**Obrada prirodnog jezika – Kolokvij 1 – LT**

**Predavanje 1**

1. Koja je najpoznatija skraćenica za obradu prirodnog jezika u engleskoj literaturi?

**NLP**, Natural Language Processing.

2. Obrada prirodnog jezika je dio (odaberite sve ispravne odgovore).

**Lingvistike** i **Umjetne inteligencije**.

3. Koji je najpoznatiji umjetni jezik?

**Esperanto**.

4. Za što se sve može koristiti obrada prirodnog jezika (odaberite sve ispravne odgovore)?

– **Automatski razgovorni agenti** (eng. chatbots, chatterbots) za korisničku podršku – ako ste ikada koristili chat funkciju na raznim web stranicama, postoji velika mogućnost da ste razgovarali sa automatskim agentom a ne sa ljudskim bićem.

– **Virtualni asistenti** (Amazon Echo, Amazon Alexa, Google Home, Apple Siri, itd.) koriste obradu prirodnog jezika i prepoznavanje govora za komunikaciju sa korisnikom.

– **Automatsko ispravljanje gramatike** koristi obradu prirodnog jezika i statističke metode za **ispravljanje teksta** kojega korisnik napiše. Na sličan način radi i automatsko generiranje teksta, npr. slijedeće riječi u rečenici.

– **Sustavi za detekciju neželjenih e-mail poruka** (eng. spam mail) koriste obradu prirodnog jezika za filtriranje tekstualnih poruka.

– **Ekstrakcija i sažimanje tekstualnih informacija** (npr. u literaturi ili medicinskoj dijagnostici).

– **Automatsko prevođenje teksta** na drugi jezik (npr. Google Translate).

– **Pretraživanje Interneta** (npr. Google Search).

– **Ekstrakcija emocija iz poruka korisnika** i određivanje njihovih mentalnih stanja.

– **Automatski nadzor tekstualnih i govornih poruka** u borbi protiv kriminala i terorizma.

– **Automatsko prevođenje** znakovnog jezika u pisani jezik.

– **Procjena rizika** za odobravanje kredita u bankama.

**– Automatska analiza i ekstrakcija informacija** iz provedenih anketa.

5. Odaberite približne godine nastanka govornog jezika.

Jezik je star minimalno **150-200 tisuća godina**. Jezik je star otprilike **100 tisuća godina**.

6. Koliko otprilike postoji jezika na svijetu?

Danas postoji otprilike nešto više od **7000 jezika** na svijetu.

7. Od ukupnog broja postojećih jezika na svijetu, polovica svjetske populacije koristi samo njih koliko?

50% svjetske populacije **koristi sveukupno 23** od tih 7000 jezika.

8. Kojim nativnim jezikom govori najveći broj ljudi na svijetu?

Najviše ljudi na svijetu govori **mandarinskim dijalektom** kineskog jezika (> milijardu ljudi).

9. U koju jezičnu porodicu spada hrvatski jezik?

**Indoeuropsku** porodicu.

10.U koju jezičnu porodicu spada engleski jezik?

**Indoeuropsku** porodicu.

11.Za koji se jezik pretpostavlja da je bio preteča indoeuropskih jezika, tj. da su jezici indoeuropske jezične skupini nastali iz njega?

Za **praindoeuropski** jezik.

12.Odaberite dva najstarija govorna jezika na svijetu.

**Sanskrit** i **Tamil**.

13.Koliko je otprilike star Sanskrit?

Star je otprilike **5000 godina**.

14.Koliko je otprilike star Tamil?

Star je otprilike **5000 godina**.

15.Koji jezik je primjer "mrtvog" jezika koji se danas ponovo koristi?

**Hebrejski** jezik.

16.Koji jezik ima gramatiku neovisnu o kontekstu?

**Sanskrit**, što znači da riječi u Sanskritu općenito ne mogu mijenjati svoje značenje bez obzira u kojem ih se kontekstu pronađe.

17.Kada je otprilike otkriveno pismo?

Otkriveno je više puta:

– u Mezopotamiji (današnji Irak) **između 3400. i 3100.** godine prije nove ere,

– u Egiptu oko **3250.** godine prije nove ere,

– u Kini oko **1200.** godine prije nove ere,

– u južnom Meksiku i Guatemali prije **500.** godine prije nove ere.

18.Koje je najstarije pismo na svijetu?

**Sumersko** pismo i **Egipatsko** pismo su najstarija prava pisma (ne proto-pisma, tj. preteče pisma) na svijetu. Sumersko pismo nastalo je otprilike 3300. godine prije nove ere na području današnjeg Iraka (Sumer je bio regija u južnoj Mezopotamiji). Egipatsko pismo nastalo je otprilike 3100. godine prije nove ere.

19.Za potrebe čega je najvjerojatnije nastao pisani jezik?

Pretpostavlja se da je pismo nastalo **za potrebe trgovine** tako da se umjesto “tokena” koji su korišteni prilikom razmjene dobara počelo te tokene crtati kada je njihov broj postao prevelik i nezgodan za razmjene.

20.Povežite sustave za pisanje sa njihovim definicijama.

**Abeceda** (eng. alphabet) je skup znakova ili simbola gdje svaki znak ili simbol predstavlja jedan glas (suglasnik ili samoglasnik).

**Abjadi** (eng. abjad,) su sustavi za pisanje slični abecedama, no uglavnom sadržavaju simbole ili znakove samo za suglasnike, ali ne i za samoglasnike.

**Silabariji** (slogovni ili silabički sustav za pisanje) (eng. syllabaries) su sustavi za pisanje gdje svaki znak predstavlja jedan slog (obično suglasnik + samoglasnik), a ne samo jedan glas.

**Logogrami** (eng. logograms) su sustavi za pisanje gdje jedan znak ili simbol može označavati jedan slog, jednu riječ ili jednu ideju.

**Abugide** (eng. abugidas) su sustavi za pisanje koji se nalaze negdje između abeceda i silabarija. Abugide imaju posebne znakove za suglasnike i za samoglasnike, no znakovi za samoglasnike se obično koriste samo kada riječ počinje sa samoglasnikom. U svim ostalim slučajevima u znakovima za suglasnike naprave se male izmjene kako bi se znalo koji samoglasnik dolazi nakon tog suglasnika.

21.Koji je sustav za pisanje najrasprostranjeniji na svijetu?

**Latinica.**

22.Odaberite abecede.

**Latinska abeceda, grčka abeceda, ćirilica**.

23.Odaberite abjade.

**Arapski** i **hebrejski**.

24.Odaberite silabarije.

U Japanu se koriste dva silabarija: **hiragana** (za nativne japanske riječi) i **katakana** (za strane i tehničke riječi).

25.Odaberite logograme.

**Mandarinski** u Kini.

26.Kada su dešifrirani egipatski hijeroglifi?

Hijeroglife je dešifrirao **1820.** godine francuski znanstvenik Jean-François Champollion.

27.Odaberite znanstvenike zaslužne za dešifriranje egipatskih hijeroglifa.

**Jean-François Champollion** i **Thomas Young**.

28.Otkriće čega je pomoglo pri dešifriranju egipatskih hijeroglifa?

Dešifriranju hijeroglifa pomoglo je **otkriće kamena iz Rosette** (eng. Rosetta stone). Kamen iz Rosette sadrži natpis na 3 jezika (egipatski hijeroglifi, egipatsko demotsko pismo, grčko pismo), a otkrila ga je Napoleonova vojska 1799. godine blizu luke zvane Rosetta (današnji Rashid).

29.Odaberite abugide.

Najčešće se koriste u Indiji i jugoistočnoj Aziji, ali koriste se i u Africi (npr. u Etiopiji) i sjevernoj Americi (npr. neki autohtoni sustavi za pisanje u Kanadi).

**Brahmi pismo, Gupta pismo, Devanagari**.

30.Odaberite sustav za pisanje koji se danas koristi za pisanje Sanskrita?

**Devanagari**.

31.Odaberite najraniji poznati zapis teksta na hrvatskom jeziku.

**Bašćanska ploča** je jedan od prvih natpisa koji sadrže tekst na hrvatskom jeziku napisan glagoljicom. Datira iz 1100. godine, a pronađena je 1851. godine na Krku u crkvi sv. Lucije. Danas se čuva u Hrvatskoj akademiji znanosti i umjetnosti.

32.Za koji jezik se pretpostavlja da je bio autohtoni ("domaći") jezik Europe?

**Baskijski** jezik, danas se govori na području sjeverne Španjolske i jugozapadne Francuske.

33.Koje je najpoznatije nedešifrirano pismo?

**Linear** **A**, nedešifrirano pismo porijeklom sa Krete koje se je koristilo od 1800. do 1450. godine prije nove ere.

**Predavanje 2**

1. Kakve vrste podataka mogu biti ulazni podaci kod obrade prirodnog jezika?

Mogu biti **tekstualni** ili **zvukovni** zapisi.

2. Što je WordNet?

To je **baza podataka** koja se koristi **u obradi prirodnog jezika**.

3. Što je tzv. synset u WordNet-u?

**Synset** je **set sinonima**, odnosno skup sličnih riječi.

4. Što je Open Multilingual WordNet?

Kroz Open Multilingual WordNet može se pristupiti bazama podataka **različitih jezika**, a tako i za hrvatski.

5. Što u slijedećem primjeru iz WordNet-a označava slovo n: "flower.n.01"?

Označava da je ta riječ **imenica** (noun).

6. Podržava li Open Multilingual WordNet hrvatski jezik?

**Podržava** kruva ti.

7. Što su regularni izrazi?

Regularni izrazi (eng. regular expressions, regex) se mogu definirati kao **jezik** koji se koristi **za specifikaciju stringova** (filtera) **za pretraživanje** **teksta**.

8. Što označava znak "\*" u regularnim izrazima?

Označava **nula ili više ponavljanja** prethodnog regularnog izraza (Kleene \*).

9. Što označava znak "+" u regularnim izrazima?

Označava **jedan ili više ponavljanja** prethodnog regularnog izraza (Kleene +).

10. Koji regularni izraz od ponuđenih će uspjeti pronaći e-mail adresu "marina@gmail.com" u nekom tekstu?

r'\w+@gmail\.com

11. Koji regularni izraz od ponuđenih će uspjeti pronaći e-mail adresu "igor26@fesb.hr" u nekom tekstu?

r'[a-z]+[0-9]\*@fesb\.hr

12. Koji regularni izraz od ponuđenih će uspjeti pronaći e-mail adrese "marina@fesb.hr" i "igor26@gmail.com" u nekom tekstu?

r'[a-z]+[0-9][\*@[a-z]+\.(?:com|hr)](mailto:*@[a-z]+\.(?:com|hr))

**Predavanje 3**

1. Odaberite sve vrste udaljenosti koje se mogu koristiti prilikom računanja udaljenosti uređivanja.

– **Levenshtein** udaljenost

– **Najduža zajednička podsekvenca niza**

– **Hamming** udaljenost

– **Damerau-Levenshtein** udaljenost

– **Jaro i Jaro-Winkler** udaljenosti

2. U kojim se sve područjima može koristiti minimalna udaljenost uređivanja?

– obradi prirodnog jezika,

– bioinformatici,

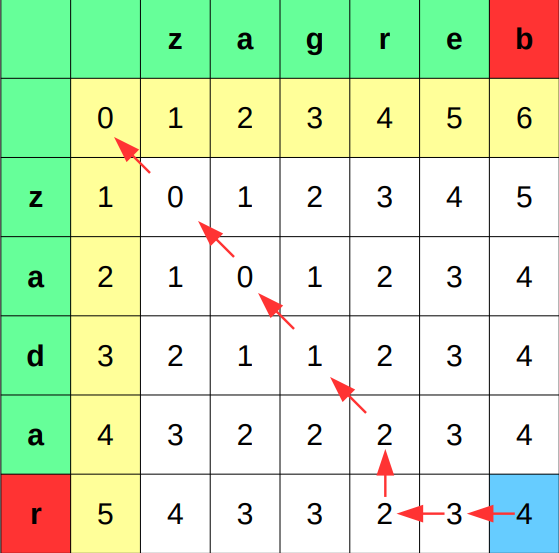
– kriptografiji,

– glazbenoj industriji (za detekciju sličnosti glazbenih djela)…

3. Izračunajte Levenshtein udaljenost uređivanja za zadane riječi. Pretpostavite da su cijene svih operacija dozvoljenih kod računanja Levenshtein udaljenosti jednake.

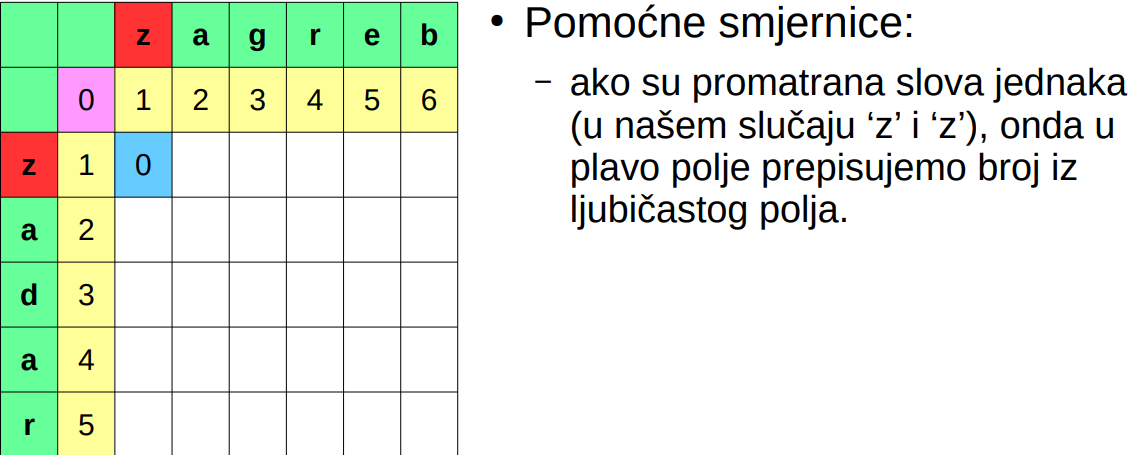
**Napomena 1:** Udaljenost između dva stringa se kod Levenshtein udaljenosti računa se kao broj **umetanja** (eng. insertions), **brisanja** (eng. deletions) i **zamjena** (eng. substitutions) znakova koje je potrebno izvršiti da bi se jedan string transformirao u drugi.

**Primjer:**

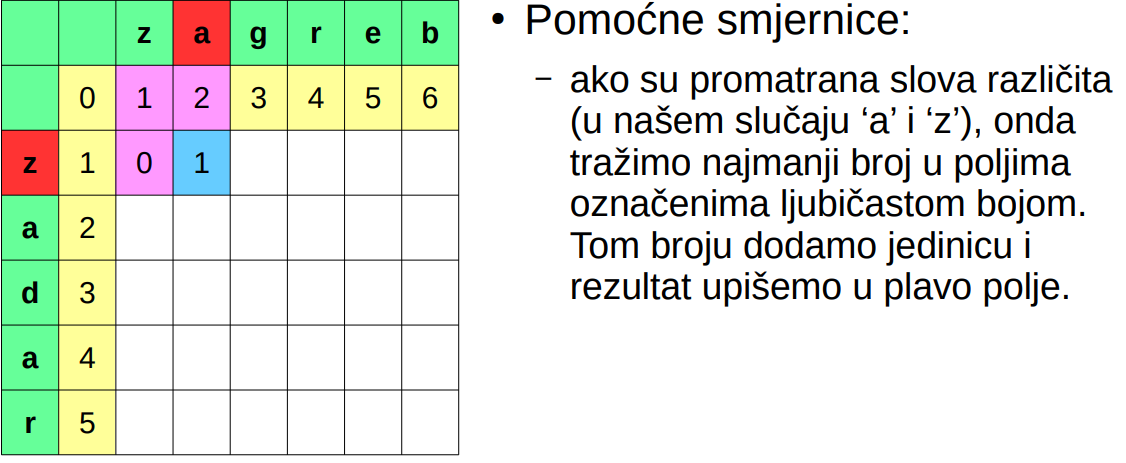


Konačni broj operacija, odnosno udaljenost je 4.

**Napomena 2:**



**Napomena 3:**



4. Odredite najdužu zajedničku podsekvencu niza za zadane riječi.

**Napomena 1:**

Dozvoljene su samo operacije **umetanja i brisanja** (nisu dopuštene zamjene).

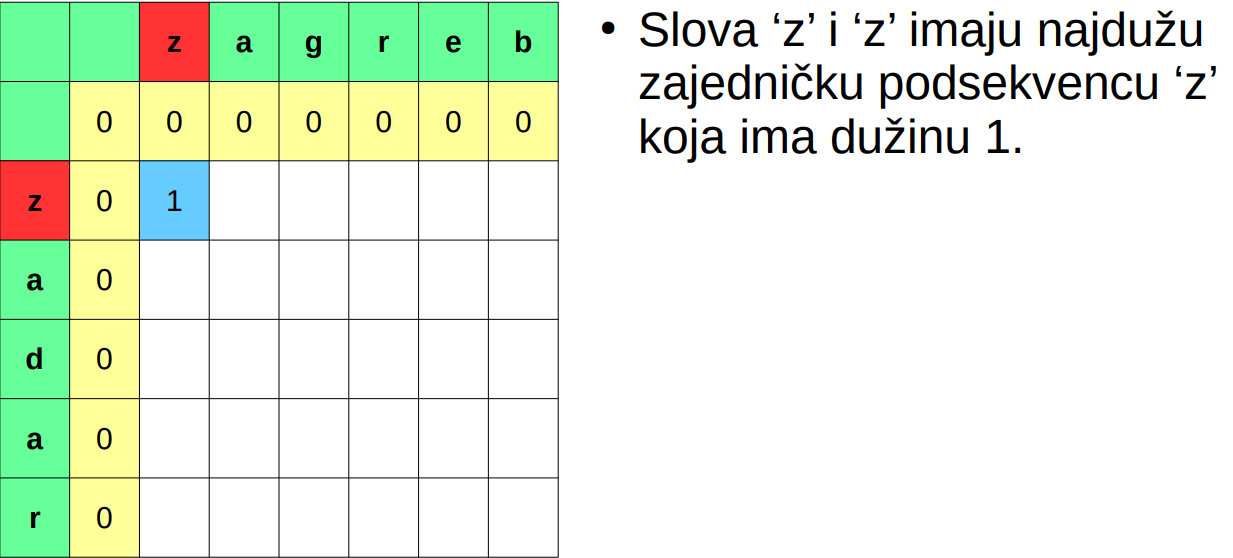
**Napomena 2:**

Kod najduže zajedničke podsekvence niza znakovi ne moraju biti jedan uz drugi, tj. dopušta se preskakanje znakova, a kod najdužeg zajedničkog podstringa niza znakovi moraju biti jedan uz drugi.

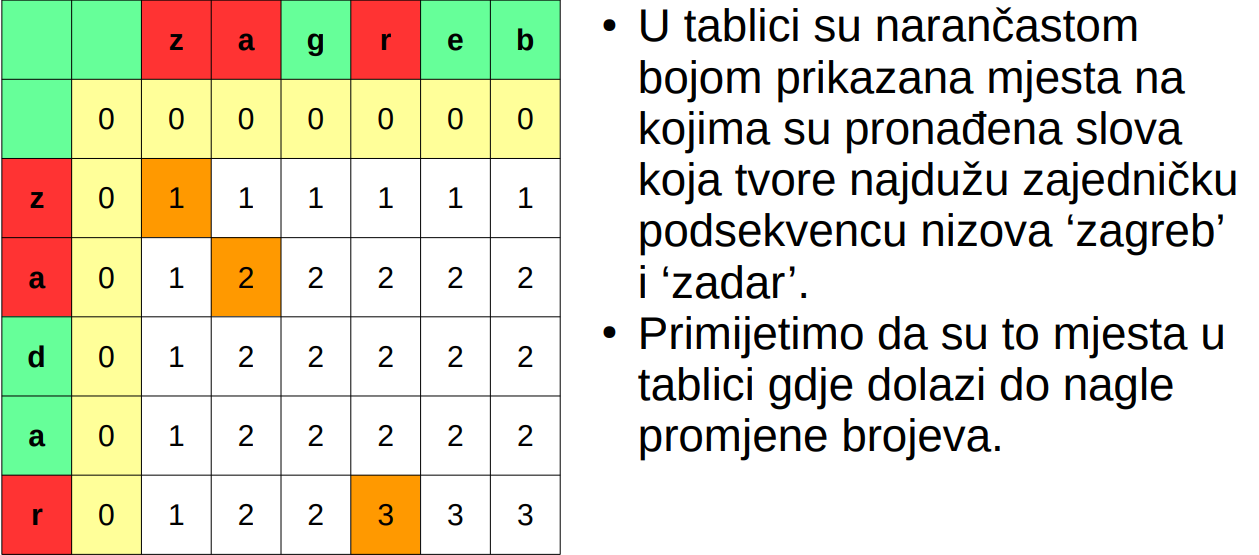
Za nizove ABGHT i ACHTI vrijedi:

* Najduža zajednička podsekvenca: AHT
* Najduži zajednički podstring: HT

**Napomena 3:**

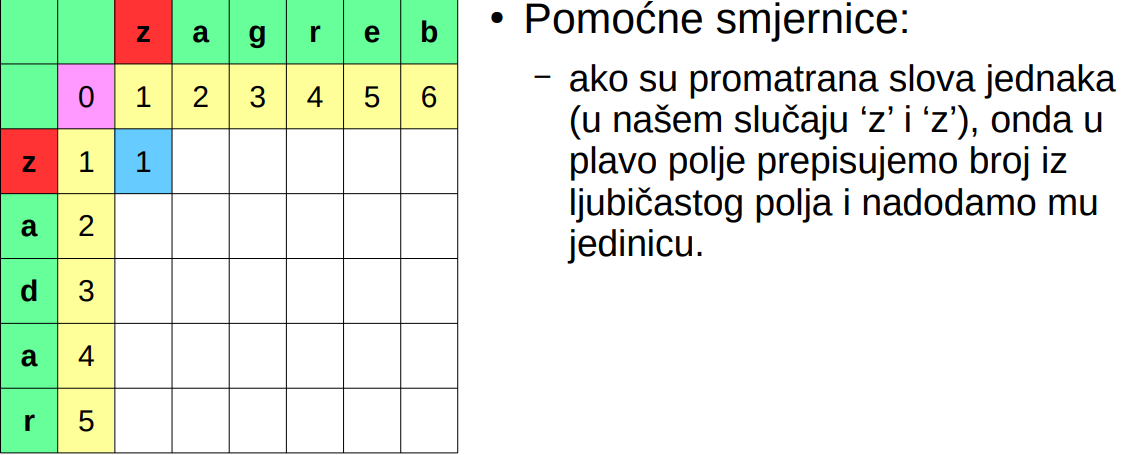


**Napomena 4:**

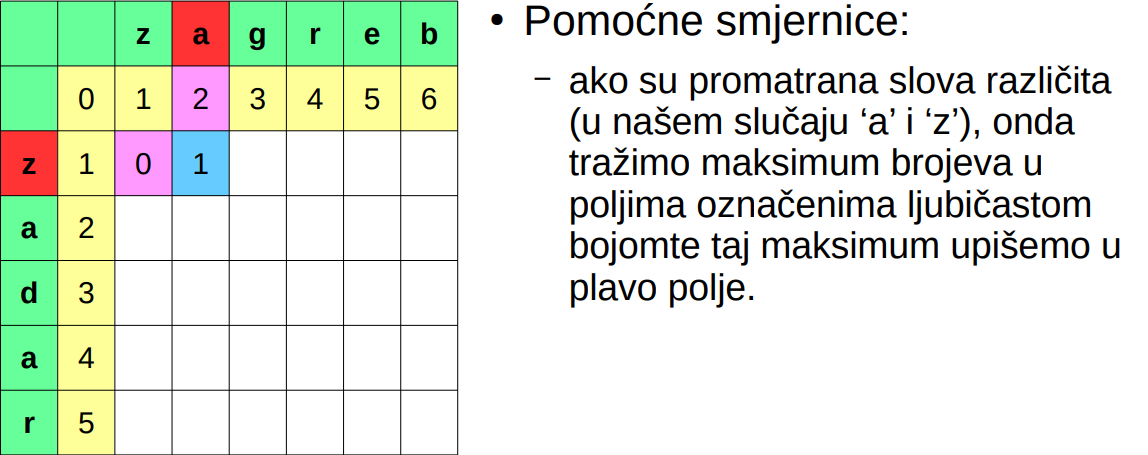


Najveća zajednička podsekvenca je 3 (zar).

**Napomena 5:**



**Napomena 6:**



**Predavanje 4**

1. Što je analiza osjećaja u obradi prirodnog jezika?

**Analiza osjećaja** iz teksta koristi se za svrstavanje teksta u (najčešće) jednu od tri klase: pozitivnu, negativnu ili neutralnu.

2. Što je detekcija emocija u obradi prirodnog jezika?

**Analiza emocija** iz teksta koristi se za pokušaj detekcije ljudskih emocija kao što su ljutnja, gađenje, strah, sreća, tuga i iznenađenje.

3. Za što se sve može koristiti analiza osjećaja i detekcija emocija u obradi prirodnog jezika?

– analiza komentara na društvenim mrežama

– analiza recenzija proizvoda

– analiza različitih anketa

– detekcija neželjene elektroničke pošte

– predviđanje rezultata političkih izbora

– analiza popularnosti javnih osoba

– detekcija lažnih vijesti (eng. fake news)

– detekcija emocionalnog stanja osoba iz poruka na društvenim mrežama

– računanje indeksa sreće (eng. happiness index) raznih zemalja

4. Odaberite najčešće modele emocija koji se koriste u obradi prirodnog jezika.

– **univerzalne emocije** - Paul Ekman;

– **cirkularni model emocija** - James. A. Russell.

5. Odaberite kategoriju kojoj pripada Ekmanov model emocija.

Ekmanov model univerzalnih emocija spada u **diskretne modele** budući da emocije dijeli na 6 (ponekad 7) strogo definiranih klasa koje se ne preklapaju.

6. Odaberite kategoriju kojoj pripada Russellov model emocija.

Russellov **cirkularni model** emocija pretpostavlja da su emocije povezane i da prema tome **ne mogu biti diskretne**.

7. Kojih se 6 emocija iz Ekmanovog modela emocija najčešće koristi u obradi prirodnog jezika?

To su : sreća, tuga, strah, gađenje, ljutnja i iznenađenost.

8. Što predstavlja x, a što y os u Russellovom modelu emocija?

Ovaj model predlaže da su emocije distribuirane u dvodimenzionalnom prostoru gdje **x-os predstavlja ugodu, a y-os aktivaciju**.

9. Koji se dio Russellovog modela emocija koristi kod analize osjećaja u obradi prirodnog jezika?

Koristi se **X os** (neugodno -> ugodno).

10.Koji se dio Russellovog modela emocija koristi kod detekcije emocija u obradi prirodnog jezika?

Koristi **obje osi**, i x i y.

11.Što je SentiWordNet?

**Jedan od rječnika emocija** koji se koristi kod detekcije emocija pomoću ključnih riječi.

12.Što je prepoznavanje ključnih riječi u obradi prirodnog jezika?

**Prepoznavanje ključnih riječi** (eng. keyword recognition) se temelji na izgrađenim rječnicima ili leksikonima emocija u kojima su riječi koje se u njima nalaze povezane sa određenim ključnim riječima koje označavaju različite emocije.

13.Što je leksički afinitet u obradi prirodnog jezika?

**Leksički afinitet** (eng. lexical affinity) uključuje dodjeljivanje određenih vjerojatnosti riječima u rječniku. Te vjerojatnosti kažu koliko je vjerojatno da određena ključna riječ pripada određenoj klasi.

14.Odaberite kategorije na koje se dijeli strojno učenje.

– **nadzirano učenje** (eng. supervised learning),

– **nenadzirano učenje** (eng. unsupervised learning),

– **polunadzirano učenje** (eng. semi-supervised learning),

– **pojačano učenje** (eng. reinforcement learning).

15.Što je prijenosno učenje?

**Prijenosno učenje** (eng. transfer learning) je dio dubokog učenja koji se koristi kada nemamo veliku količinu podataka za treniranje algoritma. Kod prijenosnog **učenja uzima se već istrenirani model** te nadograđuje podacima koje imamo. Prednost prijenosnog učenja je skraćeno vrijeme treniranja algoritma i općenito dobri rezultati.

16.Koje se arhitekture koriste u obradi prirodnog jezika za prijenosno učenje?

**Word2Vec, BERT**.

17.Odaberite trenutne izazove i otvorene probleme u obradi prirodnog jezika?

– nemogućnost jednostavne detekcije negacija u tekstu;

– rečenice koje u sebi imaju riječi vezane uz pozitivne ili negativne tekstove, no same nisu pozitivne ili negativne;

– nedovoljno razumijevanje konteksta;

– nerazumijevanje sarkazma;

– emotikoni (hijeroglifi 21. stoljeća);

– pojava novih riječi (npr. COVID-19);

– većina biblioteka i baza podataka dostupna je samo za engleski jezik;

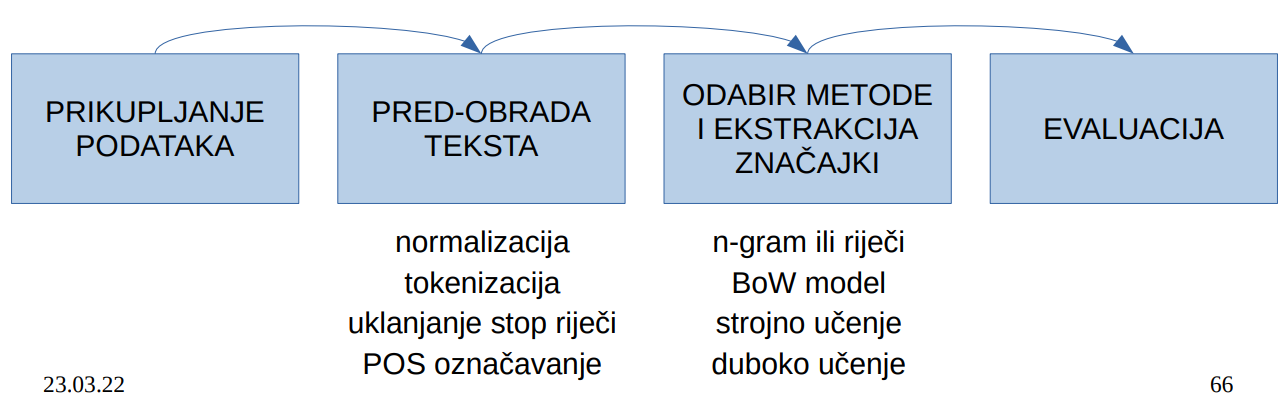
– ne možemo dobro detektirati laži (ne vidimo izraze lica).

18.Što je n-gram?

N-gram je **niz koji se sastoji od N riječi**.

19.Od kojih se koraka sastoje modeli za analizu osjećaja i detekciju emocija u obradi prirodnog jezika?

Modeli za analizu osjećaja i detekciju emocija obično se sastoje od slijedećih koraka:



**Predavanje 5**

1. Koji se algoritmi najčešće koriste za grupiranje (klasteriranje) podataka u obradi prirodnog jezika?

**K-means algoritam, algoritam očekivanja maksimizacije**.

2. Što je regresija?

To je model koji se u statistici i umjetnoj inteligenciji koristi **za predviđanje kontinuiranih vrijednosti** te koji pretpostavlja da postoji linearna veza (kod linearne regresije) između ulaznih varijabli i izlazne varijable.

3. Što je klasifikacija?

**Svrstavanje podataka** u 2 (najčešće) ili više klasa.

4. Koja se vrsta regresije često koristi za klasifikaciju?

**Logistička** regresija.

5. Kako se kod algoritma očekivanja-maksimizacije (EM algoritma) određuju početne vjerojatnosti?

Određuju se **nasumično**.

6. Koje probleme ima originalni algoritam k-sredina?

Budući da se početna središta klastera kod originalnog algoritma k-sredina određuju **nasumično**, može se dogoditi da **krajnja raspodjela podataka po klasterima** bude **loša**.

7. Po čemu je algoritam k-sredina++ bolji od originalnog algoritma k-sredina?

Nastoji **samostalno odrediti optimalna središta** klastera.

8. Što je metoda lakta?

**Metoda za računanje optimalnog broja klastera** (broj k) u koji će se dijeliti podaci kod k-means algoritma.

9. Što je jednojedinično kodiranje (eng. one-hot encoding)?

Stvaranje **vektora veličine rječnika**, koji ima **vrijednost 1** na odgovarajućem mjestu, ako se u nekom tekstu pojavljuje ta riječ iz rječnika, inače 0.

10.Ako je zadan rječnik D, kako se pomoću jednojediničnog kodiranja prikazuje rečenica E? (Napomena: u kolokviju će biti zadan određeni rječnik D i određena rečenica E, npr. D = {danas, je, sunčan, dan, prema, tome, idemo, na, marjan} i E = “danas je sunčan dan”.)

Kreiramo inicijalni vektor veličine rječnika D : [ 0 0 0 0 0 0 0 0 0 ];

Rečenica E će imati vektor oblika : [ 1 1 1 1 0 0 0 0 0 ] jer su prve 4 riječi iz rječnika.

**Predavanje 6**

1. Koja je razlika između generativnih i diskriminativnih modela?

**Generativni modeli** (eng. generative models) iz ulaznih podataka uče **zajedničke vjerojatnosti**, a **diskriminativni modeli** (eng. discriminative models) iz ulaznih podataka uče **uvjetne vjerojatnosti**.

2. Na što se odnosi riječ “naivni” kod naivnog Bayesovog klasifikatora?

Riječ “naivni” u imenu ovoga klasifikatora dolazi od naivne **pretpostavke da su značajke ili podaci koji se koriste prilikom klasifikacije neovisni** jedni o drugima.

3. Što je pretpostavka modela zbirke riječi?

Pretpostavka modela zbirke riječi (eng. Bag-of-Words assumption) je da **redoslijed riječi u dokumentu nije važan**.

4. Što je pretpostavka naivnog Bayesovog modela?

Pretpostavka naivnog Bayesovog modela ili uvjetna neovisnost je pretpostavka po kojoj je vjerojatnost pojavljivanja određene riječi u dokumentu koji pripada klasi c **neovisna** o vjerojatnosti pojavljivanja neke druge riječi u tome dokumentu.

5. Na kojim se pretpostavkama naivni Bayesov model temelji?

- **Pretpostavka 1**: redoslijed riječi u dokumentu nije važan. Ova se pretpostavka naziva **pretpostavka modela zbirke riječi** (eng. Bag-of-Words assumption).

Ova je pretpostavka pogrešna. Na primjer, nije svejedno piše li u dokumentu “you can write” ili “can you write”.

- **Pretpostavka 2**: vjerojatnost pojavljivanja određene riječi u dokumentu koji pripada klasi c neovisna je o vjerojatnosti pojavljivanja neke druge riječi u tome dokumentu. Ova se pretpostavka naziva **pretpostavka naivnog Bayesovog modela** (eng. naive Bayes assumption) ili uvjetna neovisnost (eng. class conditional independence assumption).

Ova je pretpostavka pogrešna. Na primjer, riječi poput “Sabor”, “država”, “plaće” i “zakon” se često pojavljuju zajedno u dokumentima koji pripadaju klasi “politika”, te nisu neovisne jedna o drugoj.

Iako su obje ove pretpostavke pogrešne, omogućuju pojednostavljeno i brzo računanje i svejedno mogu dovesti do dovoljno dobrih rezultata klasifikacije.

6. Po čemu se razlikuju multivarijatni i multinominalni naivni Bayesov klasifikator?

**Multivarijatni** zanemaruje višestruka pojavljivanja iste riječi u istom dokumentu (gleda samo jedinstvene riječi), a **Multinominalni** ne zanemaruje višestruka pojavljivanja iste riječi u istom dokumentu (gleda brojčane vrijednosti svih jedinstvenih riječi).

7. Što je problem nulte vjerojatnosti?

To je problem kod kojega se u dokumentu pojavljuje riječ koja nije u rječniku nastalom iz podataka za treniranje pa je vjerojatnost pojavljivanja te riječi za bilo koju klasu 0, a onda je i vjerojatnost svake klase 0.

8. Kako se problem nulte vjerojatnosti rješava?

Rješava se **Laplaceovim** **zaglađivanjem**.

9. Koje vrste problema se mogu riješiti pomoću skrivenih Markovljevih modela?

**Skriveni Markovljevi modeli** (eng. Hidden Markov Models, HMM) su probabilistički grafički modeli koji se koriste **za predviđanje sekvencijalnih** **podataka**.

**Dekodiranje, učenje, evaluacija**.

10.Za što se koriste skriveni Markovljevi modeli u obradi prirodnog jezika?

U obradi prirodnog jezika HMM modeli koriste se za **označavanje vrsta riječi u rečenici** (tzv. Part-of-Speech tagging, POS tagging, POST).

11.Što je Markovljeva pretpostavka?

**Markovljeva pretpostavka** kaže da prilikom određivanja svakog slijedećeg stanja gledamo samo trenutno stanje, a ne i prethodna stanja.

12.Što je Viterbi algoritam?

Algoritam koji se koristi kod efikasnog **rješavanja Markovljevih modela**.

13.Što su skrivena stanja kod skrivenih Markovljevih modela?

Ovo su **stanja koja otkrivamo indirektno**, tj. stanja koja nisu direktno opažena.

14. Kod označavanja vrsta riječi u obradi prirodnog jezika, što bi sadržavala transmisijska a što emisijska tablica?

**Transmisijska** bi sadržavala vjerojatnosti koje kažu koliko je vjerojatno da se određena vrsta riječi pojavi iza neke druge vrste riječi (npr. koliko je vjerojatno da se imenica pojavi iza glagola), a **emisijska** prikazuje vjerojatnosti koje kažu koliko je vjerojatno da jedinstvene riječi iz naših dokumenata pripadaju određenoj vrsti riječi (npr. koliko je vjerojatno da riječ pool pripada vrsti riječi imenica).

15.Što je “logit” kod logističke regresije?

To je ovisna **varijabla** koju predviđamo (eng. logistic unit).

16.Što je svrha algoritma gradijentnog spuštanja?

**Traženje optimalnih parametara (težina)**, odnosno algoritam gradijentnog spuštanja ili spust gradijenta je metoda numeričke optimizacije koja nastoji minimizirati neku funkciju (npr. cijenu ili vrijeme).

**Predavanje 7**

1. Što je težina riječi (eng. term weight)?

**Težina riječi** (eng. term weight) je broj koji kaže koliko je određena riječ važna prilikom pretraživanja tekstualnih dokumenata.

2. Što je frekvencija riječi (eng. term frequency)?

**Frekvencija riječi** (eng. term frequency, TF) je frekvencija s kojom se određena riječ pojavljuje u tekstualnom dokumentu.

3. Što je frekvencija dokumenta (eng. document frequency)?

**Frekvencija dokumenta** (eng. document frequency, DF) je broj koji kaže u koliko se promatranih tekstualnih dokumenata pojavljuje određena riječ.

4. Što je inverzna frekvencija dokumenta (eng. inverse document frequency)?

**Inverzna frekvencija dokumenta** (eng. inverse document frequency, IDF) se može definirati kao logaritam omjera između sveukupnog broja tekstualnih dokumenata koje promatramo (N) te broja koji kaže u koliko se tih dokumenata pojavljuje tražena riječ (df (t, D)).

5. Što je TF-IDF?

**TF-IDF** (eng. term frequency-inverse document frequency) je statistička mjera koja kombinira frekvenciju riječi (TF) i inverznu frekvenciju dokumenta (IDF). Pomoću TF-IDF mjere može se odrediti koliko je određena riječ važna u određenom dokumentu. Glavni cilj TF-IDF metode je dati veću težinu riječima koje su diskriminantne, tj. specifične za određenu temu.

6. Odaberite načine ugrađivanja riječi (eng. word embedding) koji su utemeljeni na neuralnim mrežama.

To su **Word2vec i BERT**.

7. Odaberite načine ugrađivanja riječi (eng. word embedding) koji nisu utemeljeni na neuralnim mrežama.

Die sind **GloVe, fastText, Gensim**, …

8. Koji model ugrađivanja riječi bolje modelira kontekst: BERT ili Word2vec?

**BERT**.

9. Što je CBOW (eng. Continuous Bag-of-Words) model?

To je jedan od **arhitektura umjetnih neuralnih mreža** koji se koristi kod Word2vec modela, a koji na temelju konteksta predviđaju riječ u rečenici.

10. Što je Skip-gram model?

To je jedan od **arhitektura umjetnih neuralnih mreža** koji se koristi kod Word2vec modela, a koji na temelju riječi predviđaju njezin kontekst.

11. Na kakvoj se arhitekturi neuralnih mreža temelji BERT model?

Na **Transformers** arhitekturi neuralnih mreža.