Formalni jezici i jezični procesori I
2020./2021.

Prof. dr. sc. Sanda Martinčić - Ipšić

smarti@inf.uniri.hr

Merlin pswd

- FJJP

- Predavanja počinju u 10:30

Epidemiološke upute

1. Prije polaska kod kuće **mjerite temperaturu**. U slučaju temperature ili bilo kojeg simptoma ostajete kući.
   - nastava je ionako uvijek i online
2. Ulazite kroz glavna vrata u predavaonicu **s maskama.**
3. Uvijek (na svim predavanjima) **sjedite na istom mjestu.**
4. Maske tijekom predavanja možete skinuti, ali čim se dižete s mjesta masku je potrebno staviti
5. Nosite **sredstvo za dezinfekciju** sa sobom, klupe se čiste ali neće biti vremena da se čiste između predavanja

MOLBA: Pridržavajte se svih uputa da čim dulje izdržimo u predavaonici, kad uvjeti ne budu zadovoljeni prelazimo na online

Epidemiološke upute II

- Potpisuje se **epidemiološka evidencija** prisutnih studenata
  - Papir se potpisuje prilikom ulaza u predavaonicu, kod potpisivanja održavati razmak od 1.5 m
  - papir neće kružiti

Dozvoljena mjesta

- Sjedi se isključivo na mjestima, koja su označena s naljepnicama
- **Obavezno** je **pravilno** nošenje maski:
  - **U svim učionicama**, kao i u svim zatvorenim prostorima
  - **U cijeloj zgradi**
  - **Tijekom predavanja** i vježbi

Stream link za predavanja

- https://www.twitch.tv/oiri028

- Operativne stvari pitati u chat-u
  - (čujem, ne čujem, vidim, ne vidim)
- pitanja vezana uz sadržaj postaviti na forumu u Merlin-u
  - svi studenti mogu vidjeti pitanje i čak i odgovoriti
  - Za takvu aktivnost vas mogu nagraditi dodatnim bodovima

ILI

- Teorijske osnove računarstva
- Teorijski jezici

- Theory of Computation
- Automata Theory
- Theoretical Languages

Raspored

- **predavanja srijeda 028 10:30-12:00**
- **vježbe srijeda 028 12:00-13:30**

- konzultacije – po prethodnoj najavi e-mailom (srijedom nakon predavanja) smarti@uniri.hr

- asistent: Karlo Babić, : karlo.babic@uniri.hr

Polaganje ispita jednopredmetni, matematika, fizika

- Prisutnost na nastavi: 80%
- Domaće zadaće: 4x5=**20** bodova
- 2 kolokvija: **50** bodova
- ili seminar 25 bodova
- Završni ispit: **30** bodova (minimalno 50%)
- rokovi za predaju DZ **nisu** promjenjivi odnosno ukoliko se ne preda u roku smatra se da zadaća/seminar/rad nije predan

Polaganje ispita II

- UMJESTO 2. KOLOKVIJA:
- **(rok za dogovor je do 16.12.)**

  OPCIJA: seminar - PROGRAMERSKI!!!

  - moguće prikupljanje dodatnih bodova
  - nastavak rada na seminaru u sljedećem semestru te za Završni/Diplomski rad
  - TEMA: Rad na prikupljanju podataka, pasiranjima formata, metode strojnog učenje, parseri, kompleksne mreže, teorija grafova,... ...
  - ili po dogovoru ....

Polaganje ispita -dvopredmetni

- Prisutnost na nastavi: 50%
- 2 Kolokvija: **50** bodova
- Završni ispit: **30** bodova (minimalno 50%)
- potrebno prikupiti 30 za popravni i 40 Završni
- **DODATNO NEMA Domaćih Zadaća!!!!**
- **ALI možete ih predati za dodatne bodove**
  - u tom slučaju kad prijeđete prag za završni ukupni broj bodova se zbraja i normira na 80 (dijeli s 80)
  - znači maksimalno je moguće postići 100/80
  - ocjenska ljestvica ostaje ista

Završna ocjena

- A – 90% - 100% (ekvivalent: izvrstan 5)
- B – 75% - 89,9% (ekvivalent: vrlo dobar 4)
- C – 60% - 74,9% (ekvivalent: dobar 3)
- D – 50% - 59,9% (ekvivalent: dovoljan 2)
- F – 0% - 49,9% (ekvivalent: nedovoljan 1)

Literatura

- S. Srbljić. Jezični procesori I, Element, Zagreb, 2002. ILI
- Srbljić Siniša: Uvod u teoriju računarstva. Element, Zagreb, 2007.

- M. Sipser, Introduction to the Theory of Computation, Brooks Cole, 1st edition, 1996.
- J. E. Hopcroft, J. D. Ullman. Introduction to Automata Theory, Languages and Computation, Addison-Wesley, 1979. (3rd edition 2006)

Izvedbeni plan I

DETALJNO U DINPU

21.10.2020. Rok za predaju 1. domaće zadaće

18.11.2020. Rok za predaju 2. domaće zadaće

25.11.2020. **1. Kolokvij**

16.12.2020. Rok za predaju 3. domaće zadaće

Božićni blagdani

20.01.2021. **2. Kolokvij**

20.01.2021. Rok za predaju 4. domaće zadaće

Ispitni rokovi (U ISVU)

- 11.02.2021. u 10
- 25.02.2021. u 10
- 11.03.2021. u 10
- 09.09.2021. u 10

Izvedbeni plan III

- **ROKOVI u DINPU:**
- uvijek provjeriti stanje u ISVU

- završni ispit: **4 roka**

- Prolaznost 78% (prošla ak. god.)

Sadržaj

- Uvod
  - rad jezičnog procesora
- Automati:
  - DKA, NKA, ε - NKA, Mooreov i Mealyev automat
- Regularni izrazi
- Gramatike
- Parseri
- Potisni automat
- Turingov stroj
- Chomskyjeva hierarhija jezika

Praktični sadržaji

- grep, regularni izrazi
- perl
- parseri

- seminari
- više prakse na FJJP2

Primjena formalnih metoda

- izgradnja jezičnih procesora (compiler)
- izgradnja inteligentnih sustava
- protokoli za komunikaciju
- analiza prirodnog jezika (NLP)
- analiza prirodnog govora
- računalna lingvistika
- strojno prevođenje
- ....

Formalni jezici i jezični procesori II

- Ljetni semestar
  - Izborni kolegij
  - Seminarski rad
  - Priprema za pisanje završnog rada
  - Završni rad
- Sadržaj
  - Leksička analiza
  - Sintaksna analiza
  - Semantička analiza
  - Jezični procesori / kompajleri

Računalna analiza prirodnog jezika (NLP)

- područje umjetne inteligencije (AI)
  - Turingov test
- upotrebljava
  - Konačni automati, Formalne gramatike
  - Chomskyeva hijerarhija jezika
  - Parseri
  - Statističke metode
- srodna područja
  - Obrada govora (Speech Technologies): raspoznavanje, sinteza, dijalog
  - Računalna lingvistika (Computational Linguistics)

Računalna lingvistika

- Jezične tehnologije (Language Technologies):
  - jezični resursi (korpusi, rječnici, leksikoni)
  - jezični alati (provjera gramatike i pravopisa)
  - obilježivači vrsta riječi (Part Of Speech Tagging)
  - morfološki analizatori (konačni automati)
  - prepoznavanje sintakse (parseri, gramatike)
  - prepoznavanje semantike (leksičkoga i rečeničnoga značenja)
  - prepoznavanje pragmatike (diskusrsa i dijaloga)
  - strojno prevođenje (strojno potpomognuto prevođenje)

Inteligentna računala mogu:

- rješavati teške probleme
- pomoći pri istraživanjima i konstruiranjima
- pomoći u proizvodnji
- razumjeti prirodan jezik
- razumjeti slike
- naučiti određene primjere i uzorke
- ....

Natural Language Processing-NLP
Računalna analiza prirodnog jezika

- Računala na mogu razumjeti prirodni jezik poput male djece? Zašto je to tako teško?
  - nepravilnosti u jezicima
  - razumijevanje
  - dvosmiselnost
  - ....

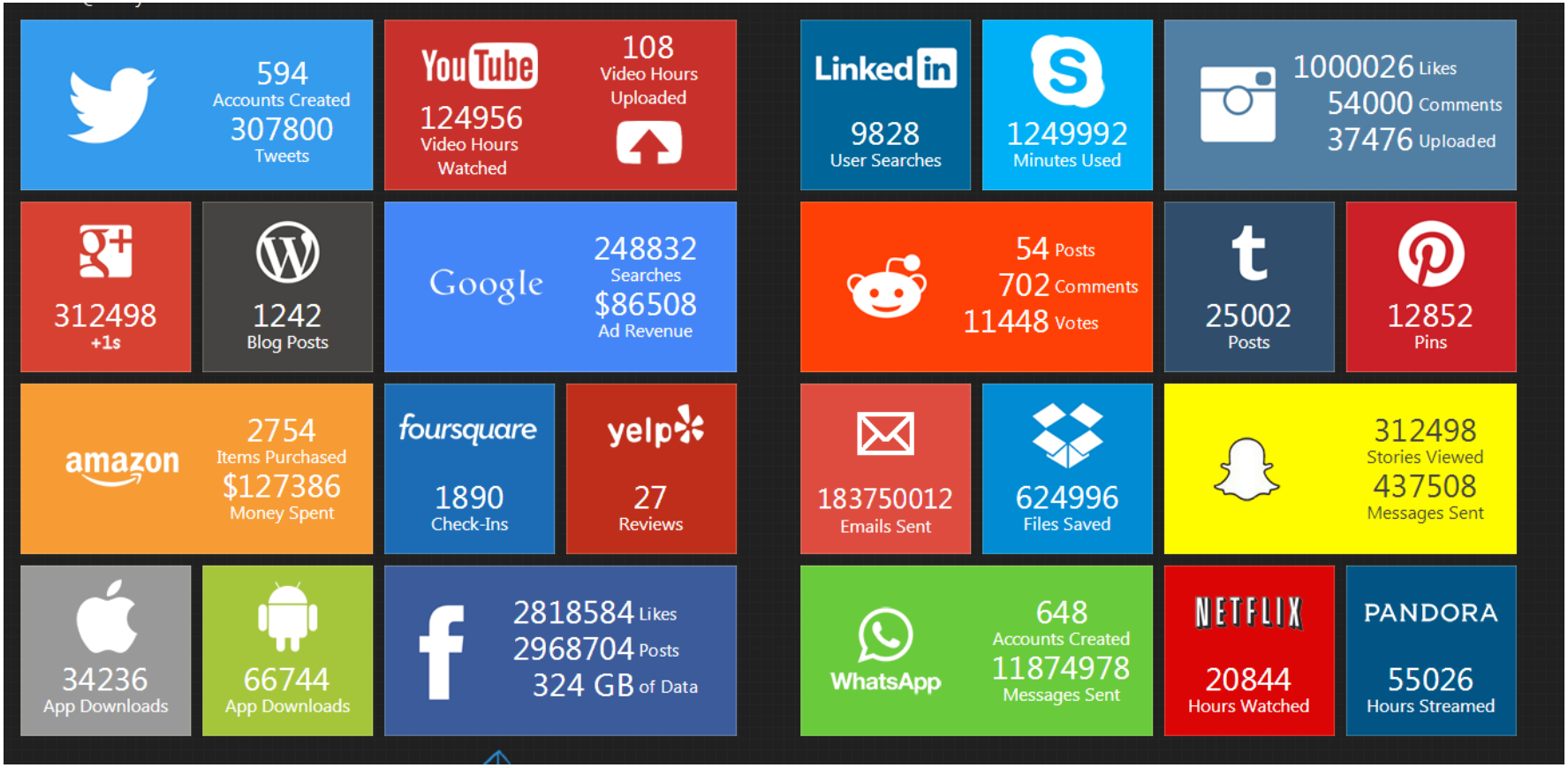
- Računala ne mogu prevesti literalno djelo?
- Računala ne mogu stvoriti literalno djelo?

NLP danas jedno od najživljih područja

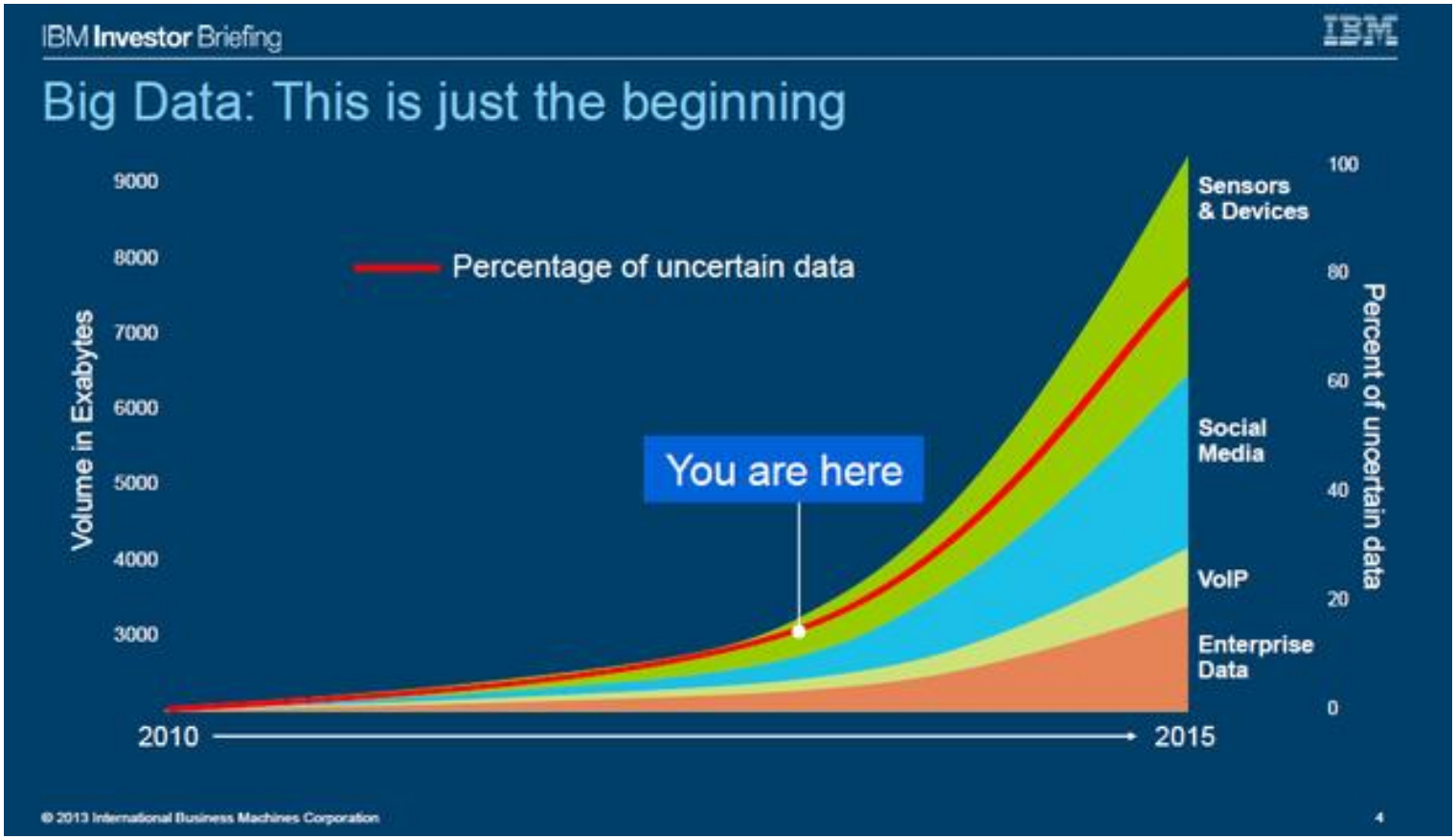
- analiza podataka velikog opsega (big data) podrazumijeva i rad s nestrukturiranim podacima – tekstovima, tweetovima, analizu mikroblogova, ...
- analiza stavova i mišljenja (sentiment analysis) – opet rad s tekstom
- sustavi za preporuku (recomender systems) – analiza komentara – opet tekst
- ekstrakcija informacija, information retrieval, question answering, text mining,....
- strojno prevođenje,...
- sve navedeno su i moguće teme za seminar!!!!

Usko vezano uz „Big data”

- https://visual.ly/community/infographic/how/internet-real-time



Struktura podataka (IBM)



Structured data

- data sources come in a **clear, predefined format**
- format is specified in details
- every piece of information is:
  - known ahead of time
  - has predefined format
  - occurs in a specified order
- this makes it easy to work with.
  - Example: RDBMS,

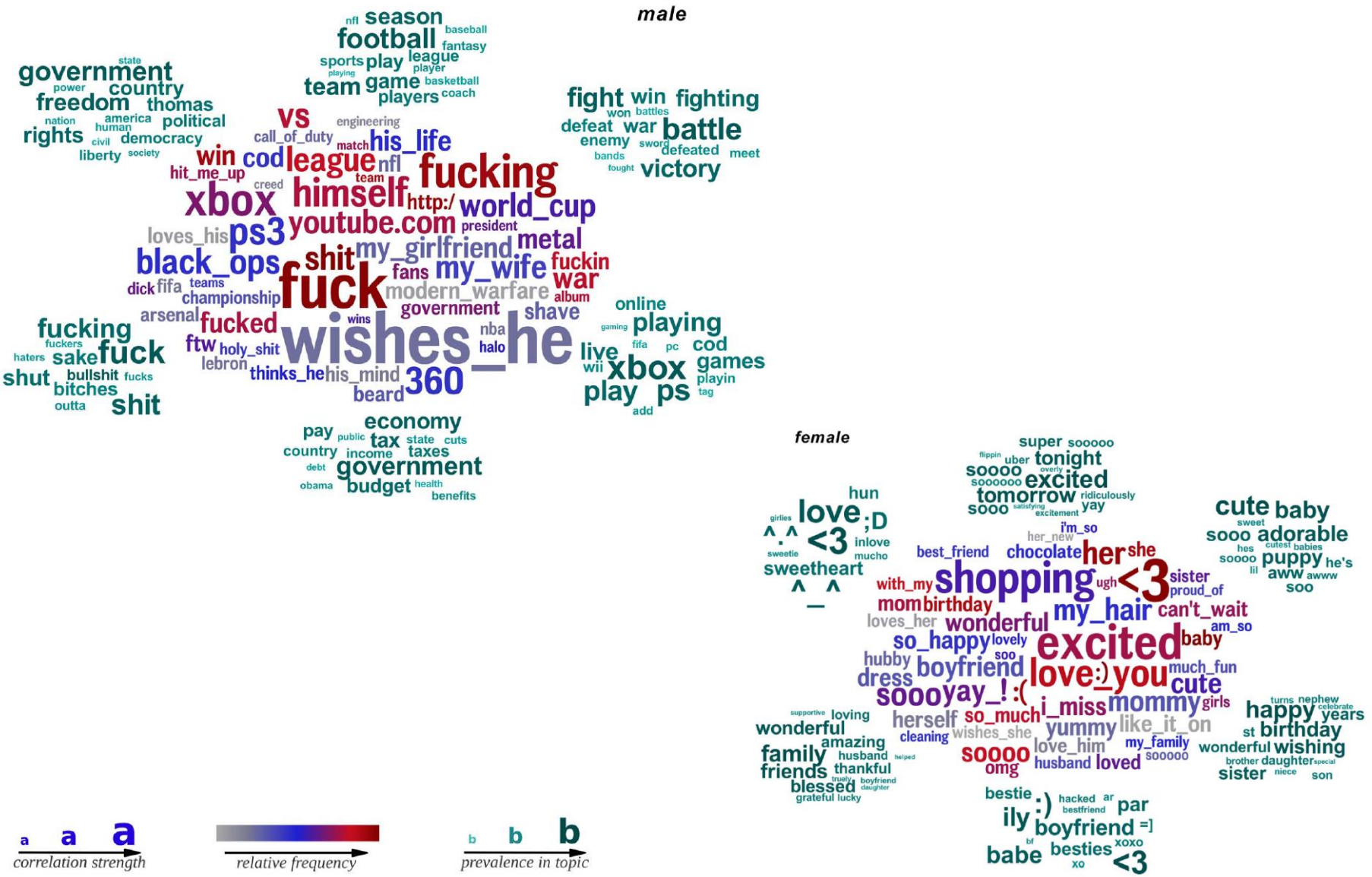
Semi-structured data

- has a logical flow and format - but the format is not user-friendly
  - a lot of noise or unnecessary data intermixed with the nuggets of high value in such a feed
- for reading is necessary to employ complex rules
  - dynamically determine how to proceed after reading each piece of information
- can contain tags that separate semantic elements
  - which includes the capability to enforce hierarchies within the data
- Example: web logs, click streams, call center logs, ..

Unstructured data

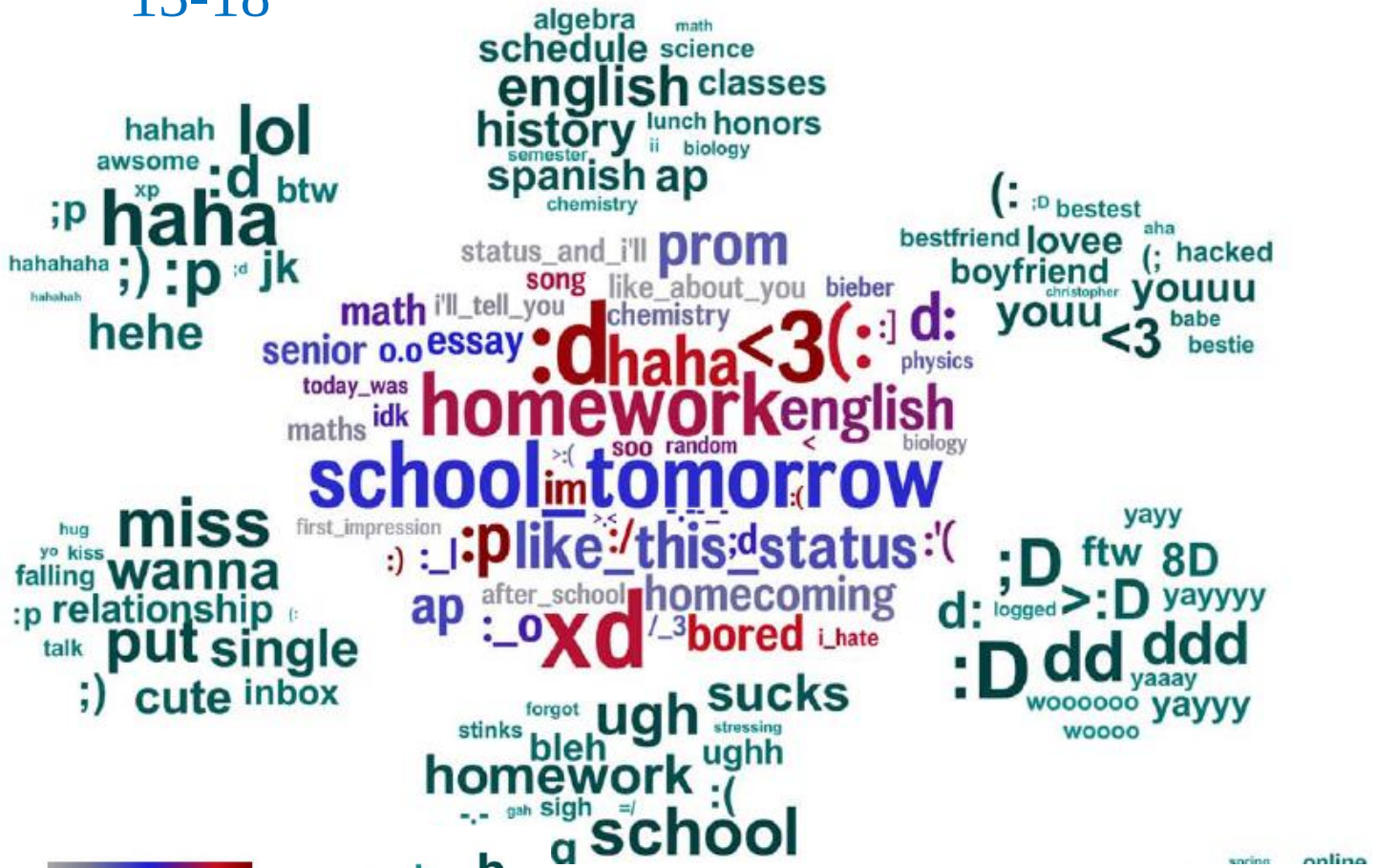
- is information that either does not have a predefined data model
  - and/or does not fit well into a relational database
- example: text, audio, video, image, geospatial and Internet data
- social networks

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, et al. (**2013**) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791



skupina?

13-18



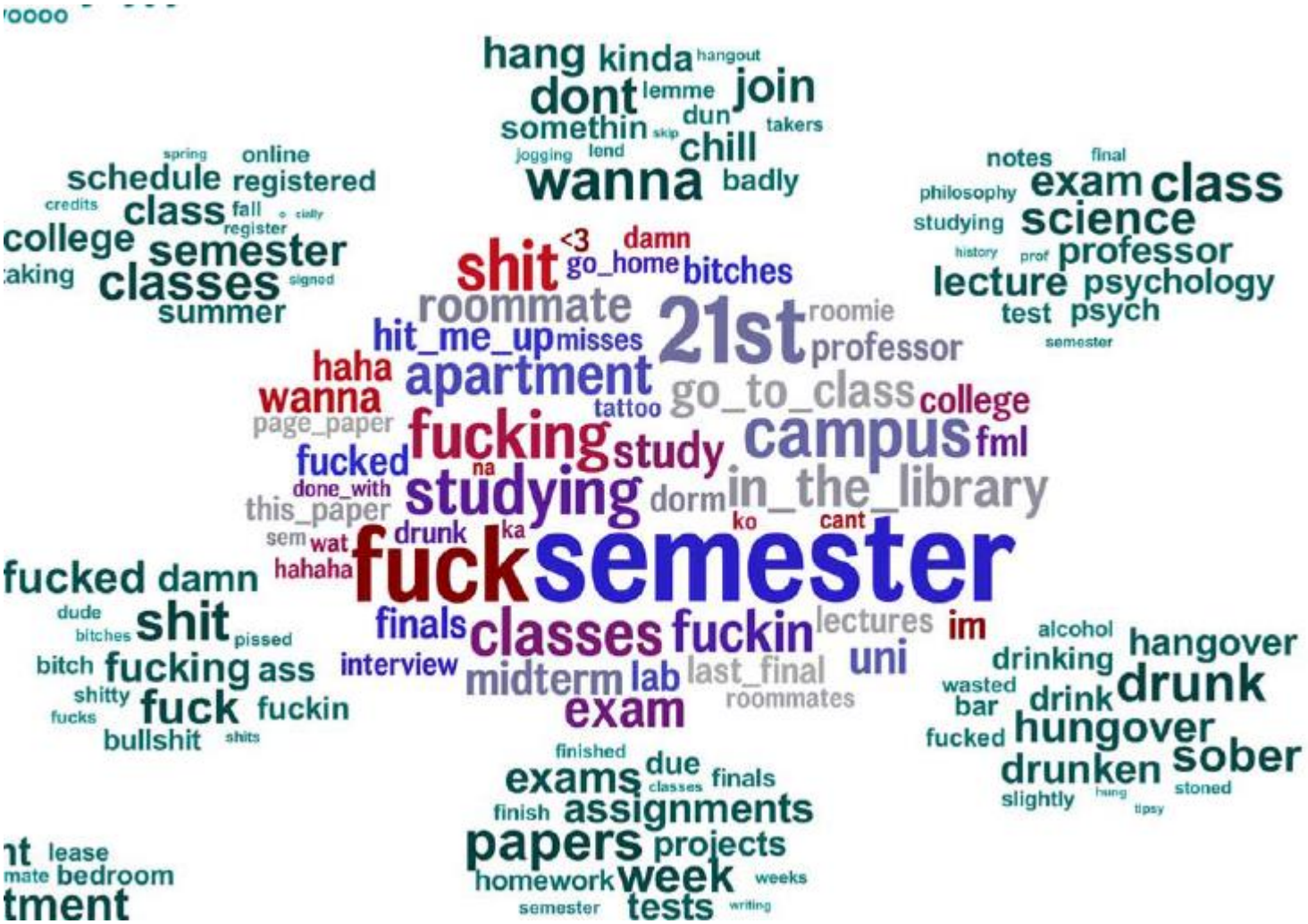
a a a
correlation strength

relative frequency

b b b
prevalence in topic

?

19-23


hang kinda hangout
dont lemme join
somethin dun takers
skip
jogging lend
chill
wanna badly
spring online
schedule registered
credits
class fall
register
college semester
taking classes signed
summer
<3 damn
shit go_home bitches
roommate 21st roomie
hit_me_up misses professor
haha apartment
wanna tattoo go_to_class college
page_paper
fucking study campus fml
fucked na
done_with studying dorm in_the_library
this_paper
sem wat drunk ka ko cant
hahaha fucksemester
fucked damn
dude
bitches shit pissed
bitch fucking ass
shitty
fucks fuck fuckin
bullshit shits
finals classes fuckin lectures im
interview midterm lab last_final uni
roommates
exam
finished
exams due finals
classes
finish assignments
papers projects
homework week weeks
semester tests writing
lease
mate bedroom
tment
notes final
philosophy exam class
studying science
history prof professor
lecture psychology
test psych
semester
alcohol
drinking hangover
wasted
bar drink drunk
fucked hungover
drunken sober
slightly hung stoned
tipsy

Pregled po disciplinama

- Lingvistika – Computational Linguistics
- Računarstvo – Natural Language Processing
  - » Artificial Intelligence
- Elektrotehnika – Pattern Recognition,
  - » Speech Recognition
- Psihologija – Computational Psyholinguistics
- Affective Computing.....

Znakovi i oznake

- skupovi
- abecede
- nizovi
- Booleova logika
- jezik
  - operacije nad jezicima
- grafovi
- stabla
- relacije

Skupovi

- objekti skupa $A=\{a_1,a_2,a_3,..,a_n\}$
- skup A je podskupskupa B: $A \subseteq B$
- beskonačni skupovi $A=\{a_1,a_2,a_3,..\}$
- prazni skup: $\emptyset$
- unija skupova: $A \cup B$
- presjek skupova: $A \cap B$

Skupovi

- kardinalni broj skupa je broj elemenata u konačnome skupu
  - ako između konačnih skupova postoji bijekcija imaju jednake kardinalne brojeve
  - ako je jedan podskup pravi podskup drugoga onda nemaju jednake kardinalne brojeve
  - prebrojivo beskonačni su skupovi za koje postoji bijekcija na skup prirodnih brojeva
  - neprebrojivo beskonačni skupovi – skup realnih brojeva
- unija: $A \cup B = \{x \mid x \in A \vee x \in B\}$
- presjek: $A \cap B = \{x \mid x \in A \wedge x \in B\}$

Abeceda

- ***Abeceda:*** konačni skup znakova
- abeceda se sastoji od simbola
  - binarna abeceda
    - B={0,1}, nad brojkama 0 i 1
  - engleska abeceda

    $\forall\Sigma_1$={a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z}
  - hrvatska abeceda

    $\forall\Sigma_2$={a,b,c,č.ć,d,dž,đ,e,f,g,h,i,j,k,l,lj,m,n,nj,o,p,q,r,s,š,t,u,v,z, ž}
  - abeceda Γ

    $\forall\Gamma$={0,1, x,y,z}

Niz

- **je konačni slijed znakova abecede postavljen jedan do drugog**
  - 011, 1011, 10110001 su nizovi nad abecedom B={0,1}
  - *duljina* niza: je broj znakova u nizu (w=101100, |w|=6)
  - *prazan* niz: je niz dužine 0 i označava se ε; | ε|=0
  - *prefiks* niza w: dobije se odbacivanjem 0, 1 ili više posljednjih znakova niza w
  - *sufiks* niza w: dobije se odbacivanjem 0, 1 ili više početnih znakova niza w

Niz II

- ***podniz*** niza w: dobije se odbacivanjem prefiks i sufiksa niza w
- ***pravi*** podniz, prefiks ili sufiks niza w: je neprazan niz x koji je podniz, prefiks ili sufiks niza w i $w \neq x$.
- ***podslijed*** niza w: dobije se odbacivanjem niti jednog, jednog ili više ne uzastopnih znakova niza w
- ***nadovezivanje*** *(konkatenacija)* nizova x i y: dobije se dodavanjem znakova niza y iza znakova niza x, bez razmaka; $w=xy$
- ***palindrom***: niz znakova koji se s obje stane čita jednako: 0, 00, 010, 1101011, abcba

Booleova logika

- Vrijednosti:
  - **TRUE** / točno / 1 i **FALSE** / netočno / 0
- AND / konjunkcija / I: $A \wedge B$
- OR / disjunkcija / ILI: $A \vee B$
- NOT / negacija / NE: $\neg A$
- XOR/ ekskluzivno ILI: $A \oplus B$
- IMPLIKACIJA $\Rightarrow / \rightarrow$
- EKVIVALENCIJA $\Leftrightarrow / \leftrightarrow$

Jezik

- ***Jezik*** je skup nizova nad abecedom
  - npr.: prazan skup $L_1 = \{\}$ ili skup nizova koji imaju paran broj nula i jedinica $L_2 = \{00,11,0011,0101,1010,1001,...\}$ L_2 nije konačan skup
  - skup svih mogućih nizova nad nekom abecedom - Σ ; nije konačan i označava se sa Σ^*
  - npr: jezik $L_3 = \Sigma^* = \{\varepsilon,0,1,01,11,00001,...\}$ je skup svih mogućih nizova znamenaka iz abecede B={0, 1}

Operacije nad jezicima

- ***Unija*** jezika L i N: $L \cup N = \{w \mid w \in L \vee w \in N\}$
- ***Presjek*** jezika L i N: $L \cap N = \{w \mid w \in L \wedge w \in N\}$
- ***Razlika*** jezika L i N:

  $L - N = L \setminus N = \{w \mid w \in L \wedge w \notin N\}$
- ***Nadovezivanje*** jezika L i N:

  $L\,N = \{xy \mid x \in L \wedge y \in N\}$
- ***Kartezijev produkt :*** $L \times N = \{(x,y) \mid x \in L \wedge y \in N\}$
- ***Partitativni skup***:

  2^L = skup svih podskupova jezika L
- ***Komplement*** jezika L: $L^C = \{w \mid w \notin L\}$

Operacije nad jezicima

- ***Kleenov operator L^*:***

$$L^* = \bigcup_{i=0}^{\infty} L^i = L^0 \cup L^1 \cup L^2 \cup \ldots$$

- ***Kleenov operator L^+:***

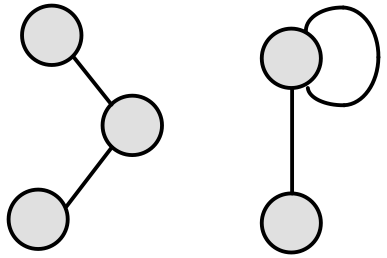
$$L^+ = \bigcup_{i=1}^{\infty} L^i$$

  - nadovezivanje jezika samim sobom: $L^2=LL$
  - definira se:

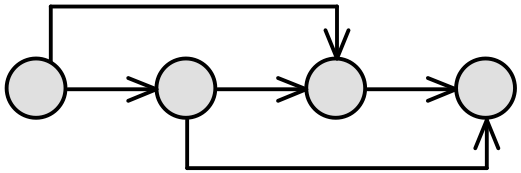
$$L^0=\{\varepsilon\}$$

$$L^i= L^{i-1}L$$

Grafovi



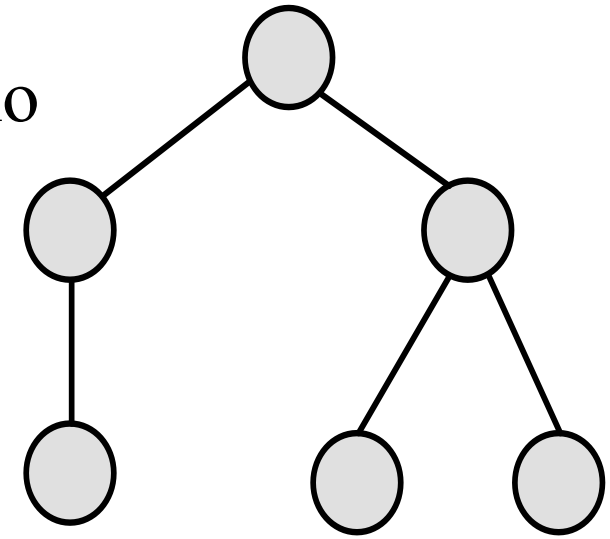
- G=(V, E), gdje su
  - V- čvorovi, E – parovi čvorova, grane
  - put grafa je niz čvorova v_1, v_2, ..v_k, ($k \geq 1$), za koji vrijedi (v_i, v_{i+1}) je grana grafa, $\forall$ $1 \leq i \leq k$
  - duljina puta: k-1



- Usmjereni graf:
  - grane usmjerenog grafa su uređeni parovi čvorova – usmjerene grane
  - neposredni prethodnik, neposredni sljedbenik

Stablo

- je usmjereni graf za kojeg vrijedi:
  - čvor u korijenu nema prethodnika, i od njega vodi put do ostalih čvorova
  - svaki čvor, osim korijena ima točno jednog neposrednog prethodnika
  - roditelj – dijete
  - preci – potomci
  - korijen- unutrašnji čvor- list
  - stupanj: broj neposrednih nasljednika



Relacije

- Binarna relacija je skup uređeni parova elemenata skupova $(a,b) \in R$: aRb
  - element a je iz skupa domene, element b je iz skupa kodomene (domena i kodomena mogu biti isti skup S)
  - ***refleksivnost:*** $\forall a \in S: aRa$
  - ***nerefleksinost***: NE $\exists\, a \in S: aRa$
  - ***simetričnost:*** $\forall a \in S, \forall b \in S : aRb \Rightarrow bRa$
  - ***asimetričnost:*** $\forall a \in S, \forall b \in S : aRb \Rightarrow$ NE bRa
  - ***tranzitivnost:*** $\forall a, b, c \in S: aRb \wedge bRc \Rightarrow aRc$
  - ekvivalencija: refleksivna + simetrična + tranzitivna relacija

Okruženje relacije

– $\mathcal{P}$ je skup svojstava relacije R

– $\mathcal{P}$ – *okruženje* relacije R je najmanja relacija R ´ koja uključuje sve parova iz R i ima svojstva iz $\mathcal{P}$

– ***tranzitivno okruženje*** relacije R: R^+

- ako je $(a,b) \in R$ onda je $(a,b) \in R^+$
- ako je $(a,b) \in R^+$ i ako je $(b,c) \in R$ onda je $(a,c) \in R^+$
- niti jedan drugi element nije u R^+
- skup R^+ uključuje skup R i ima svojstvo tranzitivnosti

– ***refleksivno i tranzitivno okruženje*** relacije R:

- $R^* = R^+ \cup \{ (a,a) \mid a \in S\}$

Formalni jezici i jezični procesori I

Prof. dr. sc. Sanda Martinčić - Ipšić

smarti@inf.uniri.hr

Formalni jezici i jezični procesori II

- Rad i izgradnja jezičnih procesora. Osnovne faze prevođenja programa.
- Analiza izvornog programa. Leksička analiza. Podatkovne strukture leksičke analize. Nejednoznačnosti i postupci oporavka kod pogreške. LEX i FLEX. Sintaksna analiza. Podatkovne strukture sintaksne analize. Sintaksna pravila. Parsiranje (od vrha prema dnu i od dna prema vrhu). YACC. Semantička analiza. Gradnja sintaksnog stabla. Prevođenje od vrha prema dnu. Rekurzivno prevođenje.
- Sinteza ciljnog programa. Dodjela memorije. Pristup nelokalnim imenima. Razmjena parametara. Generiranje međukoda. Generiranje ciljnog programa. Priprema izvođenja ciljnog programa. Optimiranje.

Jezični procesori

- jezični procesor (compiler) prevodi zapis nekog programa iz izvornog jezika u ciljni jezik
- projektiranje jezičnih procesora koristi znanja iz područja programskih jezika, arhitektura računala, teorije jezika, algoritama, projektiranja programske opreme

Zadatak JP

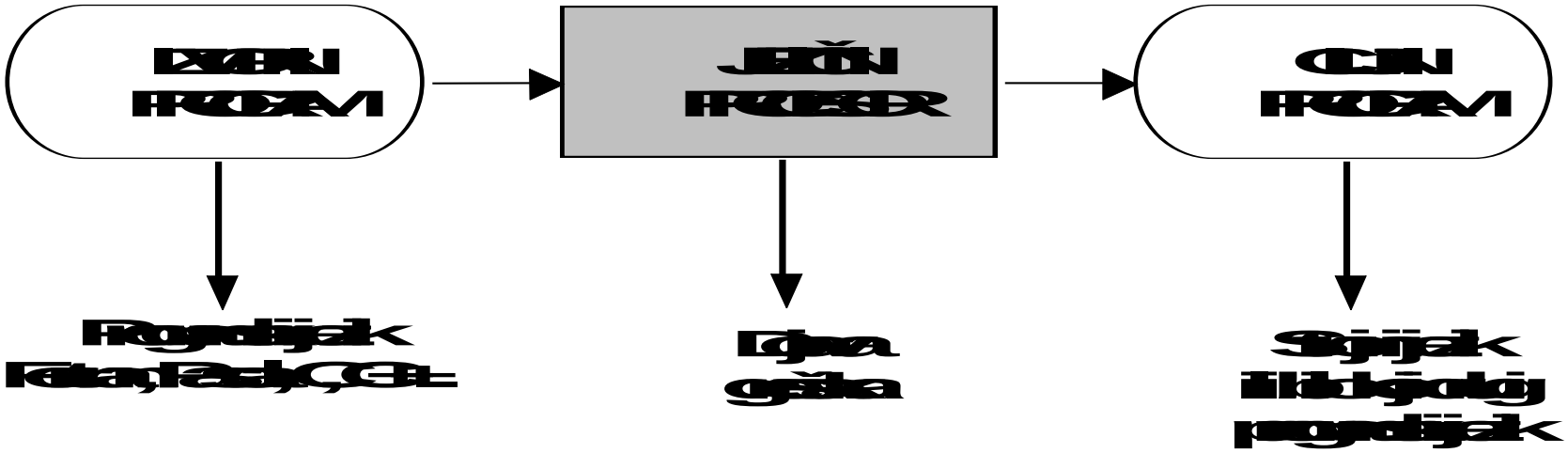
- zadatak jezičnog procesora je da svojim sučeljem približi računalo za uporabu na različitim područjima primjene
  - koristimo varijable, iterativne i rekurzivne algoritme, polja, datoteke, apstraktne tipove podataka..... a pri tome
  - ne treba poznavati sklopovsku građu računala, registre, binarni zapis, numeričke adrese memorijske hijerarhije,...
  - osloboditi jezik programiranja od arhitekture računala

Zadatak JP II

- olakšati učenje viših programskih jezika
- olakšati razvoj i razumijevanje programa
- olakšati ispravljanje i pronalaženje pogrešaka
- olakšati održavanje i dokumentiranje
- povećati prenosivost programa između različitih računala (pa čak i na mobitele)
- skratiti vrijeme razvoja programa

Jezični procesor (Compiler)

Def.: je program, koji neki program napisan u nekom određenom programskom jeziku, čita i prevodi u ekvivalentan program u nekom drugom jeziku.



Jezični procesor II

- priprema korisničkog programa za izvođenje na računalu je složen višeslojni proces postupnog prevođenja korisničkog programa u izvediv strojni program
- većinom je prevođenje automatizirano
- strojne naredbe koje je moguće izvesti su uvijek nizovi nula i jedinica

Analiza izvornog programa

1. **Linerana analiza (leksička):**
   - čitanje znakova izvornog programa s lijeva na desno i grupiraje znakova u simbole
   - SIMBOL: niz znakova, koji zajedno imaju značenje
2. **Hijerahijska analiza (sintaksna):**
   - znakovi i simboli se udružuju u grupe, koje imaju neko određeno značenje
3. **Semantička analiza:**
   - provjerava smislenost dijelova programa

Sinteza ciljnog programa

4. Generiranje međukoda:

– zapis izvornog programa u obliku koji se može izravno spremiti u memoriju računala

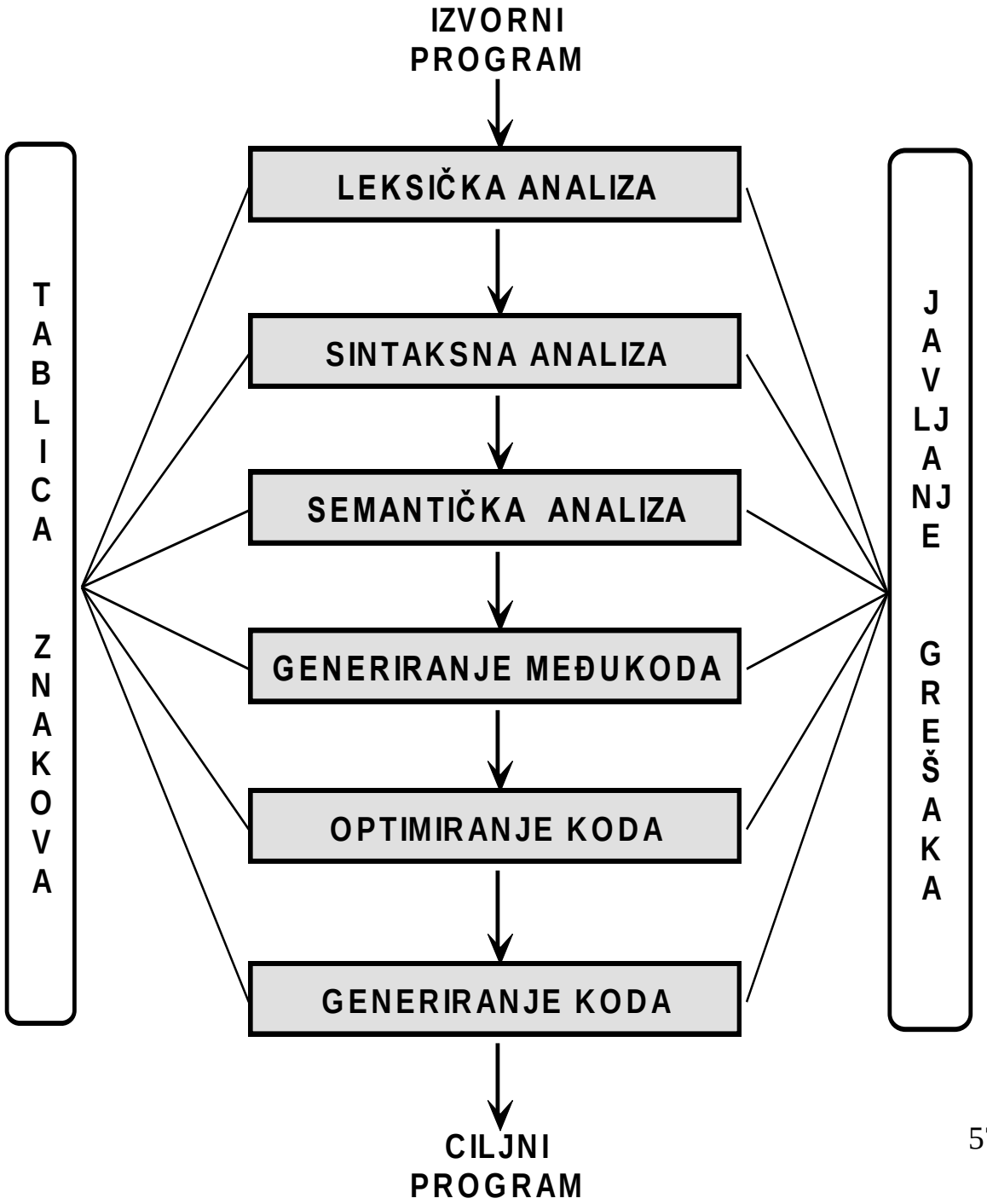
5. Optimiranje međukoda:

– poboljšanje kvalitete ciljnog programa: brzine izvođenja, veličine memorijskog prostora

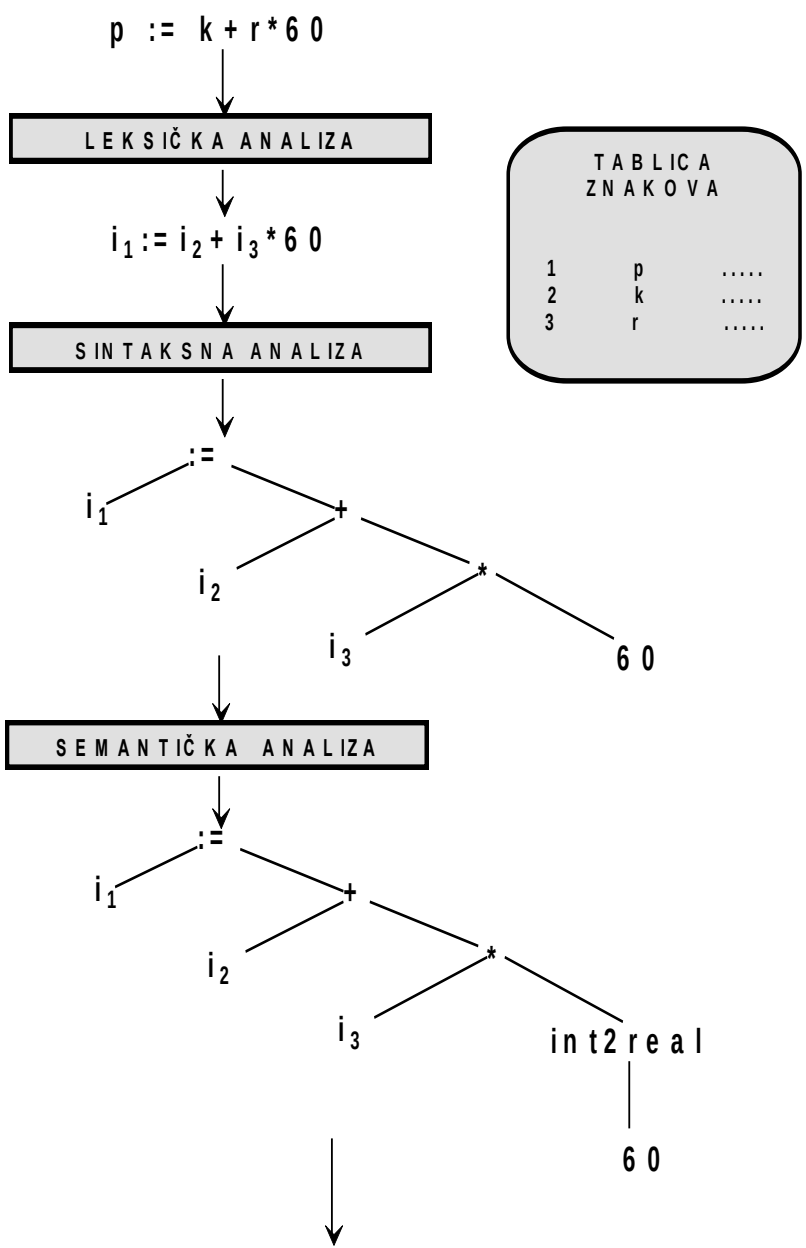
6. Generiranje koda:

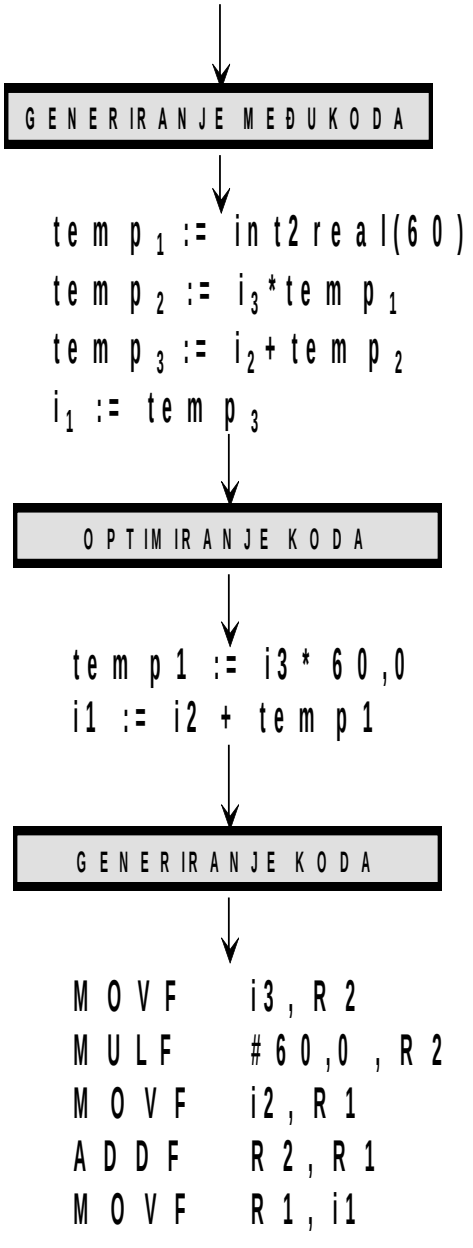
– generiranje strojnog programa

Faze rada jezičnog procesora

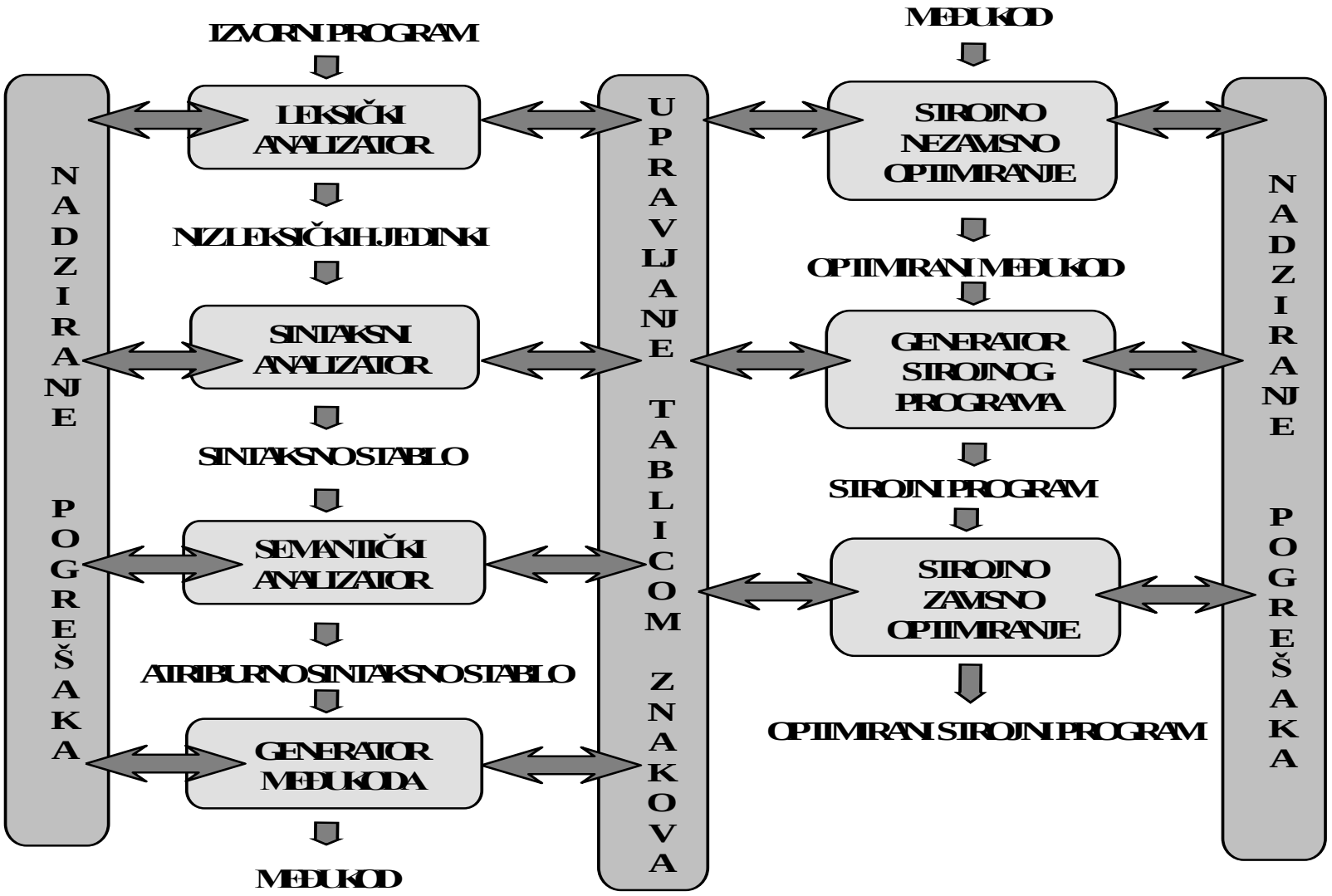


Primjer:rad jezičnog procesora





Rad jezičnog procesora



Prihvaćanje izvornog programa

- ispituje ispravnost izvornog programa pomoću **formalnih automata** s kojim se ispituje prihvaćanje nizova
  - čita se znak po znak izvornog programa

Generiranje ciljnog programa

- ciljni program se generira pomoću **formalne gramatike**