

# Autorska prava, plagijati i metode za utvrđivanje plagiranja

Vedran Juričić  
Filozofski fakultet  
Sveučilište u Zagrebu

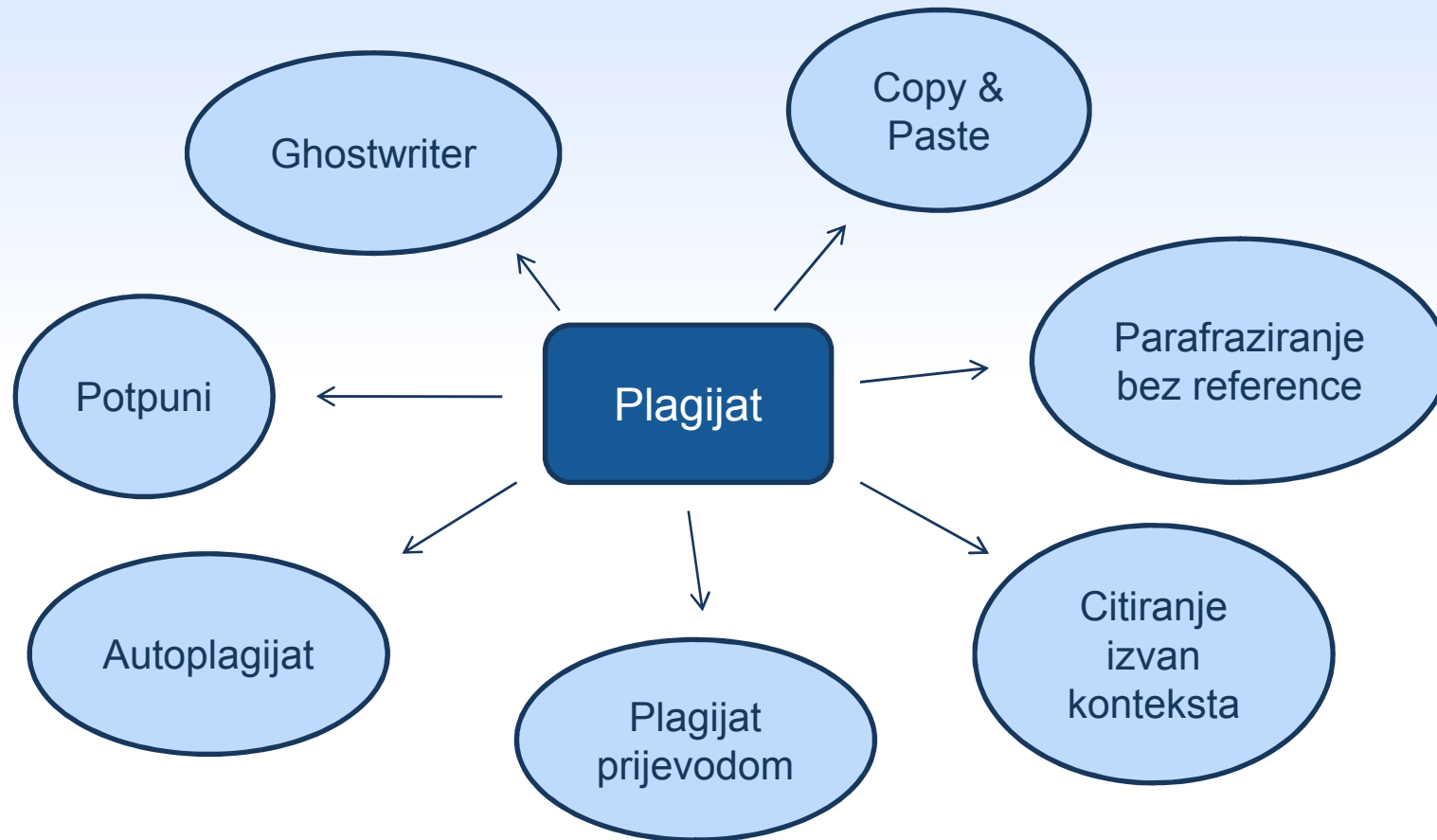
# Agenda

1. Plagijati
  - Što je plagijat, a što nije?
  - Zakoni i primjeri
2. Detekcija
  - Sustavi za detekciju
  - Problemi
  - Modeli podataka
  - Algoritmi
3. Praktični dio

# Što je plagijat?

- Čin prisvajanja ili kopiranja tuđeg pisanog, umjetničkog ili drugog kreativnog rada u svoj vlastiti, bilo u cjelini ili djelomice
- Bez prikladnog priznanja izvornog autorstva ili izvornika

# Pojavni oblici



# Pojavni oblici

## 1. *Ghostwriter*

- osoba nije autor teksta, nego je tekst napisao netko drugi u ime te osobe

## 2. **Potpuni plagijat**

- osoba potpisuje cijelo tuđe djelo svojim imenom

## 3. **Autoplaging**

- predstavljanje vlastitog prethodno objavljenog rada kao izvornog

## 4. **Plagijat prijevodom**

- osoba objavljuje prijevod tuđeg teksta bez navođenja izvora

# Pojavni oblici

## 5. ***Copy&Paste*** plagijat

- osoba preuzima dijelove tuđeg teksta bez navođenja izvora (uključujući Internet)

## 6. **Parafraziranje bez reference**

- preuzimanje tuđeg teksta ili ideja, ali ne doslovno

## 7. **Citiranje izvan konteksta**

- osoba prepisuje ili parafrazira tekst, ali ne citira precizno

# Autorsko djelo

- Zakon o autorskom pravu i srodnim pravima, Narodne novine, 167/03
- **Autorsko djelo** je originalna intelektualna tvorevina iz
  - književnoga
  - znanstvenog
  - i umjetničkog područja
- koja ima individualni karakter, bez obzira na način i oblik izražavanja, vrstu, vrijednost ili namjenu

# Što nije plagijat?

- Zakon o autorskom pravu i srodnim pravima, Narodne novine, 167/03
- Predmetom autorskog prava su **izražaji, a ne ideje**, postupci, metode rada ili matematički koncepti kao takvi.
- Copyright Law of the United States and Related Laws
- Copyright Law of the European Union

# Što nije plagijat?

- J. K. Rowling - Harry Potter
- N. N. - Mali čarobnjaci
  - Serija knjiga o školi za čarobnjake i pustolovinama skupine mladih čarobnjaka
  - Bez korištenja stvarne kopije ili likova iz Harry Pottera
- Ideja nije predmet autorskog prava

# Što je krivotvorenje?

- Imitiranje, oponašanje nečijeg rada s namjerom da se djelo predstavi kao original
- Plagiranje
  - Nezakonito pripisivanje tuđeg rada
- Krivotvorenje
  - Autentičnost rada ili djela

# Web stranice

- Općenito o plagijatima
  - <http://www.plagiarism.org>
- Plagijati pjesama
  - <http://www.plagijati.info>
- Muzej plagijata
  - Solingen, Njemačka
  - <http://www.plagiarius.de>

# Detekcija plagijata

- Proces pronalaženja plagijata u radu ili dokumentu
  1. Ručna
    - Nepraktična u slučaju velikog broja dokumenata
    - Originalni dokumenti ne moraju biti dostupni
  2. Računalno potpomognuta

# Sustavi za detekciju

- Uspoređuju “sumnjive” dokumente s velikim brojem ostalih dokumenata
- Pronalaze zajedničke dijelove
- Izračunavaju sličnost i prezentiraju rezultat
- Računalno potpomognuta detekcija
  - Nemogućnost provjere ispravno referenciranog teksta
  - Konačnu odluku donosi čovjek

# Podjela sustava

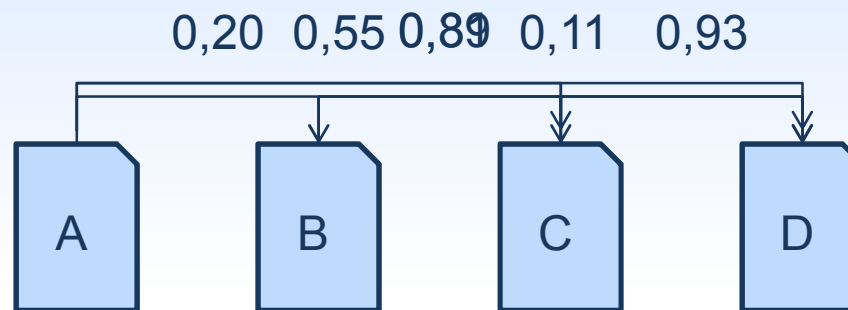
- Tip dokumenta
  1. Tekstualni dokumenti
  2. Programski kod
  3. Dizajn
  
- Cijena
  1. Besplatni
  2. Komercijalni

# Razlike u izvedbi

- Područje pretraživanja
  - Internet, lokalni podaci
- Trajanje analize
- Kapacitet
- Algoritam usporedbe
  
- Preciznost
- Odziv
- $F_1 = (2 \times P \times O) / (P + O)$

# Rad sustava

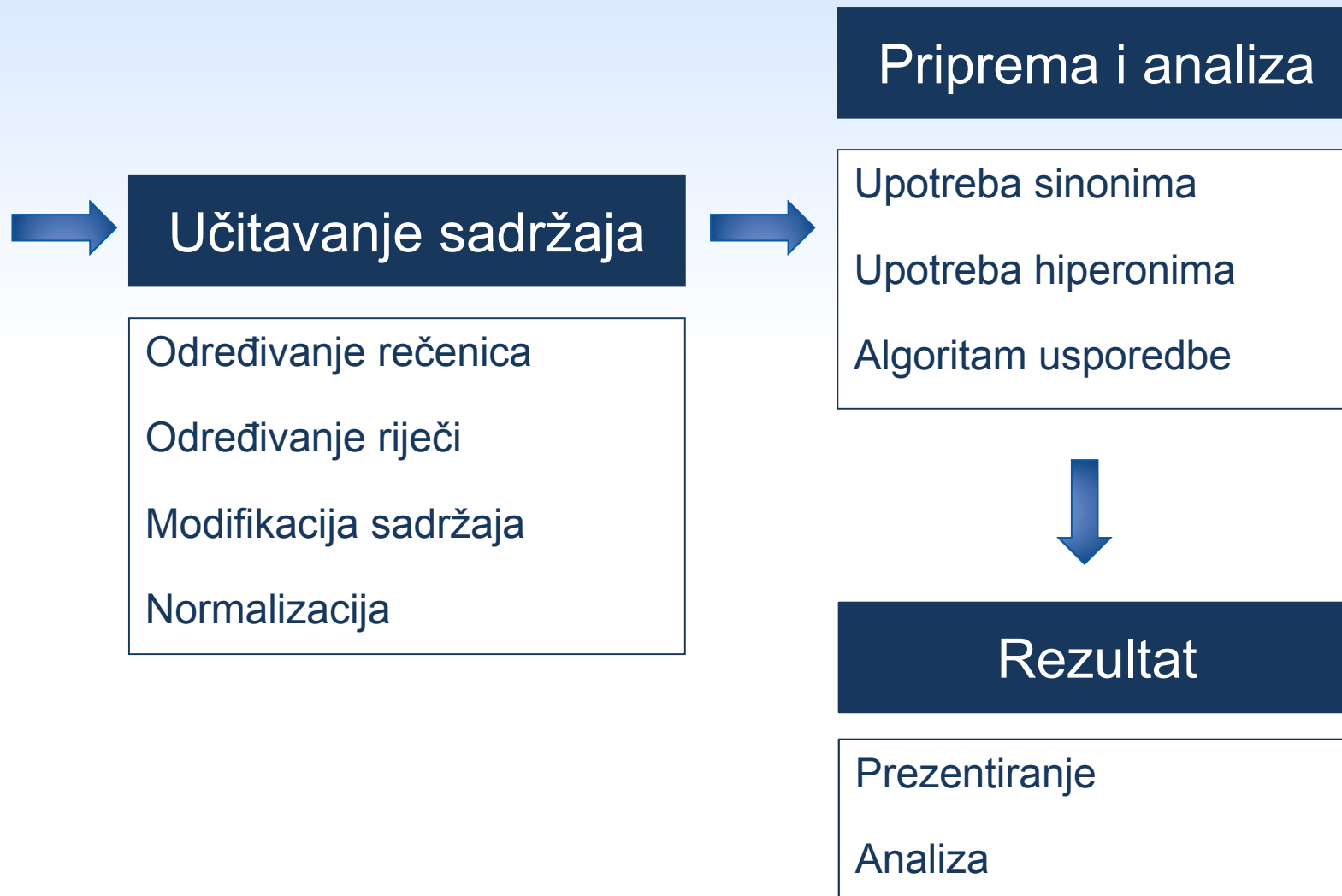
Dokumenti



Tablica  
sličnosti

	A	B	C	D
A	-	0,20	0,55	0,31
B	0,20	-	0,89	0,11
C	0,55	0,89	-	0,93
D	0,31	0,11	0,93	-

# Koraci rada



# Određivanje rečenica

- Završava s točkom, uskličnikom, upitnikom...
- Ali “Rezultati će biti objavljeni 21. lipnja u 9:00 sati.”
- engl. *sentence boundary disambiguation*
- **Popularni pristup (točnost 95%)**
  - Ako je točka, ona završava rečenicu
  - Ako je prije točke skraćunica, ona ne završava rečenicu
  - Ako riječ iza točke počinje velikim slovom, ona završava rečenicu

# Određivanje rečenica

- **Kratice**

- ... parni broj, npr. deset.
- Danas će prof. Marković održati predavanje.

- **Upravni govor**

- Ana je rekla: “Dobar dan!”

- **Veliko početno slovo**

- 7 je neparan broj. 6 je broj.
- Ana, Marko, Ivan, itd. 6 je broj.

# Modifikacija sadržaja

- Uklanjanje "suvišnih" znakova
- Zagrade
  - (), [], {}
- Točka, zarez, upitnik...
- Navodnik, apostrof
- Spojnica
- Kosa crta

# Modifikacija sadržaja

- Stop riječi
  - Riječi nevažne za analizu i usporedbu
  - Bez posebnog značenja
- Veznici
  - i, pa, ..., a, ali..., ili
- Brojevi
- Datumi
- Protiv, bilo koji, oko, po, time, dakle

# Normalizacija

- Riječi se u tekstu pojavljuju u različitim morfološkim oblicima
- Primjer
  - škola - školu, lijep - najljepši (fleksija)
  - škola - školski, lijep - ljepota (derivacija)
- Negativan utjecaj na rezultate usporedbe

# Normalizacija

## 1. Korjenovanje

- Uklanjanje afikasa kako bi se dobio korijen zajednički svim oblicima
- Korijen ne mora odgovarati pravom korijenu riječi
- (police - polic)

## 2. Lematizacija

- Pronalaženje lingvistički ispravnog kanonskog oblika neke riječi (leme)
- Može uključivati razrješavanje homografije
- (police - polica)

# Lematizacija - primjeri

Ona jede jabuku.

Mi jedemo jabuke.

Sličnost **0%**

normalizacija



Ja jesti jabuka.

Ja jesti jabuka.

Sličnost **100%**

Vjetar je počeo puhati.

Počelo je puhati.

Sličnost **33%**

normalizacija



Vjetar biti početi puhati.

Početi biti puhati.

Sličnost **100%**

# Rezultat nakon 1. koraka

Osam ljudi (mjesto radnje je neodređeno) našli su se zajedno s jednim ciljem: napraviti kaubojsku predstavu. Tih osam likova, osam karaktera pratimo od početka, tj. od audicije pa sve do konačnog proizvoda - premijere. Svi ti akteri (uglavnom gubitnici) tijekom rada na predstavi razvijaju svoje životne priče utječući neminovno jedni na druge; svi se polako hvataju za predstavu i shvaćaju je kao borbu sa samim sobom te ujedno gledaju na nju kao na priliku života.

Osam čovjek mjesto radnja neodređen naći zajedno jedan cilj napraviti kaubojska predstava. Osam lik osam karakter pratiti početak audicija konačan proizvod premijera. Akter gubitnik rad predstava razviti životna priča utjecati jedan drugi polako hvatati predstava shvatiti borba ja gledati ona prilika života.

## 2. Korak - priprema podataka

- Cilj: pripremiti niz riječi dobiven iz 1. koraka za usporedbu
- Da li se svaka riječ uzima zasebno?
- Da li će se uspoređivati podnizovi riječi?
- Da li se uzimaju u obzir i sinonimi i hipernimi?

# Ferretov model

- Tekst se pretvara u nizove od tri riječi (trigram)
- U 3. koraku (ovisno o algoritmu) traži se broj trigrama zajedničkih jednom i drugom tekstu
- Pretpostavka
  - Ako su tekstovi pisani neovisno jedan o drugom, broj podudarajućih trigrama u njima je mali
- Varijanta n-grama

# Ferretov model

Danas je lijepo vrijeme u Varaždinu.



Danas biti lijep vrijeme Varaždin.

Jučer je bilo lijepo vrijeme u Varaždinu.



Jučer biti lijep vrijeme Varaždin.

## trigrami

danas biti lijep  
biti lijep vrijeme  
lijep vrijeme Varaždin

jučer biti lijep  
biti lijep vrijeme  
lijep vrijeme Varaždin

Sličnost **66%**

# Ferretov model

Danas je lijepo vrijeme u Varaždinu.



Danas biti lijep vrijeme Varaždin.

Danas je u Varaždinu baš lijepo vrijeme.



Danas biti Varaždin lijep vrijeme.

## trigrami

danas biti lijep  
biti lijep vrijeme  
lijep vrijeme Varaždin

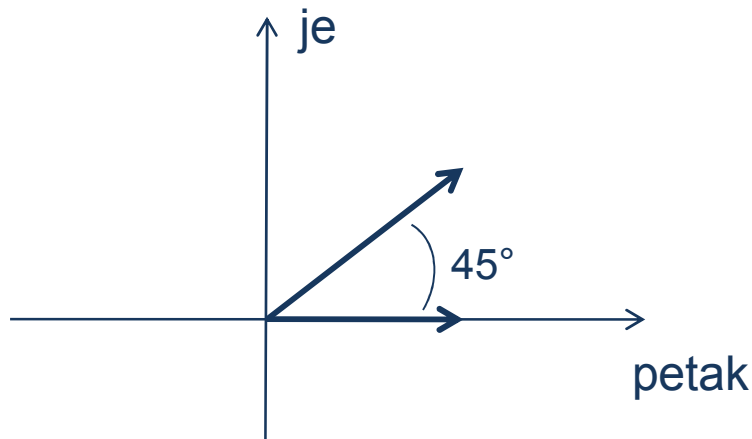
danas biti Varaždin  
biti Varaždin lijep  
Varaždin lijep vrijeme

Sličnost **0%**

- Nedostatak: osjetljiv na poredak riječi

# Model vektorskog prostora

- Tekst se prikazuje kao niz vektora u n-dimenzionalnom prostoru
- Broj dimenzija = broj različitih riječi (ili lema)



$$\vec{a} = 1 \cdot \text{petak}$$

$$\vec{b} = 1 \cdot \text{petak} + 1 \cdot \text{je}$$

Kut između vektora:  $45^\circ$

Kosinus kuta je 0,707

- Vektori su ulazni podaci za algoritam usporedbe (3. korak)
  - Skalarni produkt
  - Kut između vektora
- Sličnost se računa na temelju
  - Kosinusa kuta

$$\cos \alpha = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

# Model vektorskog prostora

Danas je lijepo vrijeme u Varaždinu.

Jučer je u Varaždinu baš lijepo vrijeme.

Danas biti lijep vrijeme Varaždin.

Jučer biti lijep vrijeme Varaždin.



$1x\text{danas} + 1x\text{biti} + 1x\text{lijep} + 1x\text{vrijeme} + 1x\text{Varaždin}$



$1x\text{jučer} + 1x\text{biti} + 1x\text{lijep} + 1x\text{vrijeme} + 1x\text{Varaždin}$

	danas	biti	lijep	vrijeme	Varaždin	jučer
(1)	1	1	1	1	1	0
(2)	0	1	1	1	1	1

Sličnost **80%**

# Model vektorskog prostora

Danas je lijepo vrijeme u Varaždinu.

Jučer je bilo lijepo vrijeme u Varaždinu.

Danas biti lijep vrijeme Varaždin.

Jučer biti Varaždin lijep vrijeme.



$1x\text{danas} + 1x\text{biti} + 1x\text{lijep} + 1x\text{vrijeme} + 1x\text{Varaždin}$



$1x\text{danas} + 1x\text{biti} + 1x\text{Varaždin} + 1x\text{lijep} + 1x\text{vrijeme}$

	danas	biti	lijep	vrijeme	Varaždin
(1)	1	1	1	1	1
(2)	1	1	1	1	1

Sličnost **100%**

# Upotreba sinonima

- Sinonimi su leksemi koji
  - Pripadaju istoj vrsti riječi
  - Imaju različite izraze
  - Značenje im se podudara potpuno ili djelomično



## istoznačnice

ljekarna - apoteka

sustav - sistem



## bliskoznačnice

muž - suprug

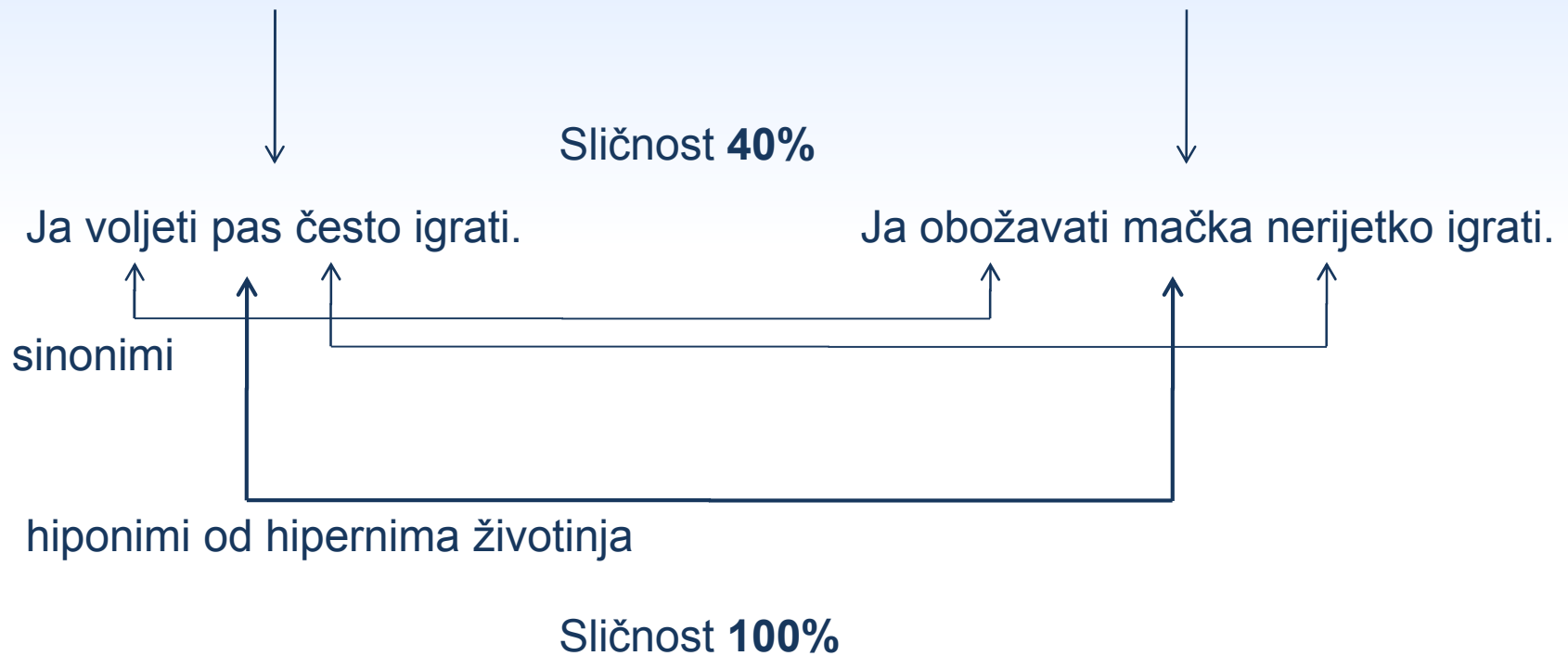
žena - supruga

# Upotreba hiperonima

- Hiperonimi su riječi ili fraze u čiji sadržaj ulaze drugi pojmovi i njihovi sadržaji
- Hiperonim je nadređen hiponimu
- Voće je hipernim od hiponima jabuka
- Bijela je hiponim od hipernima boja

# Upotreba sinonima i hipernima

Ona voli pse i često se igra s njima. On obožava mačke i nerijetko se s njima igra



# Sininimi i hiperonimi

- Engleski jezik
  - WordNet - <http://wordnet.princeton.edu/>

## Noun

S: (n) **cat**, true cat (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats)

S: (n) guy, **cat**, hombre, bozo (an informal term for a youth or man) *"a nice guy"; "the guy's only doing it for some doll"*

S: (n) **cat** (a spiteful woman gossip) *"what a cat she is!"*

S: (n) kat, khat, qat, quat, **cat**, Arabian tea, African tea (the leaves of the shrub *Catha edulis* which are chewed like tobacco or used to make tea; has the effect of a euphoric stimulant) *"in Yemen kat is used daily by 85% of adults"*

S: (n) cat-o'-nine-tails, **cat** (a whip with nine knotted cords) *"British sailors feared the cat"*

S: (n) Caterpillar, **cat** (a large tracked vehicle that is propelled by two endless metal belts; frequently used for moving earth in construction and farm work)

S: (n) big cat, **cat** (any of several large cats typically able to roar and living in the wild)

S: (n) computerized tomography, computed tomography, CT, computerized axial tomography, computed axial tomography, **CAT** (a method of examining body organs by scanning them with X rays and using a computer to construct a series of cross-sectional scans along a single axis)

## Verb

S: (v) **cat** (beat with a cat-o'-nine-tails)

S: (v) vomit, vomit up, purge, cast, sick, **cat**, be sick, disgorge, regorge, retch, puke, barf, spew, spue, chuck, upchuck, honk, regurgitate, throw up (eject the contents of the stomach through the mouth) *"After drinking too much, the students vomited"; "He purged continuously"; "The patient regurgitated the food we gave him last night"*

# Algoritmi za određivanje sličnosti

1. Hammingova udaljenost
2. Levensteinova udaljenost
3. Jaccardov koeficijent sličnosti
4. Kosinus sličnosti

# Hammingova udaljenost

- Broj bitova u kojima se razlikuju dva binarna niza
- Primjer
  - Binarni nizovi  
10011010  
10001101
  - imaju hammingovu udaljenost od 4 bita

# Levensteinova udaljenost

- Broj transformacija potrebnih da se iz jednog znakovnog niza dobije drugi
- Moguće transformacije
  - Brisanje znaka (trošak 1)  $ab - a$
  - Umetanje znaka (trošak 1)  $a - ab$
  - Zamjena znaka (trošak 1)  $a - b$
  - Kopiranje znaka (trošak 0)  $a - a$

# Levensteinova udaljenost

- Udaljenost između abcde i abcdff

	a	b	c	d	e
a	0	1	2	3	4
b	1	0	1	2	3
c	2	1	0	1	2
d	3	2	1	0	1
f	4	3	2	1	1
f	5	4	3	2	<b>2</b>

Udaljenost je 2 

Sličnost je  $1 - (2/5) = 60\%$

# Dice koeficijent

- Sličnost između dva niza riječi
- Prema formuli

$$\frac{(2 \cdot \text{Broj zajedničkih riječi})}{\text{Broj riječi u prvom nizu} + \text{Broj riječi u drugom nizu}}$$

- Primjer

$$\text{Sličnost} = 2 \cdot 2 / (5 + 4) = 44\%$$

Ana ne voli jesti jabuke.

Ivo ne voli šljive.

# Kosinus sličnost

- Sličnost dva vektora u modelu vektorskog prostora
  - Vektor X i vektor Y
  - $|X|$  je duljina vektora X,  $|Y|$  je duljina vektora Y
  - Prema formuli  $(X \cdot Y) / (|X| \cdot |Y|)$

- Primjer

Ana ne voli jesti jabuke.

Ivo ne voli šljive.

$$\text{Sličnost} = 2 / (2,23 + 2) = 47\%$$

# Praktični rad - JPlag

- Aplikacija za detekciju sličnosti između datoteka
  - Tekstualnih dokumenata
  - Programskog koda (C, C++, Java i C#)
- University of Karlsruhe
- Besplatan, potrebna je registracija
- Dostupan na
  - <https://www.ipd.uni-karlsruhe.de/jplag/>

# Jplag - primjer rada

Odabir dobitne sličnosti

JPlag - Change options

Basic options

Lang 90% - 100% 1 #####

Com 80% -

Subn 70% -

Subn 60% -

Subn 50% -

Recu 40% -

30% -

20% -

Statu 10% -

Num 0% - 1

Total number of n

Total size of subn

4.txt	->	<u>5.txt</u> (94.0%)	<u>6.txt</u> (44.5%)	<u>1.txt</u> (39.1%)	<u>3.txt</u> (21.2%)	<u>2.txt</u> (11.6%)
6.txt	->	<u>1.txt</u> (54.0%)	<u>5.txt</u> (42.3%)	<u>3.txt</u> (33.4%)	<u>2.txt</u> (10.1%)	
1.txt	->	<u>5.txt</u> (36.2%)	<u>3.txt</u> (27.1%)	<u>2.txt</u> (8.3%)		
5.txt	->	<u>3.txt</u> (19.7%)	<u>2.txt</u> (11.0%)			
3.txt	->	<u>2.txt</u> (6.7%)				

#####