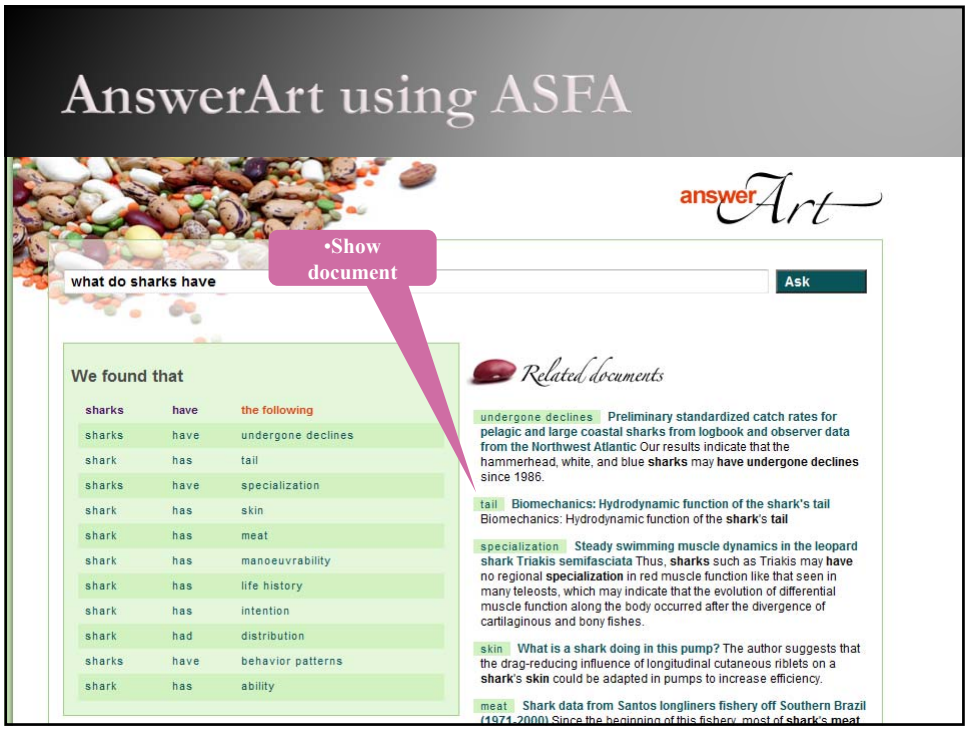




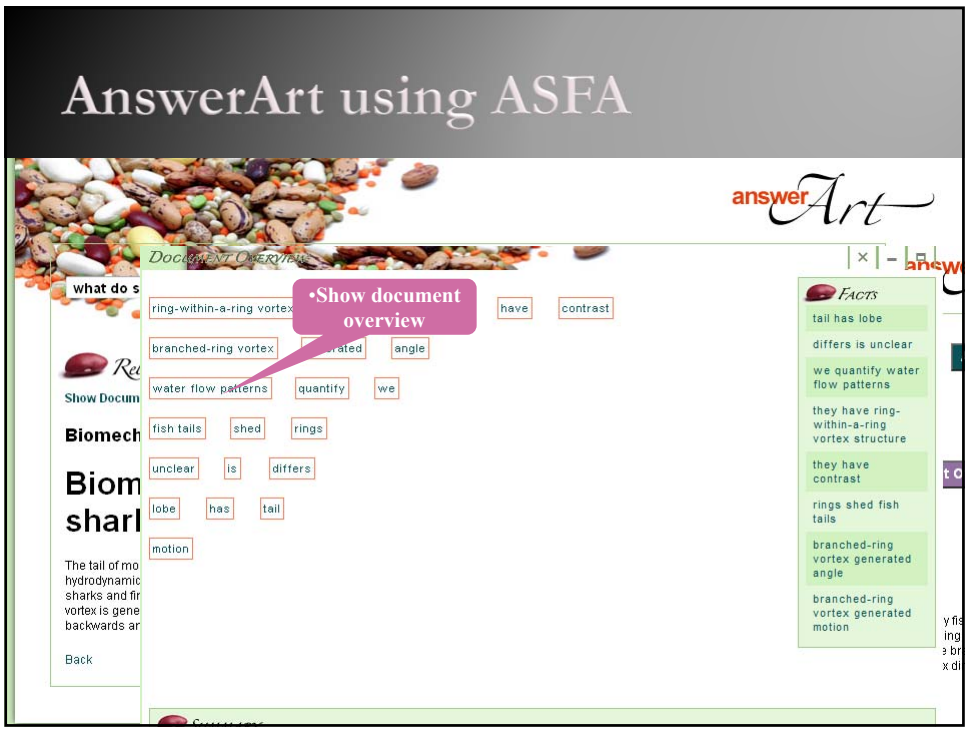

Document Overview
stroke evolution
resemble
reduction
persisted
cortex
persisted
h
correlated
infarct size
lesion size
measured
T
injection
received
rats
histology
used
MRI
useful
be
model
CBF
reduced
%
Facts
MRI used histology
rats received injection
CBF reduced %
reduction persisted h
reduction persisted 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.
Back


AnswerArt using ASFA
answerArt
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



AnswerArt using ASFA



Povzetek dosedanje snovi

Učenje na tekstovnih podatkih

- Vizualizacija tesktovnih podatkov
- Iskanje po spletu
- Izdelava povzetkov

Četrti del: Zahtevnejše metode

- Vizualizacija tekstovnih podatkov
- Iskanje na spletu
- Vizualizacija novic
- Izdelava povzetkov
- Prekojezično povezovanje dokumentov
- Analiza socijalnih omrežji
- ....

Večjezična analiza pomena stavkov

- Globoko učenje na 24 000 označnih stavkih iz baze ATOMIC
- vsak stavek je označen s
  - devet kej-potem tipov relacij za opis vzrokov in učinkov akcij glede na akterja in opazovalce
    - xIntent – Because PersonX **wanted**
    - xNeed – Before, PersonX **needed**
    - xAttr – PersonX is **seen as**
    - xReact – As a result, PersonX **feels**
    - xWant – As a result, PersonX **wants**
    - xEffect – PersonX **then**
    - oReact – As a result, others **feel**
    - oWant – As a result, others **want**
    - oEffect – Others **then**

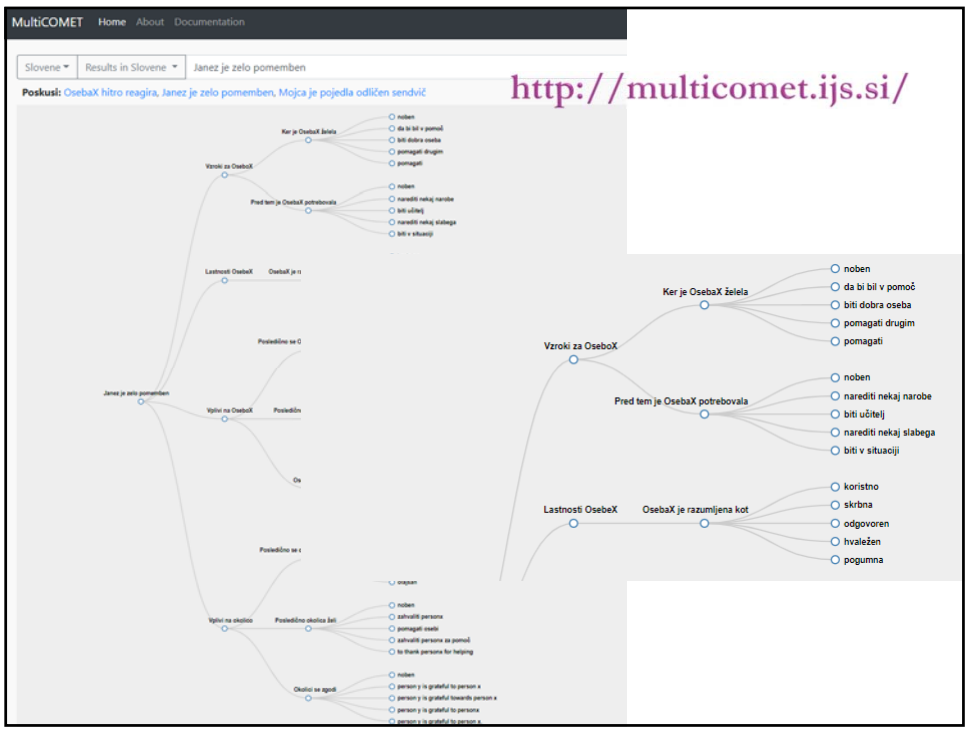
Mladenic Grobelnik, A., Mladenic, D., Grobelnik M., MultiCOMET – Multilingual Commonsense Description , SiKDD-2020

Rezultati poskusov

- Evalvacija na vzorcu 100 stavkov
- Primerjava rezultatov slovenskega modela z angleškim modelom
- Opazimo, da se uspešnost deskriptorjev se razlikuje

	Precision	Recall	F1
xIntent	0.324	0.324	0.
xNeed	0.352	0.352	0.
xAttr	0.438	0.438	0.
xReact	0.716	0.716	0.
xWant	0.21	0.21	0.
xEffect	0.456	0.456	0.456
oReact	0.706	0.706	0.706
oWant	0.468	0.468	0.468
oEffect	0.31	0.31	0.31
Average	0.442222	0.442222	0.442222

	Original	Translated/Predicted
Sentence	PersonX avoids every ___	PersonX se izogiba vsakemu ___
Descriptor Values	to stay away from people	to get away from others
	to avoid trouble	to make sure they are ok
	to stay away	to get away from the situation
	to not get caught	to be alone
	to not be noticed	to make a decision



Prekojezično povezovanje dokumentov

angleščina

španščina

Podobnost

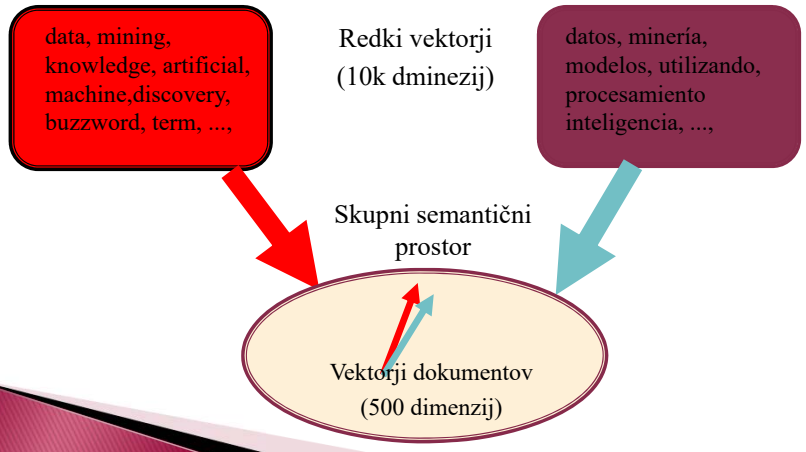
Primerljiva besedila: Wikipedia

Wikipedia poravnava

Predsatvitev besedil z vektorji

Predprocesiranje:

- Brišemo pogoste besede, TFIDF vektorji, normalizacija



Prekojezično povezovanje besedil

- **Cross-Lingual Document Retrieval through Hub Languages,** Jan Rupnik, Artificial Intelligence Laboratory, Jožef Stefan Institute

http://videolectures.net/nipsworkshops2012_rupnik_hub_languages/ (13 min)

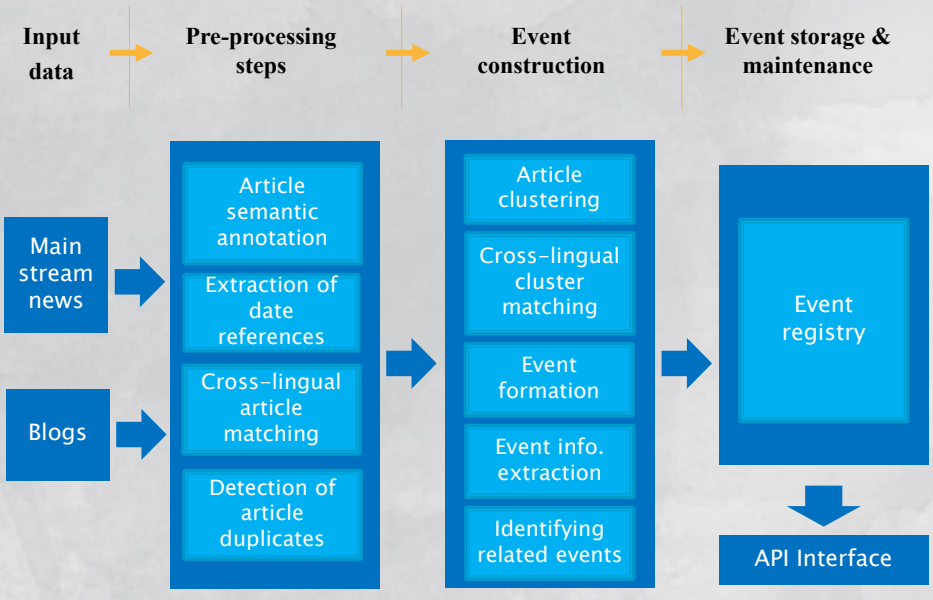
- **How much are Computers able to Understand text?,** Marko Grobelnik, Artificial Intelligence Laboratory, Jožef Stefan Institute

http://videolectures.net/i2010conf_grobelnik_hmcut/ (17 min)

Event extraction from news

- News articles in different languages
- Preprocessed and annotated enabling cross-lingual article matching
- Event extraction by cross-lingual clustering
- Enable advanced search and rich visualization options

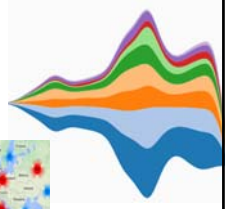
Event extraction pipeline



Related Systems/Demos

- NewsFeed (http://newsfeed.ijs.si/)
  - News and social media crawler
- Enrycher (http://enrycher.ijs.si/)
  - Language and Semantic annotation
- SearchPoint (http://searchpoint.ijs.si/)
  - Contextualized search
- XLing (http://xling.ijs.si/)
  - Cross-lingual document linking and categorization
- Event Registry (http://eventregistry.org/)
  - Event detection and topic tracking

•Sable SearchPoint



Collecting global media data

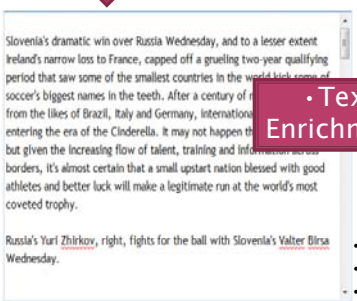
- Data collection service News-Feed
  - http://newsfeed.ijs.si/
  - …crawling global main-stream and social media
- Monitoring
  - ~70k main-stream publishers (RSS feeds + special feeds)
  - ~250k most influential blogs (RSS feeds)
  - free Twitter feed
- Data volume: ~350k articles & blogs per day (+5M tweets)
- Languages: eng (50%), ger (10%), spa (8%), fra (5%),…

How can we annotate a document?

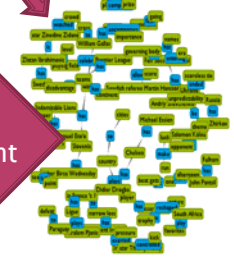
Enrycher (http://enrycher.ijs.si/)

•Plain text

•Extracted graph of triples from text



• Text Enrichment



entities

- Brazil
- Italy
- Germany
- Cinderella
- Paris
- John O'Shea
- Manchester United
- Robbie Keane
- Shay Given
- Greece
- Portugal
- Bosnia-Herzegovina
- Cristiano Ronaldo
- Uruguay

keywords

Sports, Soccer, CONCACAF, Competitions, United States, Sports and Hobbies, Kids and Teens, World Cup, Women,

categories

- Top/Kids_and_Teens /Sports_and_Hobbies /Sports/Soccer
- Top/Sports/Soccer /Competitions
- Top/Sports/Soccer /Competitions/World_Cup
- Top/Sports/Soccer /CONCACAF

•**Diego Maradona Semantics**:
•**owl:sameAs:** http://dbpedia.org/resource/Diego_Maradona
•**owl:sameAs:** http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrcN5Y29ycA
•rdf:type: http://dbpedia.org/class/yago/ArgentinaInternationalFootballers
•rdf:type: http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly
•rdf:type: http://dbpedia.org/class/yago/ArgentineFootballManagers
•rdf:type: http://dbpedia.org/class/yago/ArgentineFootballers
•**Robbie Keane Semantics**:
•**owl:sameAs:** http://dbpedia.org/resource/Robbie_Keane
•rdf:type: http://dbpedia.org/class/yago/CoventryCityF.C.Players
•rdf:type: http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly

•"**Enrycher**" is available as
•as a web-service generating
•Semantic Graph, LOD links,
•Entities, Keywords, Categories,
•Text Summarization, Sentiment

Enrycher Architecture

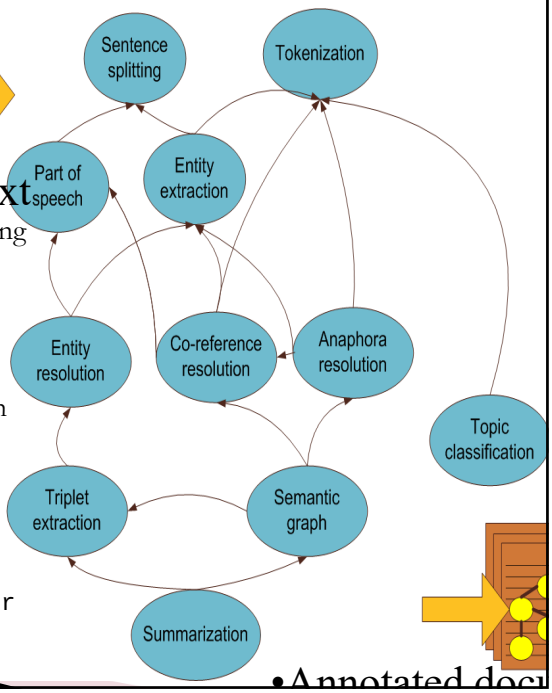
•Plain text

Enrycher is a web service consisting of a set of interlinked modules

- covering lexical, linguistic and semantic annotations
- exporting data in XML or RDF

To execute the service, one should send an HTTP POST request, with the raw text in the body:

- `curl -d "Enrycher was developed at JSI, a research institute in Ljubljana. Ljubljana is the capital of Slovenia." http://enrycher.ijs.si/run`



•Annotated docu

Pre-processing of articles

- Language independent annotation using Wikipedia
  - „…president left the White House to …“
  - http://en.wikipedia.org/wiki/White_House
  - „…un asesor de la Casa Blanca, ha…“
- Identification of date references to get event date
  - several regular expressions for each language
  - Single dates (2013/5/3), date ranges („Jun 3 - Aug 11, 2011), partial dates (June 2013)
- Cross-lingual similarity of articles

Detection of article duplicates

- Often an article is (almost) a copy of some previous article
  - Some news publishers just copy other ones
  - The same news publisher republishes slightly corrected version of existing news article
- Duplicates are detected and marked as such
  - Important for article clustering

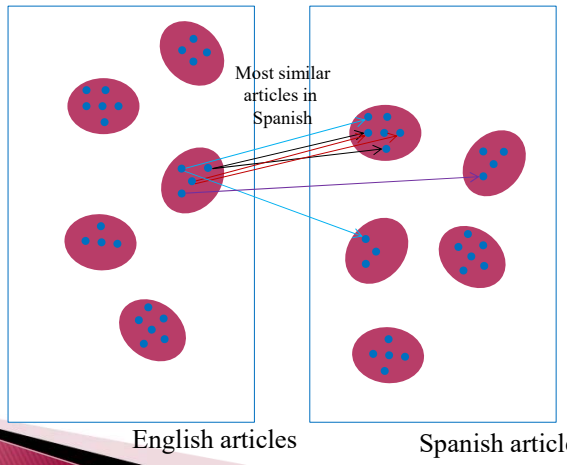
Article clustering

Identify articles that describe a single event

- Online clustering algorithm
- Grouping based on article title + article content + detected named entities
- Procedure:
  - Each new article is assigned to the closest cluster
  - Every once in a while we check if some clusters need to be split or merged
  - Old clusters are removed

Cross-lingual cluster linking

- Clusters in different languages can describe the same event
- Consider similarity of relevant concepts and date of articles

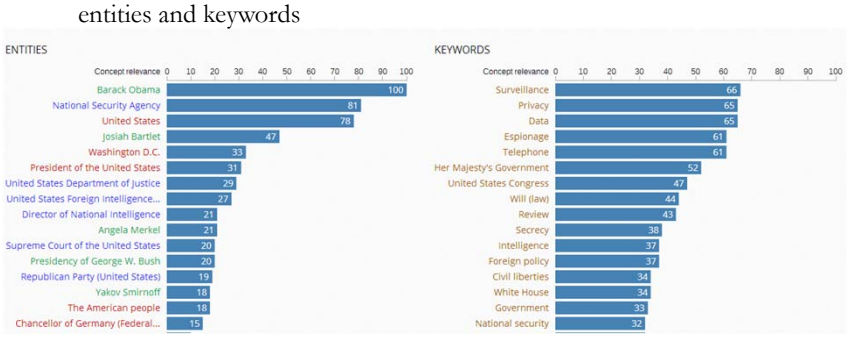


Event formation from text stream

- Event is formed from one or more linked clusters
  - as clusters evolve, they can be added or removed from the event
- Each event is assigned a unique id
- Extract event information using the articles
  - to answer questions what, when, where, who
  - title and the 1st paragraph of the medoid article
  - Date - the most frequent or average article date

Event information extraction

- Check the annotations of the articles to identify frequently occuring entities and keywords



- Event location
  - GeoNames to determine the top entity that represents a location
- Event categorization (sports, bombing attacks, earthquakes, …)
  - DMoz taxonomy for classifying articles

Event Registry

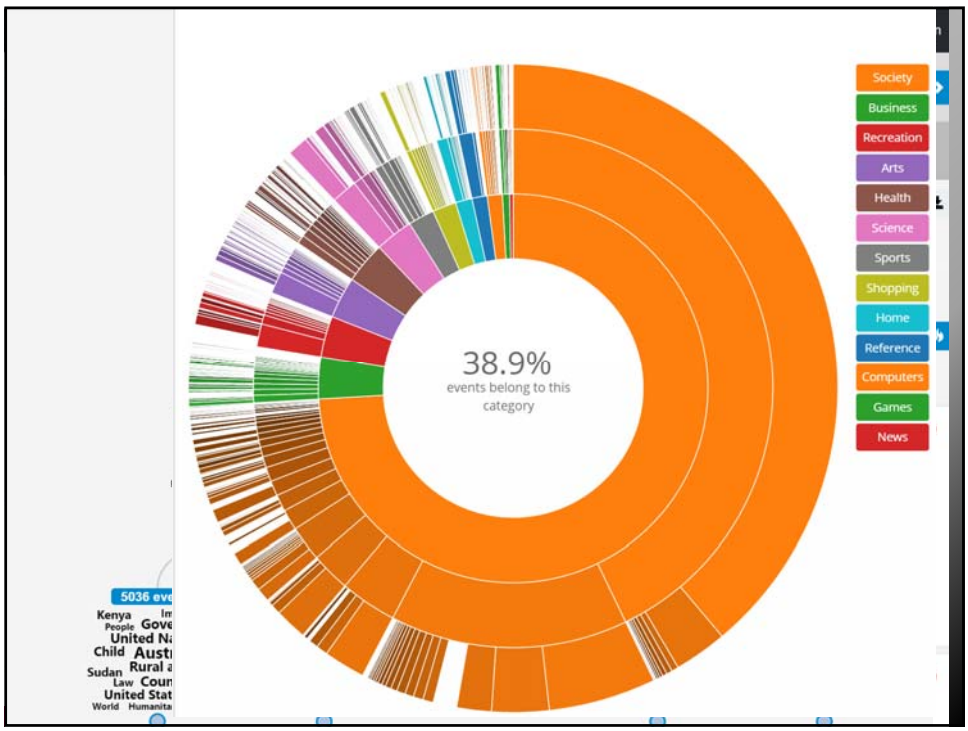
Event Registry http://eventregistry.org/

- Database of all detected events + extracted information about them
- Provides API to search for events

- Event data is also provided in structured form
  - Use of BBC Storyline ontology
- SPARQL endpoint:
  - http://eventregistry.org/rdf/search

Event Registry system for global media monitoring (http://eventregistry.org)

- Having a stream of news & social media, the task is to structure documents into events
- Event Registry allows for:
  - Identification of events from documents
  - Connecting documents across many languages
  - Tracking events and constructing story-lines
  - Describing events in a (semi)structured way
  - UI for exploration through Search & Visualization
  - Export into JSON/RDF (Storyline ontology)



Četrti del: Zahtevnejše metode

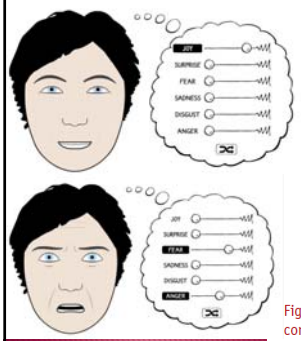
- Vizualizacija tekstovnih podatkov
- Iskanje na spletu
- Vizualizacija novic
- Izdelava povzetkov
- Prekojezično povezovanje dokumentov
- Analiza socijalnih omrežji
- ....

Socijalna omrežja

- Twitter (140 mil users, 400 mil tweets/day) storing the data for use in real-time applications
- zgradimo graf, tako da so vozlišča uporabniki, povezave pa relacije med njimi: n.pr., dva uporabnika komunicirata, uporabnik A naredi "retweet" uporabnika B, uporabnik A sledi uporabniku B
- http://www.sysomos.com/insidetwitter/twitter-stats-2010/

Ocenjevanje razpoloženja družbe

- 484 millionov tweetov (julij 2009 – januar 2012, 54 večjih mest v VB) – analiza za določanje razpoloženja (sentiment analysis)
- Osnovne emocije (radost, žalost, strah, jeza) so povezali s seznamom besed iz baze WordNet-Affect http://wordnet.princeton.edu
- http://www.grimace-project.net/



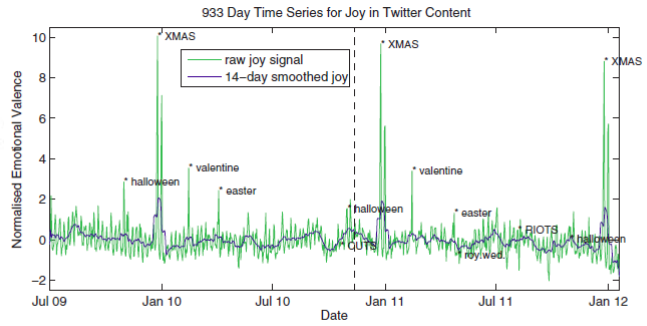


Figure 2. Plot of the time series representing levels of joy estimator over a period of 2½ years. Notice the peak corresponding to Christmas and New Year, Valentine's day and the Royal Wedding

Analiza velikih količin spletnih podatkov

- **Large Scale Learning at Twitter,** Aleksander Kołcz, Twitter, Inc. http://videolectures.net/eswc2012_kolcz_twitter/ (50 min)

Analiza omrežji

- **Deconvolution of Networks into Communities,** Jure Leskovec, Computer Science Department, Stanford University http://videolectures.net/kdd2013_leskovec_online_communities/ (23 min)
- **Graph Identification,** Lise Getoor, University of Maryland http://videolectures.net/solomon_getoor_gid/ (68 min)