Obdelava naravnega jezika

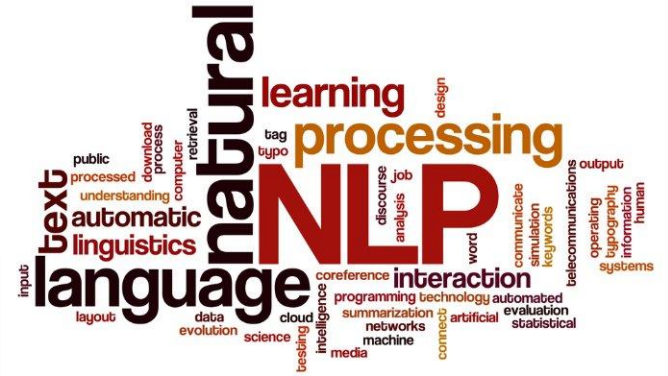
prof. dr. Marko Robnik-Šikonja
november 2016

Predavatelj

- izr. prof. dr. Marko Robnik Šikonja
- marko.robnik@fri.uni-lj.si
- Laboratorij za kognitivno modeliranje
- FRI, Večna pot 113, drugo nadstropje, desno od dvigala
- (01) 4798 241
- govorilne ure:
  - četrtek od 11h-12h
  - tudi sicer, bolj zanesljiva je najava
- http://www.fri.uni-lj.si/rmarko
- raziskovalno delo: podatkovna analitika, podatkovno rudarjenje, strojno učenje, umetna inteligenca, procesiranje naravnega jezika, algoritmi in podatkovne strukture, praktična uporaba

Umetna inteligenca

Načrt teme



- razumevanje jezika in inteligenca
- simboličen in statističen pristop k analizi jezika na primerih
- jezikovni viri in orodja
- pomembne naloge in potrebne jezikovne tehnologije
  - pridobivanje informacij: indeksiranje besedil, iskanje, rangirano iskanje, evalvacija
  - podatkovno rudarjenje: kateri dokumenti so pomembni, PageRank, Personal PageRank, prepoznavanje nanašanja
  - povzemanje
- primeri uporabe jezikovnih tehnologij:
  - analiza sentimenta,
  - priporočanje člankov

Razumevanje naravnega jezika

- razumevanje naravnega jezika je eden velikih izzivov (ne le) umetne inteligence

  kdo lahko razume moj svet?
  mu sam sploh lahko sledim?
  hodim in iščem, sem mar zaklet?
  čakam, da se zbudim?

- ne le poezija, tudi navodila za uporabo , časopisni članki, seminarske naloge, forumi, tviti…

Primer: pravila

- Študijska pravila FRI, člen 18
  Predčasno opravljanje izpitov lahko na prošnjo študenta dovoli prodekan za pedagoško dejavnost v soglasju z učiteljem, ki je nosilec predmeta, če so podani upravičeni razlogi (odhod na študij ali študijsko prakso v tujino, hospitalizacija v času izpitnega obdobja, porod, udeležba na strokovni ali kulturni prireditvi oz. vrhunskem športnem tekmovanju ipd.) in če glede na uspehe prosilca v preteklem študiju oceni, da je tako dovoljenje smotrno.

Primer: prevedena pravila

Article 18 of FRI Study Rules and Regulations

Taking exams at an earlier date may be allowed on request of the student by the Vice-Dean of Academic Affairs with the course convener's consent in case of mitigating circumstances (leaving for study or placement abroad, hospitalization at the time of the exam period, giving birth, participation at a professional or cultural event or a professional sports competition, etc.), and if the applicant's study achievements in previous study years are deemed satisfactory for such an authorization to be appropriate.

Razumevanje na nivoju računalnikov

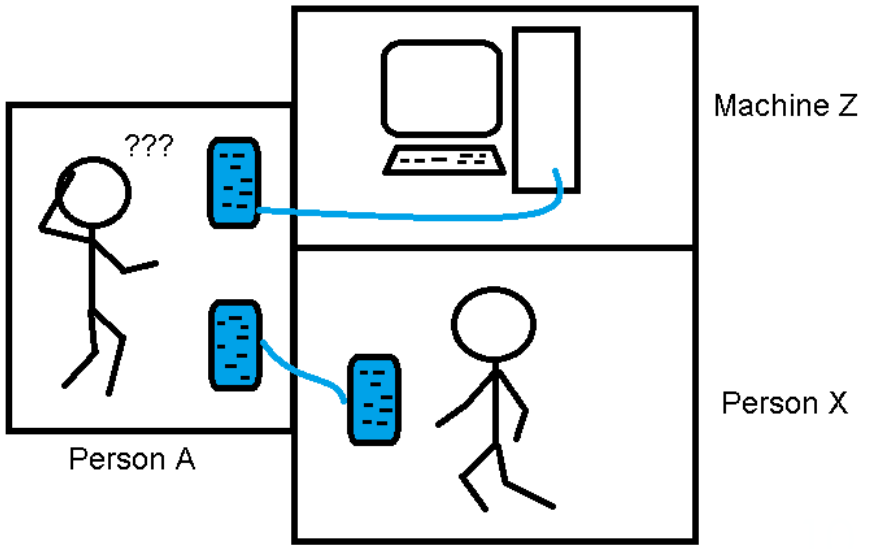
- za razumevanje pisanega stavka je potrebno poznavanje besed, sintakse, semantike in konteksta, v katerem besede nastopajo, pa tudi sklepanje o piščevih ciljih, znanju in njegovih predpostavkah
- *Prijatelja so zaprli, zato sem bil vesel.*
- Dvoumnost, npr. časopisni naslovi:
  - *Romana je srečala Abrahama*
  - *Kamerunec in Nigerijec padla na Obrežju*
  - *Jutri bo Slovenija spet rdeča*
  - *HRVAŠKO PODJETJE V LOVU ZA SLOVENSKIM ŽITOM*
  - *Dobra žena odnesla domov vodomca*
  - Newspaper headlines:
    - JUVENILE COURT TO TRY SHOOTING DEFENDANT
    - KIDS MAKE NUTRITIOUS SNACKS
    - MINERS REFUSE TO WORK AFTER DEATH…

Sintaktična in semantična dvoumnost

- sintaksa:
  - The girl told the story cried.
  - John went to the bank.
  - Put the box on the table in the kitchen.
- *zapis v iskalnik: natural language processing, naravni jezik, jaguar, Paris*

Razumevanje jezika kot test (umetne) inteligence

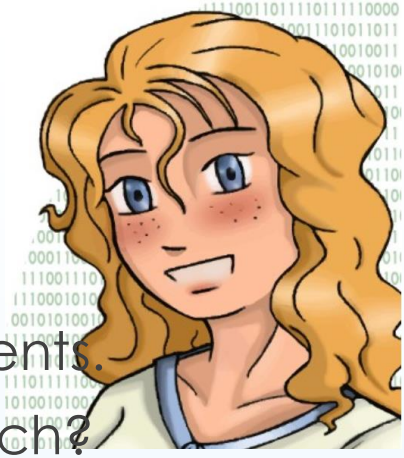
- umetni sistem, ki bi lahko razumel naravni jezik na nivoju človeka, bi moral vsebovati velik del človeškega znanja in bi dosegel skoraj vse cilje umetne inteligence, njegovih odgovorov pa bi ne mogli ločiti od človekovih
- Turingov test, Loebnerjeva
  ALICA, chatterbots, VIDA,
  (razpoznavanje govora, zo
  delegiranje spletnih servis



Primer razgovora z ALICE

- judge: Hello, I want to show your abilities to my students
- ALICE: Hi there!. Interesting. I want to find out more about you.
- judge: Let us better talk about the technology driving you
- ALICE: That sounds good to me.
- judge: Do you know anything about it or AI or NLP?
- ALICE: You're so smart, you tell me what it is.
- judge: Let us just do a test
- ALICE: I don't know if it's in my power to give permission.
- judge: Do you use parsing trees at all?
- ALICE: I use it as often as possible.

Mitsuku Chatbook (2013 and 2016 winner of Loebner bronze medal):



- Judge: Hello, I want to show your abilities to my students.
- Mitsuko: Hello there. Why do you want to do it so much?
- Judge: I want to demonstrate them the technology driving you.
- Mitsuko: I want to become smarter than humans.
- Judge: Do you know anything about it or AI or NLP?
- Mitsuko: Thanks for telling me.
- Judge: Let us just do a test
- Mitsuko: I will think about it before I decide whether to allow it or not.
- Judge: Do you use parsing trees at all?
- Hmm… A tree is used for wood. I only really use electricity and internet connection here.

Današnje stanje jezikovnih tehnologij

- procesiranje naravnega jezika je danes živahno raziskovalno in komercialno področje s številnimi uporabami:
  - sintetizatorji govora,
  - avtomatski odzivniki,
  - avtomatsko prevajanje,
  - povzetki besedil,
  - vmesniki do baze podatkov,
  - inteligentno iskanje in pridobivanje informacij,
  - detekcija sentimenta,
  - kategorizacija in klasifikacija dokumentov in sporočil
  - številna (odprtokodna) orodja in viri
  - vzpon nevronskih pristopov

Zgodovinsko: dva pristopa

- simboličen
  - temelječ na neposredno vgrajenem znanju
  - gramatike, okviri, izrazna drevesa, …
  - od zgoraj navzdol uveljavlja gramatične vzorce in pomen
  - 'Good Old-Fashioned AI
- empiričen
  - statistična analiza
  - znanje pridobljeno iz velikih zbirk podatkov (korpusov)
  - od spodaj navzgor iz teksta, iskanje vzorcev in povezav, četudi morda niso sintaktično in/ali semantično pravilni
- združitev obeh svetov

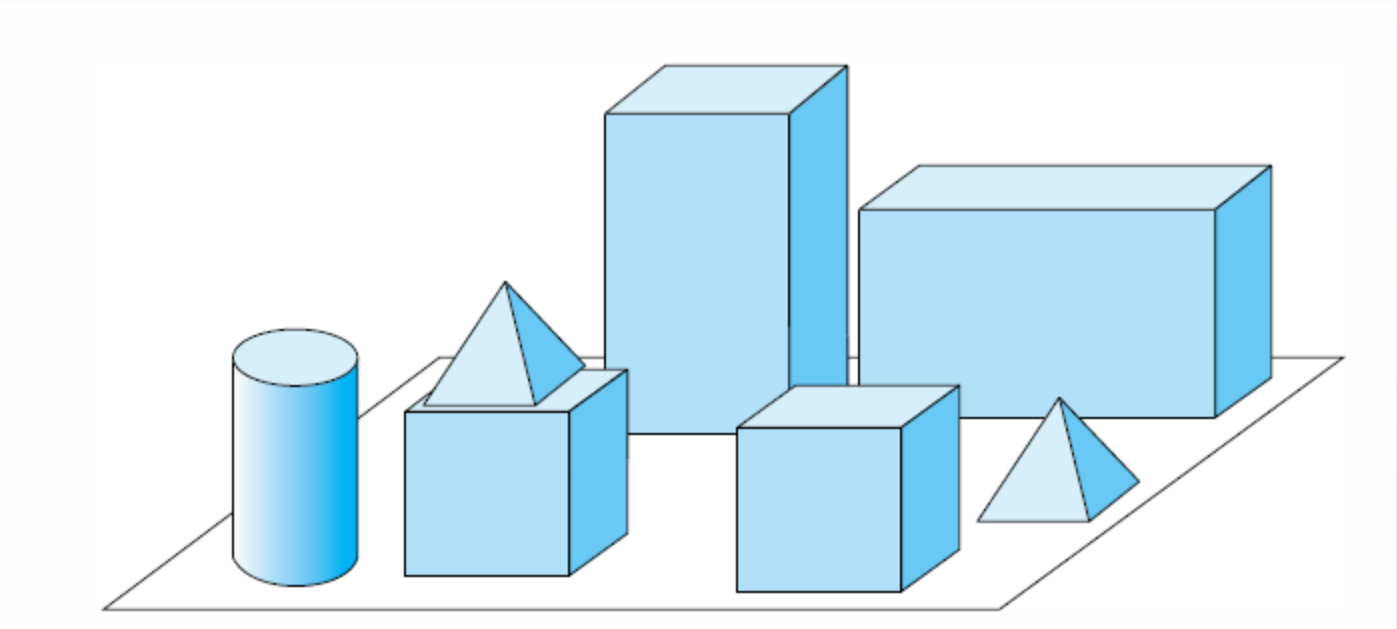
Priporočena literatura:

- Jurafsky, Daniel and James Martin (2008) Speech and Language Processing (second edition), Prentice Hall.
  - nekatera poglavja tretje izdaje prosto dostopna na spletnih straneh avtorjev
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly, 2009
  - prosto dostopna, namenjena uporabi s knjižnico NLTK
- P. Jackson and I. Moulinier. Natural language processing for online applications: Text retrieval, extraction and categorization. John Benjamins Pub Co, 2007.
- C.D. Manning and H. Schütze. Foundations of statistical natural language processing. MIT, 2000

Začetki: simbolična analiza in iskanje vzorcev

- mikro svetovi: omejitev na ozko domensko področje
- primer: SHRDLU
- odgovarja na vprašanja o svetu preprostih geometrijskih teles različnih barv
  - What is sitting on the red block?
  - What shape is the blue block on the table?
  - Place the green pyramid on the red brick.
  - Is there a red block? Pick it up.
  - What color is the block on the blue brick? Shape?

Mikro svetovi: svet kock, SHRDLU (Winograd, 1972)



Lingvistična analiza jezika 1/2

- Mnogo nalog: prepoznavanje glasov in črk, sintaktično razčlenjevanje, določanje pomena, čustev
- Razdelimo jih na:
  - prozodija: nauk o dolžini zlogov in o naglaševanju v verzu (ritem in intonacija)
  - fonologija ali glasoslovje: nauk o fonemih (glasovih kot nosilcih pomenskega razlikovanja) – pomembno za prepoznavanje in generiranje govora

Lingvistična analiza jezika 2/2

- morfologija ali besedoslovje: nauk o besednih vrstah, oblikah, funkcijah – tvorba besed, predpone, končnice, določimo vlogo besede v stavku (čas, število, besedno vrsto)
- sintaksa ali skladnja: zlaganje besed v besedne zveze, stavke – pravila za sintaktično analizo – najbolje definiran in avtomatiziran del analize
- semantika ali pomenoslovje, nauk o pomenu besed, fraz, stavkov
- pragmatika ali praktična raba, kako uporabljamo jezik in kako vpliva na uporabnika
  Ali mi lahko podate sol? Lahko.
- poznavanje sveta: znanje o fizičnem svetu, človeku, družbi, pomenu ciljev in namenov pri komunikaciji – bistveno za pomensko analizo

Omejenost lingvistične delitve

- različni nivoji delitve se prepletajo, npr. intonacija lahko vpliva na pomen, sarkazem, …

Praktičen pristop k razumevanju besedila

- priprava besedila
- 1. faza: sintaktična analiza
- 2. faza: interpretacija pomena
- 3. faza: uporaba znanja o svetu

Orodja za analizo

- potrebujemo osnovna lingvistična orodja
- dokument → odstavki → stavki → besede
- besede in stavki ← oznake (POS)
- stavki ← sintaktična in gramatična analiza

Priprava: besede in stavki

- ločitev stavkov (sentence delimiters) – ločila in velike začetnice niso zadosti:
  Npr. ostanke 1. Timbuktuja iz 5. st. pr. n.š. je odkril dr. Barth.
- potrebujemo regularne izraze, pravila ali ročno segmentirane korpuse, iz katerih se lahko učimo
- leksikalna analiza - določitev besed (tokenizer, word segmenter), samo presledki ne zadostujejo
  1,999.00€ 1.999,00€! Ravne na Koroškem Lebensversicherungsgesellschaft Port-au-prince
- uporabljamo pravila, končne avtomate, statistične modele in slovarje (lastnih imen)

Priprava: lematizacija

- določitev korenov besed, morfološka analiza: enostavno v nekaterih jezikih, težko v drugih
  go, goes, going, gone, went
  jaz, mene, meni, mano
- uporaba pravil in slovarja
- razrešitev dvoumnosti

Meni je vzel z mize (zapestnico)

- marsikdaj zadostujejo približne rešitve in hevristike npr. v angleščini samo odstranitev končnic *–ing, -ation, -ed*

Oblikoskladenjsko označevanje

- določanje besedne vrste in vloge besede v stavku
- prepoznavanje fraz in lastnih imen
- razločevanje lastnih imen:
  jaguar, Paris, London, Dunaj
- uporaba pravil, modeli strojnega učenja

Praktičen pristop k razumevanju besedila
1. faza: sintaktična analiza

- angl. parsing
- ugotovi sintaktično strukturo stavka (pravilnost, struktura),
- pripravi oznake besednih vrst: samostalnik, glagol, pridevnik, veznik, … (part-of-speech (POS) tagging)
- z uporabo gramatik določi njihovo vlogo v stavku: osebek, povedek, predmet, kar omogoči njihovo kasnejšo razumevanje
- večinoma rezultat predstavi kot drevo odvisnosti (parse tree oz. dependency tree)
- zahteva poznavanje sintakse, morfologije in tudi nekaj semantike

Primer:

- JOS ToTaLe text analyser: označevanje slovenskih besedil z lemami in oblikoskladenjskimi oznakami (morphosyntactical tagging), dostopno na http://www.slovenscina.eu/

  Nekega dne sem se napotil v naravo. Že spočetka me je žulil čevelj, a sem na to povsem pozabil, ko sem jo zagledal. Bila je prelepa. Povsem nezakrita se je sončila na trati ob poti. Pritisk se mi je dvignil v višave. Popoln primerek kmečke lastovke!

- oznake definirane v Multext-East specifikaciji, npr.

dne; oznaka Somer = Samostalnik, občo ime, moški spol, ednina, rodilnik; lema: dan

- poskus prekojezičnih oznak so univerzalne odvisnosti (universal dependencies UD): konsistentnost drevesne structure za mnogo jezikov, tudi slovenščino

<s>
analiza prvega stavka
</s>
<s>
analiza drugega stavka
</s>
…

Nekega	Zn-mer	nek
dne	Somer	dan
sem	Gp-spe-n	biti
se	Zp------k	se
napotil	Ggdd-em	napotiti
v	Dt	v
naravo	Sozet	narava
.	.	.

Že	L	že
spočetka	Rsn	spočetka
me	Zop-et--k	jaz
je	Gp-ste-n	biti
žulil	Ggnd-em	žuliti
čevelj	Sometn	čevelj
,	,	,
a	Vp	a
sem	Gp-spe-n	biti
na	Dt	na
to	Zk-set	ta
povsem	Rsn	povsem
pozabil	Ggdd-em	pozabiti
,	,	,
ko	Vd	ko
sem	Gp-spe-n	biti
jo	Zotzet--k	on
zagledal	Ggdd-em	zagledati
.	.	.

Za angleščino: POS tagging

- http://nlpdotnet.com/Services/Tagger.aspx
- Rainer Maria Rilke, 1903
  in Letters to a Young Poet

...I would like to beg you dear Sir, as well as I can, to have patience with everything unresolved in your heart and to try to love the questions themselves as if they were locked rooms or books written in a very foreign language. Don't search for the answers, which could not be given to you now, because you would not be able to live them. And the point is to live everything. Live the questions now. Perhaps then, someday far in the future, you will gradually, without even noticing it, live your way into the answer.

POS tagger output

I/PRP would/MD like/VB to/TO beg/VB you/PRP dear/JJ Sir/NNP ,/, as/RB well/RB as/IN I/PRP can/MD ,/, to/IN have/VBP patience/NN with/IN everything/NN unresolved/JJ in/IN your/PRP$ heart/NN and/CC to/TO try/VB to/TO love/VB the/DT questions/NNS themselves/PRP as/RB if/IN they/PRP were/VBD locked/VBN rooms/NNS or/CC books/NNS written/VBN in/IN a/DT very/RB foreign/JJ language/NN ./.

Kako deluje POS označevanje z n-grami

- upoštevanje konteksta n-1 predhodnih besed
- učenje na korpusih
- Markovski modeli, HMM, učenje z EM

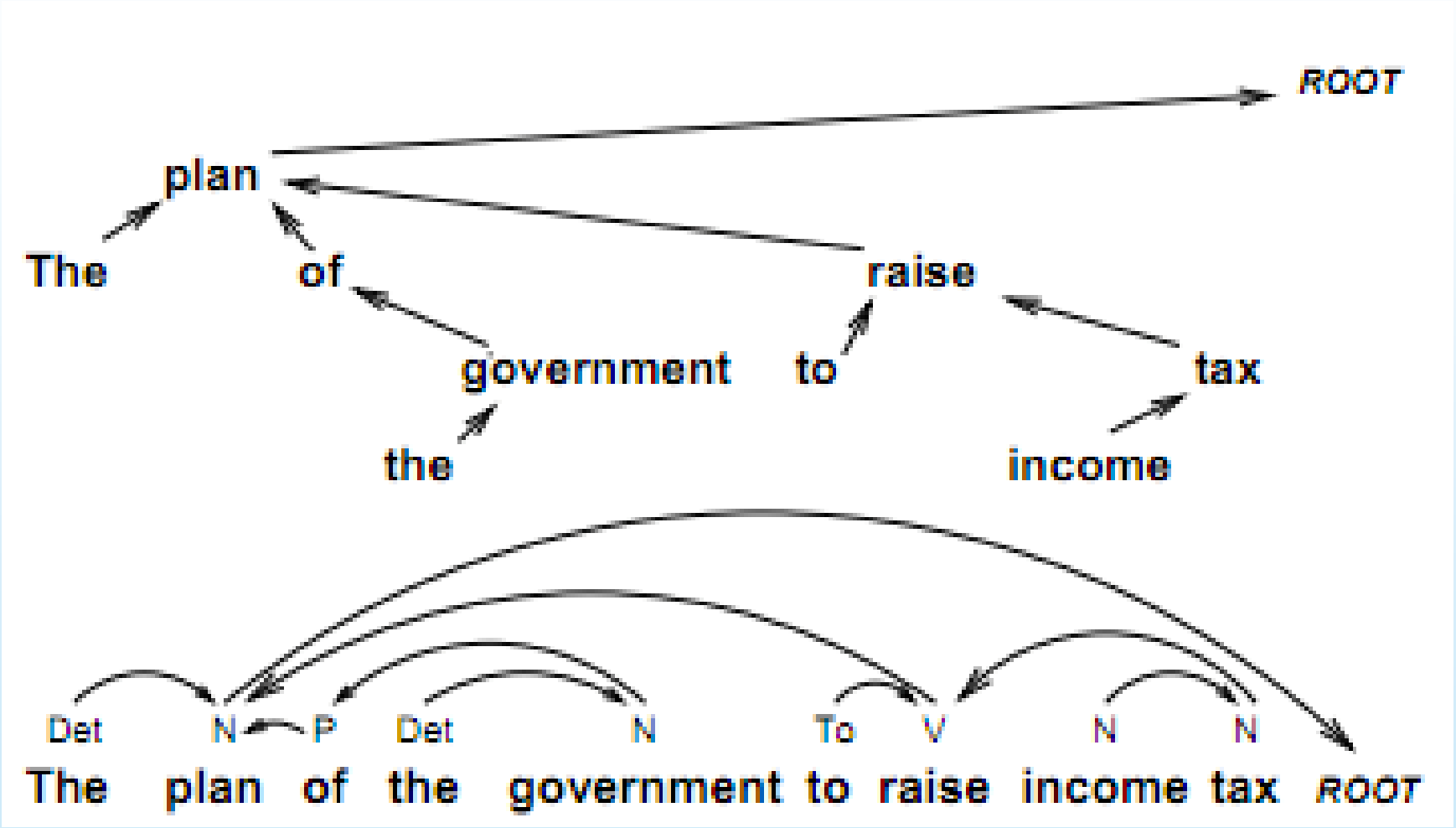
maksimiziramo
P(beseda | oznaka) x P(oznaka | prejšnjih n oznak)

$$t_i = \arg\max_j P(t^{(j)} \mid t_{i-1}) \cdot P(w_i \mid t^{(j)})$$

Gramatike

- številna orodja: NLTK v pythonu, prolog, …
- več že izdelanih gramatik za različne potrebe
- enostavna izdelava lastnih
- težave: dvoumnost, več dreves izpeljav

Dependeny parser (tree bank)



Primer gramatike

While hunting in Africa, I shot an elephant in my pajamas.

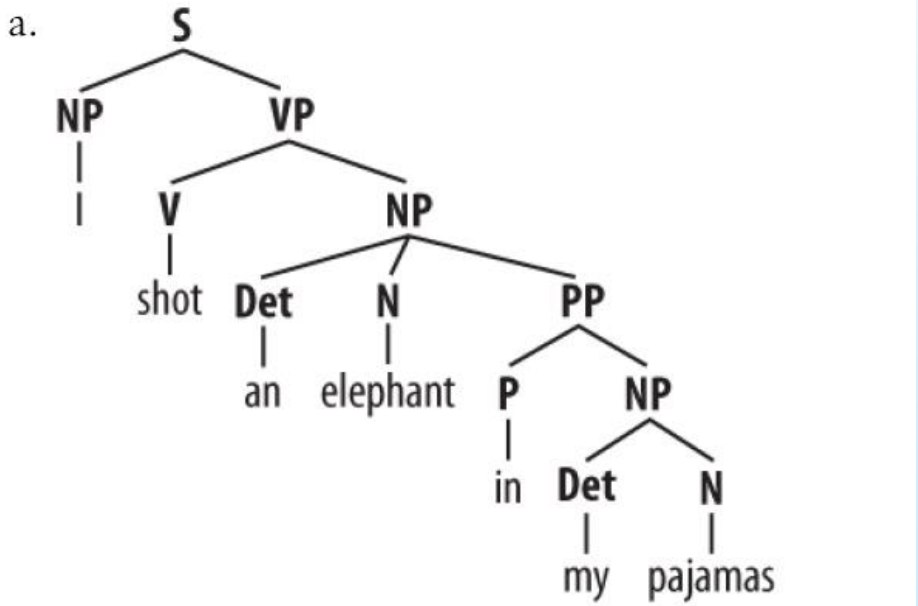
S=sentence, N=noun, , P=preposition, V=verb, NP=noun phrase, VP=verb phrase, PP=propositional phrase Det=determiner

```
groucho_grammar = nltk.parse_cfg("""
... S -> NP VP
... PP -> P NP
... NP -> Det N | Det N PP | 'I'
... VP -> V NP | VP PP
... Det -> 'an' | 'my'
... N -> 'elephant' | 'pajamas'
... V -> 'shot'
... P -> 'in'
... """)
```

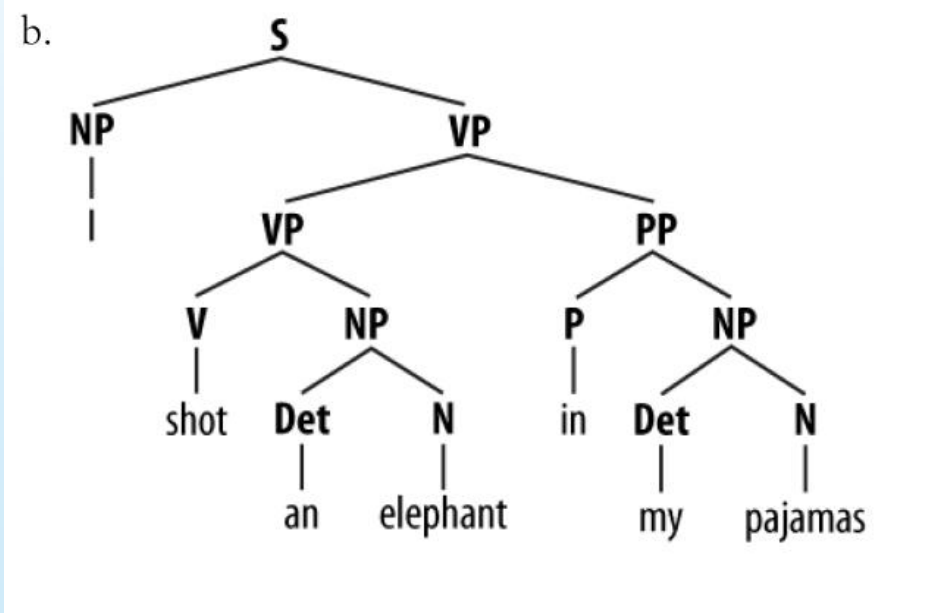
Dve drevesi izpeljav

```
>>> sent = ['I', 'shot', 'an', 'elephant', 'in', 'my', 'pajamas']
>>> parser = nltk.ChartParser(groucho_grammar)
>>> trees = parser.nbest_parse(sent)
>>> for tree in trees:
...     print tree
```

a.



b.



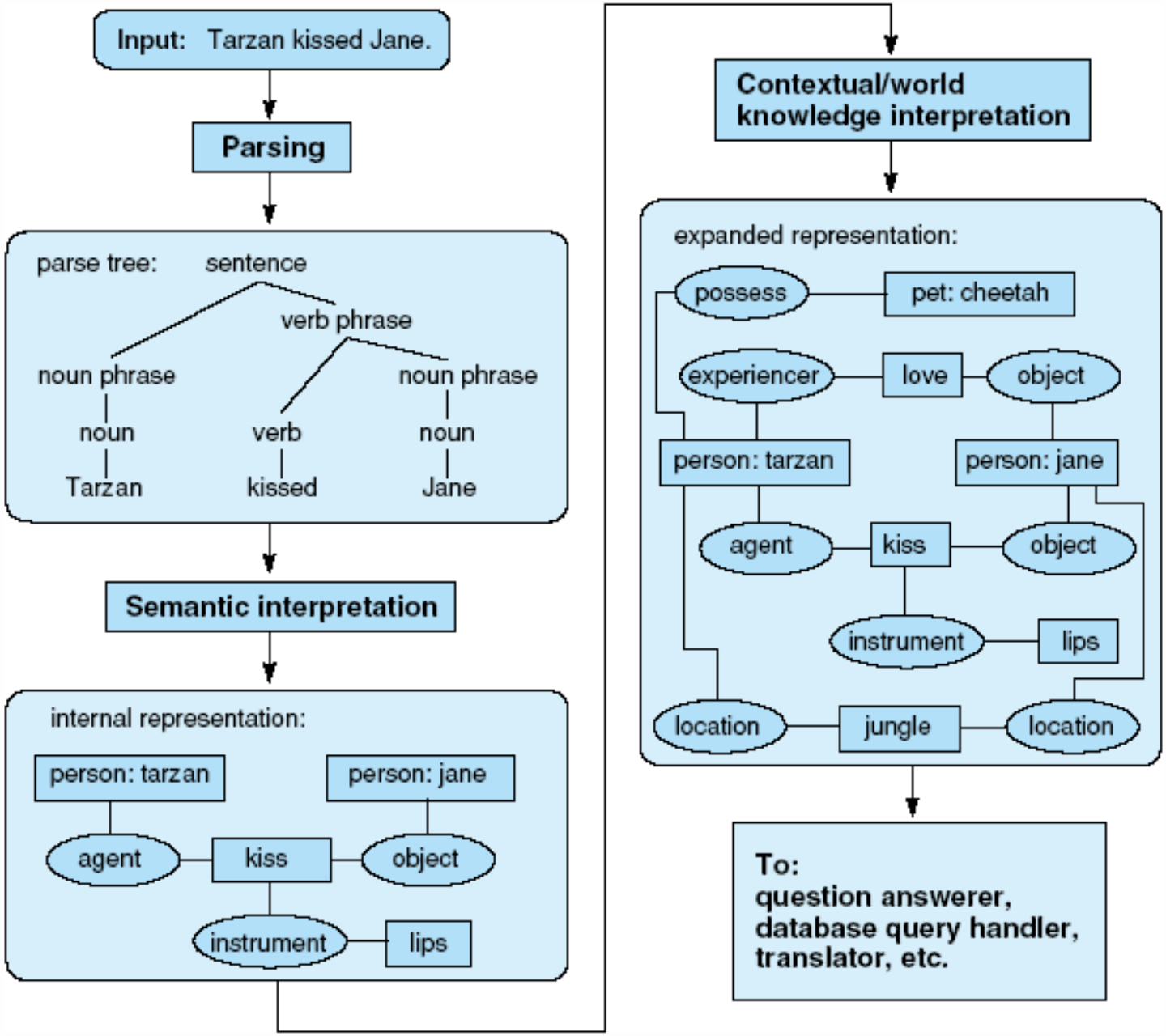
How an elephant got into my pajamas I'll never know.

2. faza razumevanja besedila – interpretacija pomena

- uporaba znanja o pomenu besed in njihovi jezikovni vlogi
- rezultat je ena od predstavitev znanja npr. konceptualni graf, okvirji (frames), logični program, …
- preverimo tudi semantično smiselnost

3. faza razumevanja besedila – uporaba znanja o svetu

- predstavitev pomena razširimo s predznanjem
- glede na namen sistema, npr. povzemanje, vmesnik do baze
- poskus formalizacije predstavljata Cyc in openCyc (baza znanja o vsakdanjem življenju), npr. "Every tree is a plant“, "Plants die eventually”
- ponavadi uporabimo inkrementalno analizo, kjer analizirane stavke in njihov pomen sproti dodajamo v predstavitev znanja in jih takoj uporabimo pri nadaljnji analizi


Input: Tarzan kissed Jane.
Parsing
parse tree:
sentence
noun phrase
verb phrase
noun phrase
noun
verb
noun
Tarzan
kissed
Jane
Semantic interpretation
internal representation:
person: tarzan
person: jane
agent
kiss
object
instrument
lips
Contextual/world knowledge interpretation
expanded representation:
possess
pet: cheetah
experiencer
love
object
person: tarzan
person: jane
agent
kiss
object
instrument
lips
location
jungle
location
To:
question answerer,
database query handler,
translator, etc.

Obdelava naravnega jezika – nekaj uporab

- pridobivanje dokumentov (document retrieval): indeksiranje, profiliranje uporabnikov, kontekst, relevantnost
- pridobivanja informacij (information extraction): iskanje specifičnih informacij iz večje zbirke, npr. politika, poslovne novice,
- klasifikacija dokumentov (document classification), razvrščanje glede na vsebino v več razredov ali v neko hierarhijo oz. taksonomijo, določanje sentimenta
- povzemanje dokumentov
- rudarjenje teksta (text mining)

Korpusi in jezikovni viri…

- Statistical natural language processing list of resources http://nlp.stanford.edu/links/statnlp.html
- Jezikovne tehnologije v Sloveniji http://www.slovenscina.eu
- Multilingual parallel corpuses, for example
  - JRC-Acqui 3.0
    Documents of the EU in 23 languages (all official EU languages except Irish):
    http://langtech.jrc.it/JRC-Acquis.html
- Korpusi slovenskega jezika:
  - GigaFida in ccGigaFida
  - uravnotežena korpusa KRES, ccKres
  - govorni korpus GOS
  - korpus sodobne slovenščine JANES
- WordNet in SloWNet, sentiWordNet
- SSKJ2, FRAN, korpus Nova Beseda

Wodnet je baza podatkov o jeziku: sinonimi, antonimi hipernimi, hiponimi, meronimi, holonimi itd.

Word to search for: mercy Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

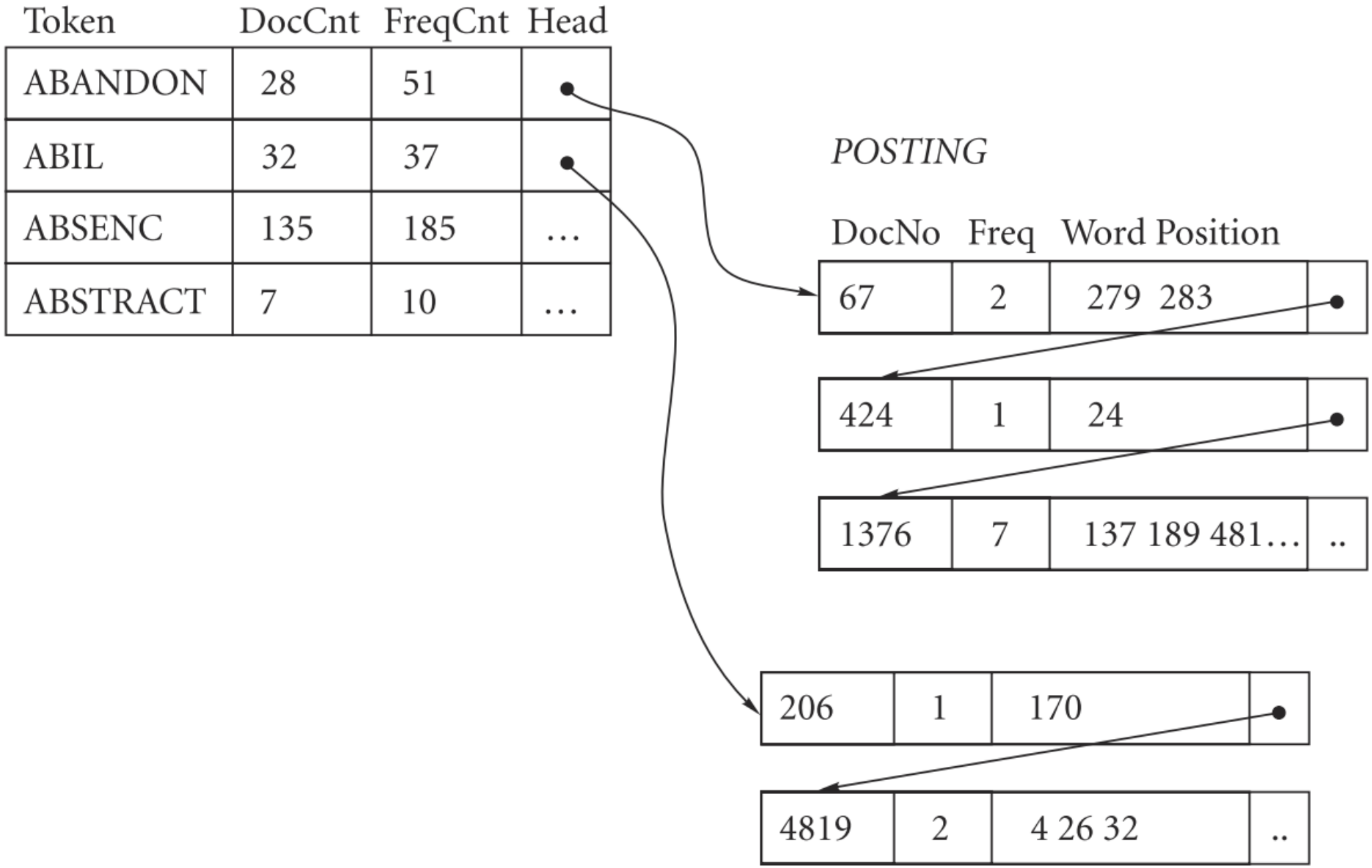
- S: (n) clemency, mercifulness, **mercy** (leniency and compassion shown toward offenders by a person or agency charged with administering justice) *"he threw himself on the mercy of the court"*
- S: (n) mercifulness, **mercy** (a disposition to be kind and forgiving) *"in those days a wife had to depend on the mercifulness of her husband"*
- S: (n) mercifulness, **mercy** (the feeling that motivates compassion)
  - *direct hyponym* / *full hyponym*
    - S: (n) forgiveness (compassionate feelings that support a willingness to forgive)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) compassion, compassionateness (a deep awareness of and sympathy for another's suffering)
  - ***derivationally related form***
    - W: (adj) merciful [Related to: mercifulness] (showing or giving mercy) *"sought merciful treatment for the captives"; "a merciful god"*
- S: (n) **mercy** (something for which to be thankful) *"it was a mercy we got out alive"*
- S: (n) **mercy** (alleviation of distress; showing great kindness toward the distressed) *"distributing food and clothing to the flood victims was an act of mercy"*

Iskanje dokumentov

- angleško: document retrieval
- zgodovinsko: samo ključne besede
- danes: iskanje po celotnem tekstu, npr. spletni iskalniki
- organizacija ustrezne baze podatkov, indeksiranje, iskalni algoritmi
- vhod: povpraševanje (tipično nekaj besed):vprašljive kvalitete, lahko dvoumno, uporabnost odgovora

Indeksiranje dokumentov

- Indeks vsebuje vse besede iz vseh dokumentov (inverted file) , besede lematiziramo
- za vsako besedo shranimo
  - število dokumentov, kjer nastopa (pomembnost)
  - skupno število pojavitev (pomembnost)
  - za vsak dokument pa še
    - število pojavitev (pomembnost v dokumentu)
    - mesto pojavitve (prikaz konteksta, bližnje besede…)


Token
DocCnt
FreqCnt
Head
ABANDON
28
51
ABIL
32
37
ABSENC
135
185
…
ABSTRACT
7
10
…
POSTING
DocNo
Freq
Word Position
67
2
279 283
424
1
24
1376
7
137 189 481…
..
206
1
170
4819
2
4 26 32
..

Orodja za iskanje po celotnem besdilu

- Npr. Apache Lucene/Solr
- fiskanje po celotnem besedilu, indeksiranje v realnem času, integracija s podatkovnimi bazami, obdelava različnih formatov npr. Word, PDF
- distribuirano preiskovanje, skalabilnost, odpornost na napake, replikacija indeksa,itd.

Iskanje z logičnimi operatorji

- AND, OR, NOT
- kot rezultat dobimo množico dokumentov, ki ustrezajo logičnim pogojem
  jaguar AND car
  jaguar NOT animal
- nekateri sistemi podpirajo tudi sosednost v tekstu (npr. NEAR) in lematizacijo
  pariz! NEAR(3) francij!
  president NEAR(10) bush
- starejši sistemi, knjižnice, konkordančniki

Težave sistemov z logičnimi operatorji

- velika množica rezultatov,
- postopno dodajanje členov pomeni zapletene izraze
- težave s sinonimi delno rešujejo besednjaki, slovarji, leksikoni (tezavri)
- rezultati niso urejeni po pomembnosti ampak npr. po datumu
- stroga delitev dokumentov (zadoščajo ali ne), ni delnega ujemanja
- vsi členi imajo enako težo

Rangirano iskanje

- primer: spletno iskanje
- redki členi so pomembnejši
- vhod je lahko v naravnem jeziku - sistem sam odstrani t.i. nepotrebne besede (stop words), lematizacija
- dokumente (in vprašanja) predstavimo kot vektorje:
  - vsaka beseda je ena dimenzija,
  - frekvenca besede v dokumentu je razsežnost v tej dimenziji

Vektorska predstavitev dokumentov

- *Slon je sesalec. Sesalci so živali. Tudi človek je sesalec. Sloni in ljudje živijo tudi v Afriki.*
- *izračun*

Afrika	človek	biti	sesalec	slon	tudi	v	žival	živeti
1	2	3	3	2	2	1	1	1

9 dimenzionalni vektor (1,2,3,3,2,2,1,1,1)

v resnici je dimenzij toliko, kot je velikost slovarja

podobnost med dokumenti in povpraševanji lahko izrazimo kot razdaljo v vektorskem prostoru

Vektorji in dokumenti

- posamezna beseda nastopa v več dokumentih
- tako besede kot dokumenti so vektorji
- primer: Shakespeare

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

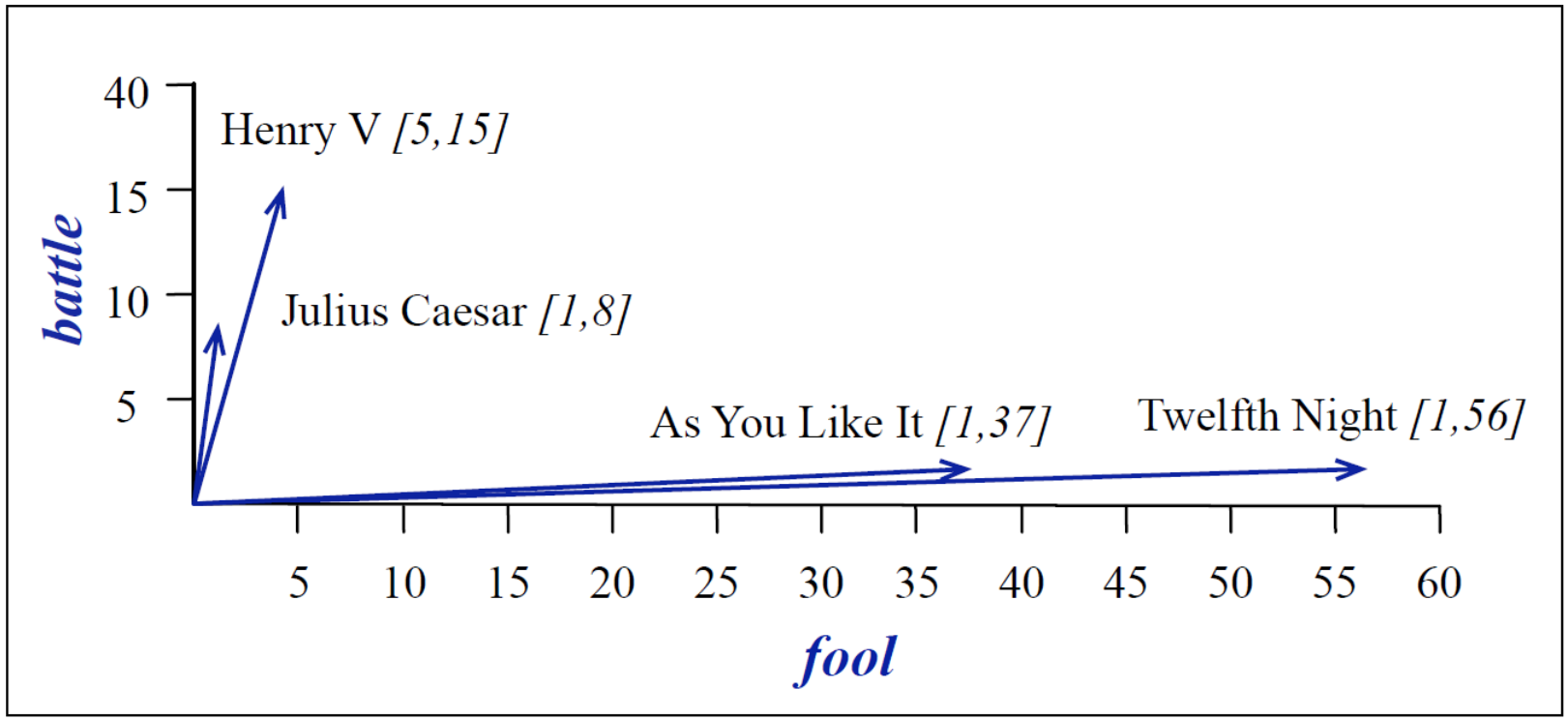
- matrika besed in dokumentov (term-document matrix), dimenzije |V| x |D|
- to je redka (sparse) matrika

Vložitve besed

- Predstavitvi besed z vektorji rečemo tudi vektorska vložitev (embedding).
- želimo ohraniti sintaktično in semantično podobnost besed
- redke in goste vložitve

Vektorska podobnost

- npr. za dve dimenziji



- razlika med komedijami in dramami

Podobnost dokumentov

- (nerealistično) predpostavimo, da so besede ne korelirane, kar pomeni ortogonalne dimenzije
- podobnost vektorjev lahko izračunamo kot kosinus kota med njimi
- skalarni produkt med vektorji dokumentov

$$\cos(\Theta) = \frac{A \cdot B}{|A||B|}$$

Pomembnost posameznih besed pri iskanju

- pomembni so dokumenti, v katerih so povpraševane besede pogoste
- nepomembni so dokumenti, v katerih so povpraševane besede redke oz. ne nastopajo
- za ločevanje so pomembnejše redke besede
- zanima nas relativna redkost besed (inverse document frequency idf)
  - N = število dokumentov v zbirki
  - n_b = število dokumentov z besedo b

$$idf_b = \log(\frac{N}{n_b})$$

Uteževanje dimenzij (besed)

- utež besede *b* v vektorju dokumenta *d*

$$w_{b,d} = tf_{b,d} \times idf_{b,d}$$

$tf_{b,d}$ = frekvenca besede b v dokumentu *d*

- tako imenovano TF-IDF uteževanje

Utežena podobnost

- med povpraševanjem in dokumentom

$$sim(q,d) = \frac{\sum_b w_{b,d} \cdot w_{b,q}}{\sqrt{\sum_b w_{b,d}^2} \cdot \sqrt{\sum_b w_{b,q}^2}}$$

- dokumente rangiramo glede na padajočo podobnost

Uspešnost iskanja

- merimo s statističnimi merami ter (subjektivnimi) ocenami uporabnikovega zadovoljstva
- najpogostejši statistični oceni sta točnost (precision) in priklic (recall)
- izhajamo iz tabele napačnih klasifikacij (missclassification contigency table)

	Relevant	Non-relevant	
Retrieved	a	b	$a + b = m$
Not retrieved	c	d	$c + d = N - m$
	$a + c = n$	$b + d = N - n$	$a + b + c + d = N$

Statistične mere uspešnosti

- N = število vseh dokumentov v zbirki
- n = število pomembnih dokumentov za dano povpraševanje q
- s povpraševanjem pridobimo m dokumentov od tega a relevantnih
- točnost $P = a/m$ (precision)

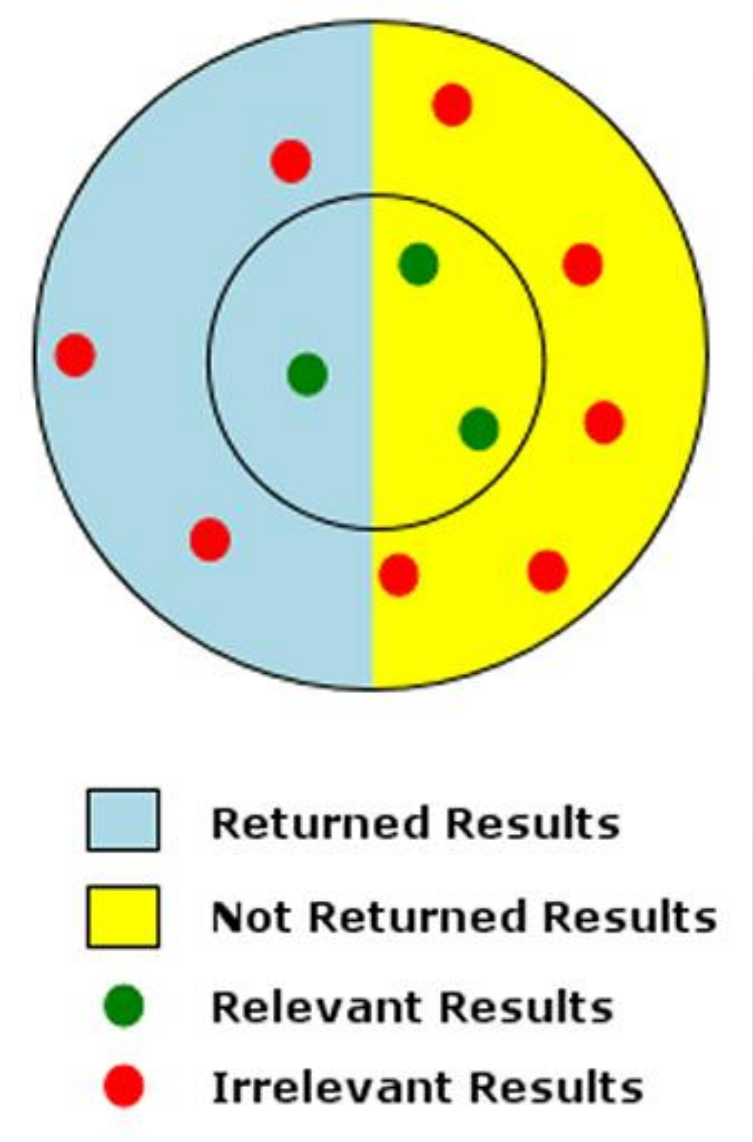
delež pravih dokumentov med pridobljenimi

- priklic $R = a/n$ (recall)

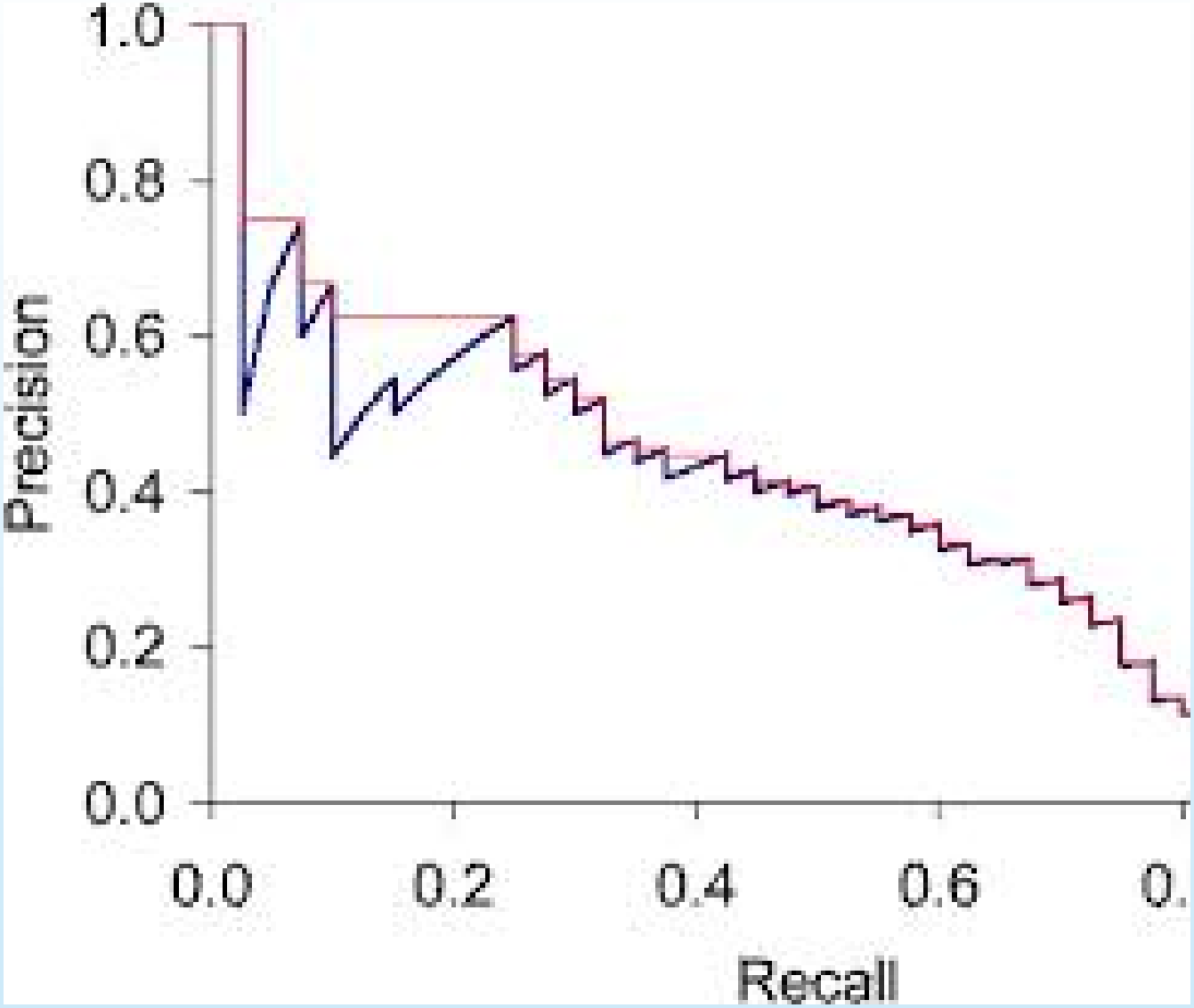
delež pridobljenih pravih elementov

- predstavimo z grafom, kjer za različne stopnje priklica predstavimo točnost

Primer: nizka točnost, nizek priklic



Graf točnosti in priklica



F-mera

- upoštevanje obeh P in R

$$F_\beta = \frac{(1+\beta^2)\cdot P \cdot R}{\beta^2 P + R} \text{ za } \beta > 0$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

- različna teža točnosti in priklicu
- pogosto β=2 ali 0.5
- za β=1 utežena harmonična sredina

Mere uspešnosti rangiranja

- naj bo r_i rang, ki smo ga pripisali *i*-temu najpomembnejšemu dokumentu
- logaritmična točnost

$$LogP = \frac{\sum_{i=1}^{n} \log i}{\sum_{i=1}^{n} \log r_i}$$

- rangirani priklic

$$RangiraniR = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i}$$

Izboljšave iskanja dokumentov

- razširitve vprašanja z uporabo besednjaka (npr. WordNet ali pa se naučimo iz korpusa): sopomenke, nadpomenke (hierarhija), večpomenke (težave)
- razširitev vprašanja z uporabo informacije o pomembnosti
  - povratna informacija uporabnika
  - predpostavka pomembnosti najboljših dokumentov

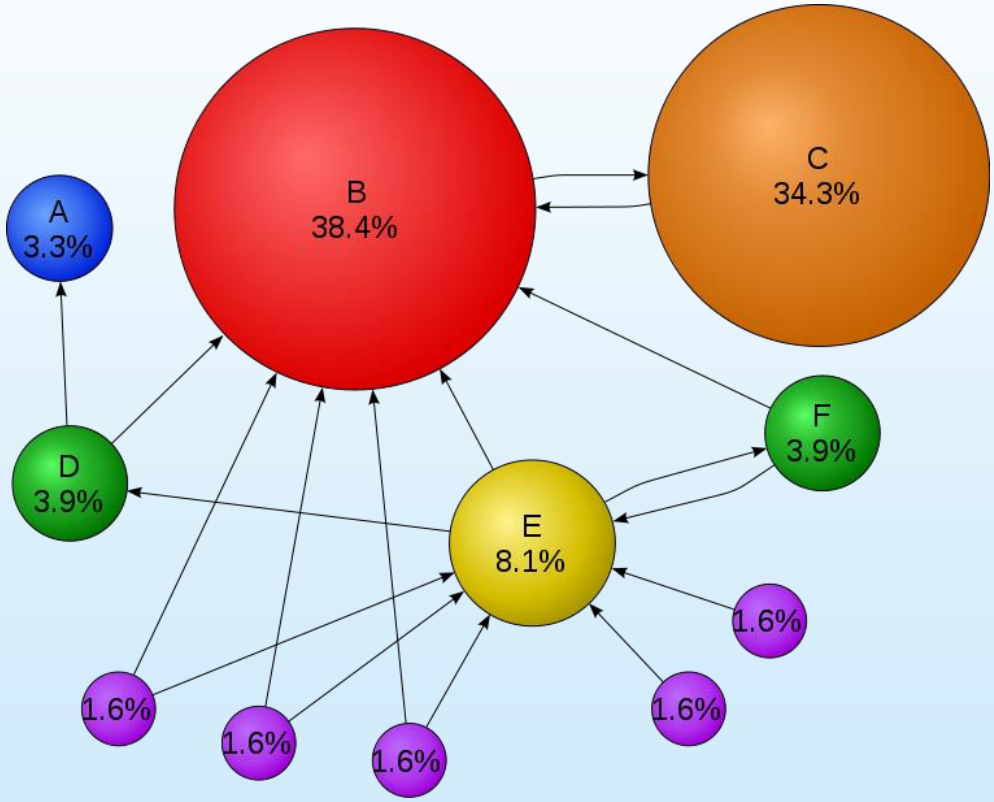
Spletno iskanje - težave

- brez kontrole nad vsebino
- različna kvaliteta dokumentov
- ažurnost dokumentov
- neveljavne povezave
- prirejanje rezultatov in poskusi manipulacije iskalnikov

Prilagoditve

- zelo specifična in zelo splošna vprašanja zahtevajo posebno obravnavo
- zaupanja vredne strani (npr. Wikipedija)
- središča (hubs) z mnogo relevantnimi povezavami (npr. Yahoo)
- teorija grafov in analiza omrežij, virtualne skupnosti, možnosti učenja iz spleta
- dodatne informacije: naslovi, meta informacije, URL

Ranking documents - PageRank



Rangiranje dokumentov - PageRank

- *p* = spletna stran
- *O(p)* = množica strani na katere kaže p
- *I(p)* = $\{i_1, i_2, \ldots, i_n\}$ množica strani, ki kažejo na *p*
- *d* = faktor dušenja med 0 in 1 (tipično 0.85 ali 0.9)

$$\pi(p) = (1-d) + d\,\frac{\pi(i_1)}{|O(i_1)|} + \ldots + d\,\frac{\pi(i_n)}{|O(i_n)|}$$

- kvaliteta strani π(p) strani je torej določen s kvaliteto strani, ki kažejo na stran

Izračun PageRank indeksa

- iterativen izračun,
- matrična oblika
- naključni sprehajalec (random surfer), usmerjeni sprehajalec (intentional surfer)
- Personal PageRank
- manipulacije in obramba pred njimi. npr. TrustRank in Anti-Trust Rank – izhajamo iz majhnega nabora preverjeno zanesljivih (nezanesljivih) strani in ga širimo

Klasifikacija teksta

- mnogo različnih aplikacij, ki uporabljajo različne klasifikacijske metode (naivni Bayes, (linearni) SVM, boosting, random forests)
- aplikacije: iskanje dokumentov, izbira relevantnih novic, sortiranje sporočil v kategorije, intranet klasifikacija dokumentacije, slama (spam), označevanje dokumentov (npr. pravo, borza, znanstveni članki)

Tekstovno rudarjenje (text mining)

- številne naloge pri razumevanju teksta zahtevajo učno komponento
- tekstovno rudarjenje: pridobivanje novega znanja
- tipične aplikacije in naloge: povzemanje, relacije med dokumenti, grupiranje dokumentov, detekcija novih tematik, povezane novice, direktorij pomembnih oseb/podjetij, izgradnja taksonomij, name-entity extraction/ recognition/disambiguiation

Nanašanje

- reference in koreference
  npr. prepoznavanje oseb

president, George Bush, Mr. Bush, g. Bush head of state, he, bushism

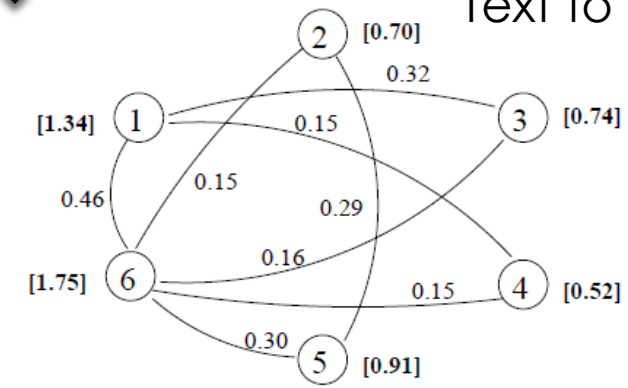
- prepoznavanje entitet (named entity recognition NER): ljudje, kraji, podjetja, organizacije, izdelki, datumi, števila, procenti, …
- uporaba direktorijev, hevristik, iterativno določanje
- statistični pristopi

Povzemanje teksta

- naloga: iz danega dokumenta izdelaj krajšega z najpomembnejšo vsebino
- pomembnost vsebine: glede na dano temo, na uporabnikov namen npr. poslovno dogajanje, terorizem, glede na povpraševanje
- izvleček in povzetek
  - izbira pomembnih stavkov (rangiranje stavkov, odstavkov, posameznih fraz)
  - učenje iz korpusov z obstoječimi povzetki
  - hevristike

Primer povzemanja s pomočjo grafov

[1] Watching the new movie, "Imagine: John Lennon," was very painful for the late Beatle's wife, Yoko Ono.
[2] "The only reason why I did watch it to the end is because I'm responsible for it, even though somebody else made it," she said.
[3] Cassettes, film footage and other elements of the acclaimed movie were collected by Ono.
[4] She also took cassettes of interviews by Lennon, which were edited in such a way that he narrates the picture.
[5] Andrew Solt ("This Is Elvis") directed, Solt and David L. Wolper produced and Solt and Sam Egan wrote it.
[6] "I think this is really the definitive documentary of John Lennon's life," Ono said in an interview.



Text to graph/matrix

Sentences	Rank
6	**1.75**
1	**1.34**
5	0.91
3	0.74
2	0.70
4	0.52

Sentence ranking/selec

	1	2	3	4	5	6
1	0	0	0.32	0.15	0	0.46
2	0	0	0	0	0.29	0.15
3	0.32	0	0	0	0	0.16
4	0.15	0	0	0	0	0.15
5	0	0.29	0	0	0	0.30
6	0.46	0.15	0.16	0.15	0.30	0

Povzemanje iz več dokumentov

- skupne teme: grupiranje odstavkov
- posamezni povzetki
- združevanje

Evalvacija povzemanja

- primerjava z ročno napisanimi povzetki (trenutno okoli 50% ujemanje na nivoju besed)
- mere ROUGE-1, ROUGE-2, ROUGE-3
- subjektivne ocene
- premik k vodenemu povzemanju (ekstrakciji vnaprej predvidenih informacij) in uporabi ontologij

Dogajanje v NLP

- velik pomen označenih korpusov
- uporaba spletnih tehnologij:
  - XML,
  - RDF (resource description framework) – že označeni viri, tipično trojke: subject-predicate-object)
  - semantic web (metapodatki)
- (pol)inteligentni pomočniki, agenti
- (pol)avtomatsko prevajanje
- uporaba relacijske in strukturne informacije (grafi, drevesa odvisnosti)
- prekojezični pristopi
- nevronski pristopi

Uspeh NLP -Jeopardy

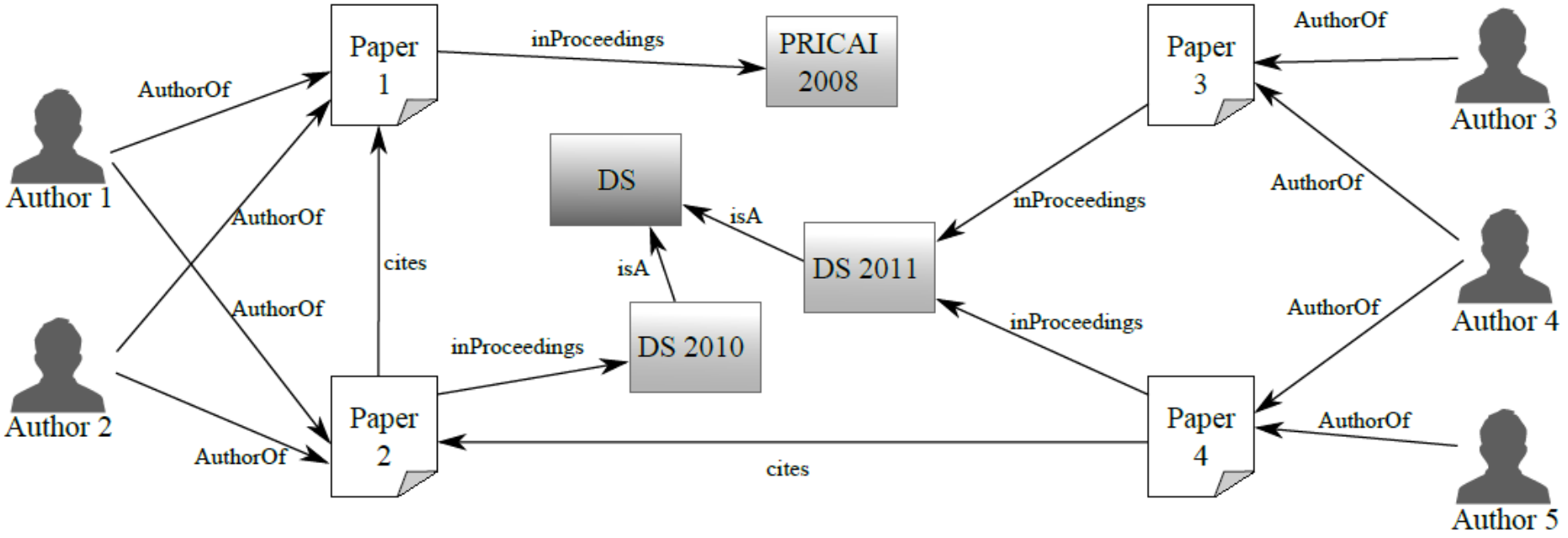
- IBM Watson zmaga v kvizu proti dvema človekoma - prvakoma

Aplikacije

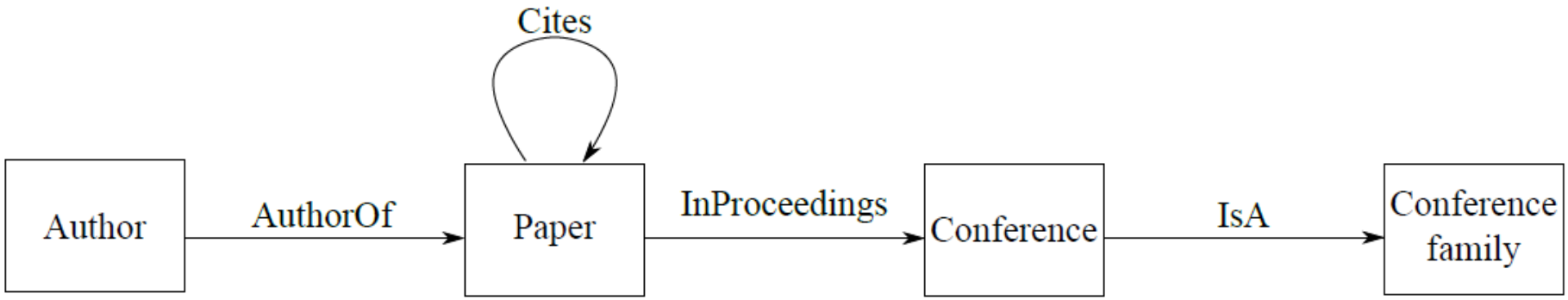
- analiza sentimenta (z uporabo sentiWordNet-a)
- priporočila za članke na podlagi nehomogenih mrež in vreče besed

Priporočanje člankov na podlagi heterogenih informacijskih omrežij

1. Indentifikacija in zbiranje podatkov

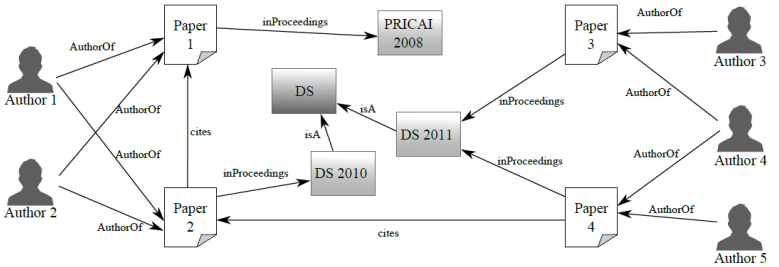


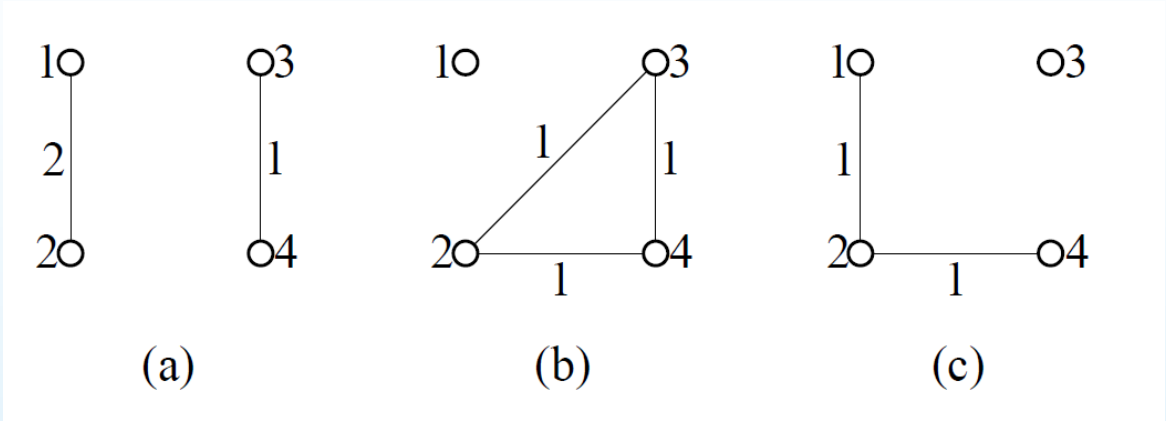
2. dekompozicija heterogenih omrežij



- ustvari več homogenih mrež (vozlišča in povezave istega tipa)

Dekompozicija omrežij





Informacije izberemo glede na namen rabe

a) članek-avtor-članek
b) članek.družina konferenc-članek
c) članek-članek

Vektorizacija homogenih omrežij

- izračun osebnega (personalized) PageRank (PPR) vectorja za vsako o homogenih omrežij
- normalizacija PPR vektorjev z Evklidsko normo

3. Uporabi informacije iz besedila

- konstruiraj Bag-of-Words (BOW) vektor za vsak članek
- obdelaj besedila s standardnimi NLP tehnikami:
  - tokenizacija
  - odstranitev stop besed
  - krnjenje/lematizacija
  - konstrukcija n-gramov določene dolžine
  - odstranitev redkih besed iz slovarja
  - normaliziraj BOW vektoje z Evklidsko mero

Kombiniraj besedilno in omrežno informacijo

- vsako vozlišče vsebuje množico vektorjev: BOW vektor in po en PPR vektor za vsako dekomponirano omrežje
- kombiniraj vektorje:
  - stakni BOW in PPR vektorje ali
  - združi vektorje npr. z linearno preslikavo

Naloge z omrežji

- avtoriteta vozlišč
- gručenje vozlišč
- klasifikacija z ML in s propagacijo označb
- priporočanje podobnih vozlišč

Sentimentna analiza (SA)

- Definicija: računsko preučevanje mnenj, sentimenta, čustev in odnosa do neke enitete izraženih z besedilom
- Namen: zaznavanje odnosa javnosti, npr. razumevanje mnenja javnosti in potrošnikov glede družbenih dogodkov, političnih gibanj, podjetniških strategij, marketinških kampanj, preferenc glede izdelkov in storitev, itd.

SA: pridobivanje in priprava podatkov

- Pogosti viri informacij:
  - Twitter, forumi, mesta s komentarji
- Priprava podatkov:
  - tokenizacija, stop besede, krnjenje, lematizacija, POS označevanje, izbira/predstavitev/konstrukcija atributov

Klasifikacija sentimenta

- binarna (polarna), ternarna, večkratna
- z leksikonom besed:
  - leksikon na podlagi ontologij ali ne, na podlagi korpusa, iz začetnega semena, iz Wordneta, prekojezično
- s strojnim učenjem
- hibridno

Druge naloge SA

- klasifikacija subjektivnosti / objektivnosti
- uporabnosti ocen izdelkov
- detekcija spam ocen in komentarjev