

UNIVERSITY OF LJUBLJANA
FACULTY OF COMPUTER AND INFORMATION SCIENCE

Romana Grilj

**Performance comparison of open
source and commercial information
extraction tools**

MASTER'S THESIS
THE 2ND CYCLE MASTER'S STUDY PROGRAMME
COMPUTER AND INFORMATION SCIENCE

SUPERVISOR: doc. dr. Slavko Žitnik
CO-SUPERVISOR: akad. prof. dr. Martin Krpan

Ljubljana, 2023

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Romana Grilj

**Primerjava uspešnosti odprtokodnih
in komercialnih orodij za luščenje
podatkov**

MAGISTRSKO DELO
MAGISTRSKI ŠTUDIJSKI PROGRAM DRUGE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

MENTOR: doc. dr. Slavko Žitnik
SOMENTOR: akad. prof. dr. Martin Krpan

Ljubljana, 2023

To delo je ponujeno pod licenco *Creative Commons Priznanje avtorstva-Deljenje pod enakimi pogoji 2.5 Slovenija* (ali novejšo različico). To pomeni, da se tako besedilo, slike, grafi in druge sestavine dela kot tudi rezultati diplomskega dela lahko prosto distribuirajo, reproducirajo, uporabljajo, priobčujejo javnosti in predelujejo, pod pogojem, da se jasno in vidno navede avtorja in naslov tega dela in da se v primeru spremembe, preoblikovanja ali uporabe tega dela v svojem delu, lahko distribuira predelava le pod licenco, ki je enaka tej. Podrobnosti licence so dostopne na spletni strani creativecommons.si ali na Inštitutu za intelektualno lastnino, Streliška 1, 1000 Ljubljana.



Izvorna koda diplomskega dela, njeni rezultati in v ta namen razvita programska oprema je ponujena pod licenco GNU General Public License, različica 3 (ali novejša). To pomeni, da se lahko prosto distribuira in/ali predeluje pod njenimi pogoji. Podrobnosti licence so dostopne na spletni strani <http://www.gnu.org/licenses/>.

ACKNOWLEDGMENTS

Worth mentioning in the acknowledgment is everyone who contributed to your thesis.

Romana Grilj, 2023

To all the flowers of this world.

*"The only reason for time is so that
everything doesn't happen at once."*

— Albert Einstein

Contents

List of used acronmys

acronym	meaning
CA	classification accuracy
DBMS	database management system
SVM	support vector machine
...	...

Abstract

Title: Performance comparison of open source and commercial information extraction tools

This sample document presents an approach to typesetting your BSc thesis using L^AT_EX. A proper abstract should contain around 100 words which makes this one way too short. A good abstract contains: (1) a short description of the tackled problem, (2) a short description of your approach to solving the problem, and (3) (the most successful) result or contribution in your thesis.

Keywords

Data analysis, Information Retrieval, structural data, Web Mining

Povzetek

Naslov: Primerjava uspešnosti odprtokodnih in komercialnih orodij za luščenje podatkov

V vzorcu je predstavljen postopek priprave magistrskega dela z uporabo okolja L^AT_EX. Vaš povzetek mora sicer vsebovati približno 100 besed, ta tukaj je odločno prekratek. Dober povzetek vključuje: (1) kratek opis obravnavanega problema, (2) kratek opis vašega pristopa za reševanje tega problema in (3) (najbolj uspešen) rezultat ali prispevek magistrske naloge.

Ključne besede

analiza podatkov, ekstrakcija podatkov, strukturni podatki, spletno rudarjenje

Chapter 1

Uvod

Datoteka `magistrska_naloga.tex` na kratko opisuje, kako se pisanja magistrskega dela lotimo z uporabo programskega pateka `LATEX`. V tem dokumentu bomo predstavili nekaj njegovih prednosti in hib. Kar se slednjih tiče, mi pride na misel ena sama. Ko se srečamo z njim, nam izgleda kot kislo jabolko, nismo prepričani, da bi želeli vanj ugrizniti. Lahko pa z njim pripravimo odličen zavitek ali pa pridemo na okus.

V Poglavju ?? bomo na hitro spoznali besedilne konstrukte kot so izreki, enačbe in dokazi. Naučili se bomo, kako se na njih sklicujemo. V Poglavju ?? se bomo srečali s sklicevanjem na besedilne konstrukte. Poglavje ?? bo predstavilo vključevanje plovk: slik in tabel. V Poglavju ?? se bomo srečali s sklicevanjem na literaturo. Sledil bo samo še zaključek.

Bodite pozorni, da se v glavni mapi nahajata še datoteki `declaration.tex` in `izjava.tex`. Ti datoteki se ločeno prevedeta, ju podpišete in oddate v referat ločeno od magistrske naloge.

Chapter 2

Sklicevanje na besedilne konstrukte

Matematična ali popolna indukcija je eno prvih orodij, ki jih spoznamo za dokazovanje trditev pri matematičnih predmetih.

Izrek 2.1 *Za vsako naravno število n velja*

$$n < 2^n. \tag{2.1}$$

Dokaz. Dokazovanje z indukcijo zahteva, da neenakost (??) najprej preverimo za najmanjše naravno število — 0. Res, ker je $0 < 1 = 2^0$, je neenačba (??) za $n = 0$ izpolnjena.

Sledi indukcijski korak. S predpostavko, da je neenakost (??) veljavna pri nekem naravnem številu n , je potrebno pokazati, da je ista neenakost v veljavi tudi pri njegovem nasledniku — naravnem številu $n + 1$. Izračun zapišemo s tremi vrsticami, ki jih končamo s piko, saj do del tega stavka:

$$n + 1 < 2^n + 1, \tag{2.2}$$

$$\leq 2^n + 2^n, \tag{2.3}$$

$$= 2^{n+1}.$$

Neenakost (??) je posledica indukcijske predpostavke, neenakost (??) pa enostavno dejstvo, da je za vsako naravno število n izraz 2^n vsaj tako velik kot 1. S tem je dokaz Izreka ?? zaključen. \square

Opazimo, da je \LaTeX številko izreka podredil številki poglavja.

Chapter 3

Plovke: slike in tabele

Slike in daljše tabele praviloma vključujemo v dokument kot plovke. Pozicija plovke v končnem izdelku ni pogojena s tekom besedila, temveč z izgledom strani. \LaTeX bo skušal plovko postaviti samostojno, praviloma na vrh strani, na kateri se na takšno plovko prvič sklicujemo. Pri tem pa bo na vsako stran končnega izdelka želel postaviti tudi sorazmerno velik del besedila. V skrajnem primeru, če imamo res preveč plovk, se bo odločil za stran popolnoma zapolnjeno s plovkami.

3.1 Formati slik

Bitne slike, vektorske slike, kakršnekoli slike, z \LaTeX om lahko vključimo vse. Slika ?? je v `.pdf` formatu. Pa res lahko vključimo slike katerihkoli formatov? Žal ne. Programski paket \LaTeX lahko uporabljamo v več dialektih. Ukaz `latex` ne mara vključenih slik v formatu Portable Document Format `.pdf`, ukaz `pdflatex` pa ne prebavi slik v Encapsulated Postscript Formatu `.eps`. Strnjeno v Tabeli ??.

Nasvet? Odločite se za uporabo ukaza `pdflatex`. Vaš izdelek bo brez vmesnih stopenj na voljo v `.pdf` formatu in ga lahko odnesete v vsako tiskarno. Če morate na vsak način vključiti sliko, ki jo imate v `.eps` formatu, jo vnaprej pretvorite v alternativni format, denimo `.pdf`.



Figure 3.1: Herschelov graf, vektorska grafika.

Table 3.1

ukaz/format	.pdf	.eps	ostali formati
pdflatex	da	ne	da
latex	ne	da	da

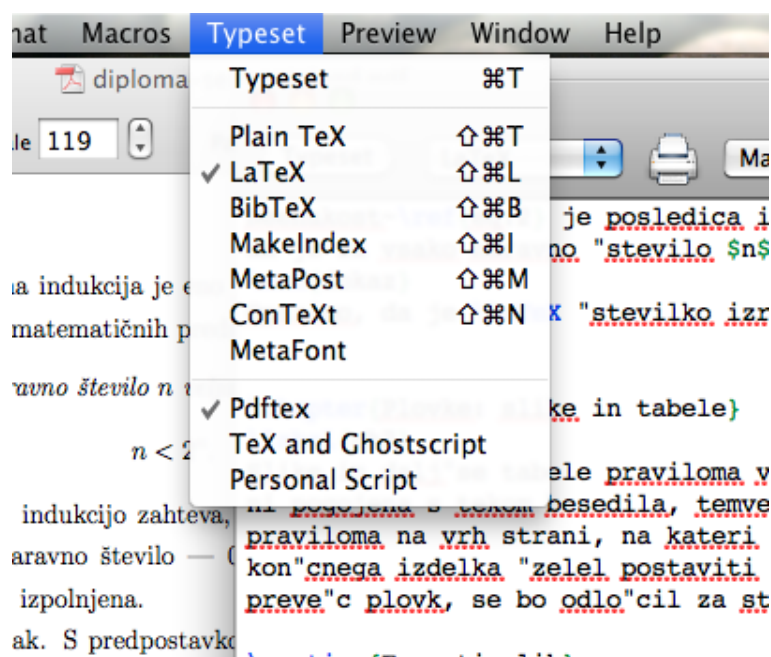


Figure 3.2: Kateri dialekt uporabljati?

Včasih se da v okolju za uporabo programskega paketa \LaTeX nastaviti na kakšen način bomo prebavljali vhodne dokumente. Spustni meni na Sliki ?? odkriva uporabo \LaTeX a v njegovi pdf inkarnaciji — `pdflatex`. Vključena Slika ?? je seveda bitna.

Chapter 4

Razno

4.1 Notacije

Za notacijo spremenljivk ter skalarjev uporabimo običajno notacijo, t.j., spremenljivka x in skalar a . Pri notaciji matrik ter vektorjev pa se poslužujemo krepega fonta. Torej, matrika \mathbf{A} ter vektor \mathbf{v} ,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & & & \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix}.$$

4.2 Lepe tabele in psevdokoda

Psevdokoda ?? prikazuje primer delovanja genetskega algoritma, medtem ko Tabela ?? prikazuje primer lepe tabele brez vertikalnih črt.

Table 4.1: Primer enostavne tabele.

Ime	Vrednost	Opis
a	0.03	skalar
x	-1	spremenljivka

Algorithm 1 Psevdokoda genetskega algoritma

```
1:  $t \leftarrow 0$ 
2:  $InitPopulation[P(t)] \leftarrow$  inicializiraj populacijo
3:  $EvalPopulation[P(t)] \leftarrow$  evaluiraj populacijo
4: repeat
5:    $P'(t) \leftarrow Variation[P(t)] \leftarrow$  generiraj novo populacijo
6:    $EvalPopulation[P'(t)] \leftarrow$  evaluiraj novo populacijo
7:    $P(t+1) \leftarrow ApplyGeneticOperators[P'(t) \in Q]$ 
8:    $t \leftarrow t + 1$ 
9: until prekinitev
10: if rezultat dovolj dober then
11:   shrani rezultat
12: end if
```

Chapter 5

Kaj pa literatura?

Kot smo omenili že v uvodu, je pravi način za citiranje literature uporaba `BIBTEXa` [?]. Programski paket `LATEX` je prvotno predstavljen v priročniku [?] in je v resnici nadgradnja sistema `TEX` avtorja Donalda Knutha, znanega po denimo, če izpustim njegovo umetnost programiranja, Knuth-Bendixovem algoritmu [?].

Vsem raziskovalcem s področja računalništva pa svetujem v branje mnenje L. Fortnowa [?].

Chapter 6

Opis funkcionalnosti

6.1 Prepoznavanje imenskih entitet

Prepoznavanje imenskih entitet je tehnika na področju obdelave naravnega jezika, ki se uporablja za prepoznavanje in klasifikacijo besed v besedilu. Te posebne vrste so imenovane entitete, kot so imena oseb, organizacij, lokacij, datumov, števil, denarnih zneskov in drugih specifičnih poimenovanj.

Cilj je prepoznati in določiti začetek in konec posameznih entitet v besedilu ter jim pripisati ustrezno kategorijo. Na primer, v stavek "Janez Novak je rojen 10. avgusta 1985 v Ljubljani" bi sistem prepoznal "Janez Novak" kot ime osebe, "10. avgust 1985" kot datum in "Ljubljana" kot lokacijo.

Številne praktične uporabe:

1. Avtomatsko označevanje imenskih entitet v novicah, člankih in drugih besedilnih vsebinah.
2. Razumevanje strukture in vsebine dokumentov za informacijsko iskanje in kategorizacijo.
3. Pomoč pri analizi sentimenta, kjer se želimo razumeti, kako se osebe, organizacije ali druge entitete omenjene v besedilu nanašajo na določeno temo ali izdelek.

6.2 Analiza sentimenta

Analiza sentimenta je proces določanja čustvenega odziva, nagnjenosti ali stališča zapisanega besedila. Cilj analize sentimenta je ugotoviti, ali je določeno besedilo pozitivno, negativno ali nevtrarno. To je lahko koristno pri analizi mnenj strank, razumevanju čustvenega odziva na izdelke, blagovne znamke, dogodke in druge. Na primer, če imamo naslednji stavek: "Ta film je fantastičen, vreden ogleda!", bi analiza sentimenta prepoznala, da je izraz pozitiven. Ta analiza temelji na uporabi naravnojezikovnega procesiranja in strojnega učenja. Obstaja več pristopov k analizi sentimenta, vključno z naslednjimi:

1. **Pravilni pristopi:** Uporabljajo se predvsem pravila in vzorci za identifikacijo pozitivnih in negativnih izrazov v besedilu. Na primer, besede, kot so "dobro", "fantastično", "radostno" itd., bi bile označene kot pozitivne, medtem ko bi bile besede, kot so "slabo", "žalostno", "neznosno" itd., označene kot negativne.
2. **Strojno učenje na podlagi besedila:** Ta pristop vključuje uporabo algoritmov strojnega učenja, ki so naučeni prepoznati čustveni naboj besed v besedilu na podlagi velikega števila označenih podatkov (besedil s čustvenimi oznakami).
3. **Analiza sentimenta s čustvenimi slovarji:** Ta pristop vključuje uporabo slovarjev z besedami in izrazoslovjem, ki so povezani z določenimi čustvi. Besedilo se nato preveri in oceni glede na prisotnost pozitivnih ali negativnih besed iz čustvenih slovarjev.
4. **Algoritmi globokega učenja:** V zadnjem času so se pojavili tudi pristopi, ki temeljijo na globokem učenju.

6.3 Povzemanje besedila

Pri povzemanju besedila gre za postopek ustvarjanja krajšega in jedrnatega povzetka izdaljšega besedila, kot je članek ali dokument. Namen povzemanja je izluščiti ključne informacije in ideje iz izvirnega besedila ter jih predstaviti na bolj pregleden in krajši način. To je zelo koristno v velikih količinah podatkov, ko želimo hitro pridobiti bistvo informacij, ne da bi brali celotno besedilo.

NLP tehnike za povzemanje uporabljajo različne algoritme in metode, ki vključujejo strojno učenje in obdelavo naravnega jezika, da bi učinkovito izluščile ključne besede, stavke ali odstavke, ki predstavljajo osrednje ideje v izvirnem besedilu. Rezultat je običajno kratek povzetek, ki ohranja pomembne informacije iz izvirnega besedila. Ta tehnologija ima širok spekter uporab, kot so samodejno povzemanje novic, generiranje opisov izdelkov, izdelava povzetkov raziskovalnih člankov in še veliko več.

6.4 Izvleček besedne zveze

Nanaša se na besede ali izraze, ki so najpomembnejši ali najbolj značilni za določeno besedilo ali dokument. Te besede so običajno tiste, ki nosijo ključne informacije ali so bistvene za razumevanje vsebine.

Identifikacija ključnih besed je pomembna naloga, saj nam omogoča, da hitro ugotovimo, o čem govori določen tekst. Te besede so lahko uporabne tudi za avtomatsko indeksiranje dokumentov, iskanje relevantnih informacij in razumevanje teme besedila brez potrebe po branju celotnega besedila.

6.5 Klasifikacija besedila

Klasifikacija besedil je tehnika, pri kateri avtomatizirano določimo kategorijo ali razred določenega besedila na podlagi vsebine besedila. To je lahko zelo uporabno, saj nam omogoča razvrščanje besedil v različne skupine glede na njihovo vsebino. Na primer, lahko klasificiramo e-poštna sporočila kot

”spam” ali ”ne-spam”, novice glede na tematiko, uporabniške komentarje glede na ton (pozitiven, negativen, nevtralen), itd.

Postopek klasifikacije besedil se običajno začne s pripravo in čiščenjem besedil. To vključuje odstranjevanje nepotrebnih znakov, šumnikov, posebnih znakov, pretvorbo vseh črk v male črke, lahko pa tudi odstranjevanje pogostih besed, ki nimajo velikega pomena za klasifikacijo (npr. ”in”, ”ali”, ”je”, ”na”, ”s”, itd.).

Nato se besedila predstavijo v obliki, ki jo lahko uporabimo za učenje modela. Pogosto se uporablja metoda imenovana vreča besed (”Bag-of-Words”), kjer se besedilo pretvori v nabor besed, ki se pojavljajo v njem, in število pojavitev teh besed. Ta postopek lahko ponazorimo s pomočjo vektorja.

6.6 Zaznava objektov

Je tehnika, ki se uporablja za avtomatsko zaznavanje in identifikacijo objektov na digitalnih slikah ali video posnetkih. Namen te tehnike je, da prepozna in označi različne objekte v podobi ter jih loči od ozadja ali drugih objektov.

Postopek objektnega zaznavanja običajno vključuje naslednje korake:

1. Zaznavanje: Model preučuje sliko ali video posnetek in identificira regije, kjer bi se lahko nahajali objekti.
2. Lokalizacija: Po tem, ko so bile regije prepoznane, algoritem določi omejitveno okvirje (bounding boxes), ki natančno označujejo položaje in mejne točke objektov na sliki.
3. Klasifikacija: Ko so objekti omejeni z omejitvenimi okviri, analizira vsebino znotraj teh okvirov in jih razvrsti v različne kategorije (npr. avto, pes, zgradba, itd.).
4. Sledenje: V video posnetkih je lahko zaželeno, da algoritem sledi objektom skozi različne kadre in tako beleži njihovo gibanje.

Objektno zaznavanje se uporablja v številnih aplikacijah, kot so samovozeča vozila za zaznavanje drugih vozil in pešcev, nadzorne kamere za varnostne namene, prepoznavanje obrazov, identifikacija prometnih znakov, analiza medicinskih slik in še veliko drugega. Gre za enega ključnih elementov umetne inteligence.

Chapter 7

Dataseti

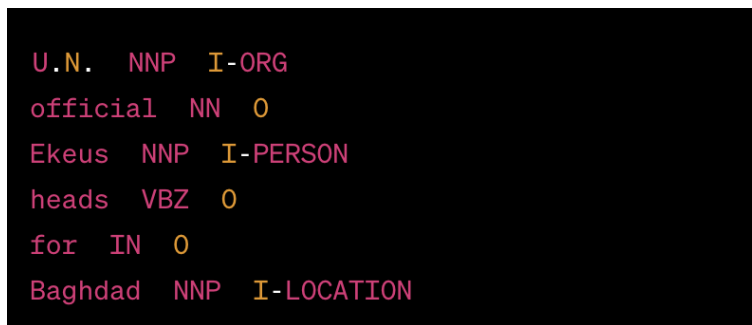
7.1 Kaj je dataset?

Dataset je zbirka podatkov, ki so organizirani in shranjeni v strukturirani ali ne-strukturirani obliki ter označeni za namen analize, raziskav, učenja modelov ali drugih postopkov obdelave podatkov. Dataseti vsebujejo različne vrste podatkov, od števil, besedil, slik, zvokov, videoposnetkov do drugih tipov informacij. V kontekstu računalniškega znanstvenega modeliranja in strojnega učenja so dataseti ključnega pomena, saj služijo kot osnova za razvoj, treniranje in evalvacijo modelov. Modeli se učijo na teh podatkih, tako da prepoznajo vzorce in povezave med vhodnimi podatki in ciljnim izhodi. Na primer, v naravnojezikovni obdelavi dataset vsebuje besedilne podatke, ki so lahko članki, knjige, novinarski članki ali socialni mediji. Te podatke lahko uporabimo za različne naloge, kot je sentimentna analiza, klasifikacija tem, generiranje besedil itd. V poslovnem okolju se dataseti uporabljajo za analizo strank, trženjske kampanje, obdelavo naravnega jezika v storitvah za stranke in še veliko drugih aplikacij. Pomembno je, da so dataseti pravilno pripravljeni, imajo ustrezne metapodatke in so primerni za ciljno nalogo, da bi omogočili kakovostno analizo in doseganje uporabnih rezultatov.

7.2 Uporabljeni dataseti

7.2.1 CoNLL 2003

Je zbirka podatkov, ki se uporablja za razvoj in evalvacijo sistemov za obdelavo naravnega jezika, zlasti za nalogo imenovanja imenovalnih entitet. Imenuje se po konferenci CoNLL (Conference on Computational Natural Language Learning) leta 2003, kjer je bil ta nabor podatkov predstavljen v okviru tekmovanja za prepoznavanje imenovanih entitet. Dataset CoNLL 2003 je priljubljen referenčni dataset za prepoznavanje poimenovanih entitet naravnega jezika v obdelavi naravnega jezika. Uporabljen je bil v skupni nalogi na konferenci o računalniškem učenju naravnega jezika (CoNLL) leta 2003.



```
U.N. NNP I-ORG  
official NN O  
Ekeus NNP I-PERSON  
heads VBZ O  
for IN O  
Baghdad NNP I-LOCATION
```

Figure 7.1: CoNLL2003 dataset

Poimenovane entitete so razdeljene v štiri glavne kategorije:

1. Oseba (PER): Posamezna imena ljudi.
2. Organizacija (ORG): Imena podjetij, ustanov ali organizacij.
3. Lokacija (LOC): Imena geografskih lokacij, kot so mesta, države ali regije.
4. Razno (MISC): Druge poimenovane entitete, ki ne spadajo v zgoraj navedene kategorije, na primer datumi, odstotki ali denar.

Podatki v datasetu so predstavljeni v obliki ene besede na vrstico, kjer vsaka vrstica predstavlja besedo in pripadajočo oznako v stavku. Besede in oznake so ločene z belim prostorom. Dataset CoNLL 2003 se pogosto uporablja za evalvacijo zmogljivosti modelov za prepoznavanje poimenovanih entitet in že več let je standardni benchmark za raziskovalce in strokovnjake v skupnosti obdelave naravnega jezika. Ostaja dragocen vir za razvoj in preizkušanje novih algoritmov in sistemov za NER. Dataset je sestavljen/razdeljen na tri različne skupine in sicer: CoNLL2003 podatkovna zbirka je običajno razdeljena na tri sklope:

1. učni (train) z 14.000 vrsticami primerov
2. validacijski (validation) z 3.250 vrsticami primerov
3. preizkusni (test) z 3.450 vrsticami primerov

7.2.2 IMDb Reviews dataset

IMDB podatkovna zbirka, znana tudi kot IMDB Movie Reviews Dataset, je priljubljen benchmark podatkovni niz v področju obdelave naravnega jezika. Ta niz je sestavljen iz pregledov filmov, ki so jih prispevali uporabniki na spletni strani IMDb (Internet Movie Database).

Podatki vsebujejo ocene in besedilne komentarje, ki jih je ustvarila skupnost uporabnikov IMDb. Vsak pregled vsebuje besedilni komentar in oceno filma, ki se giblje med 1 (najslabša) in 10 (najboljša). Cilj te podatkovne zbirke v naravnem jeziku je razviti modele, ki lahko avtomatsko analizirajo besedilne komentarje in napovedo, ali je pregled pozitiven ali negativen glede na oceno in besedilo. IMDb podatkovna zbirka je običajno razdeljena na dva sklopa:

1. učni (train) z 25.000 vrsticami primerov
2. preizkusni (test) z 25.000 vrsticami primerov

Vsak sklop vsebuje tisoče pregledov filmov. To je idealna podatkovna zbirka za naloge analize čustvenega tona besedil (sentiment analysis), kjer modeli ocenjujejo, ali je mnenje v besedilu pozitivno, negativno ali nevtrarno.

7.2.3 COCO dataset

COCO (Common Objects in Context) je široko uporabljen nabor podatkov v področju računalniškega vida in detekcije objektov. Namenjen je zagotavljanju celovite in raznolike zbirke slik za različne naloge, vključno z detekcijo objektov, segmentacijo in podnaslavljanjem. Nabor podatkov naj bi odražal scenarije iz resničnega sveta in vsebuje slike, ki so kompleksne ter vključujejo več objektov v različnih kontekstih.

Nabor podatkov COCO je obsežen in vsebuje deset tisoče slik z milijoni označenih posameznih objektov. Slike prihajajo iz različnih virov, zajemajo raznolike prizore, ozadja, svetlobne pogoje in velikosti objektov.

Tukaj je nekaj ključnih značilnosti nabora podatkov COCO:

1. Kategorije slik: Nabor podatkov COCO vsebuje slike, ki zajemajo 80 različnih kategorij objektov, od splošnih objektov, kot so "oseba," "avto" in "pes," do bolj specifičnih objektov, kot so "mobilni telefon," "zobna ščetka" in "zmaj."
2. Anotacije: Vsaka slika v naboru podatkov COCO je opremljena z oznakami na ravni objekta in koordinatami okvirja. To pomeni, da je vsak posamezen objekt določene kategorije znotraj slike označen, okoli njega pa je narisano območje z okvirjem, ki označuje njegovo lokacijo. Informacije o anotacijah so ključnega pomena za usposabljanje modelov za detekcijo objektov in segmentacijo.
3. Segmentacija objektov: Poleg anotacij območja z okviri nabor podatkov COCO prav tako zagotavlja maske segmentacije na ravni slikovnih pik za vsak posamezen objekt. To pomeni, da so objekti ne le lokalizirani z okviri, ampak so natančno določene tudi meje objektov na ravni slikovnih pik.

4. Izzivi in tekmovanja: Nabor podatkov COCO je spodbudil številne izzive in tekmovanja v skupnosti računalniškega vida. COCO izziv je priljubljen dogodek, na katerem raziskovalci in inženirji predstavijo svoje modele za detekcijo objektov, segmentacijo in podnaslavljanje, s čimer premikajo meje tega, kar je mogoče v teh področjih.

COCO podatkovna zbirka je običajno razdeljena na dva sklopa:

1. učni (train) z 117.000 primeri
2. preizkusni (test) z 4.950 primeri

7.2.4 CNN/Daily Mail Dataset

CNN/Daily Mail je zbirka novičarskih člankov skupaj s povzetki, ki se uporablja za usposabljanje in preizkušanje modelov za avstraktivno povzemanje besedil. Ta nabor podatkov vsebuje različne novičarske članke in njihove povzetke, zaradi česar je primeren za naloge avstraktivnega povzemanja, kjer se ustvarijo povzetki v lastnih besedah, ne le izbirajo stavke iz izvirnega besedila. Nabor podatkov vsebuje na tisoče člankov s pripadajočimi povzetki, kar omogoča raziskovalcem obsežno usposabljanje in preizkušanje modelov. Tukaj je nekaj ključnih značilnosti nabora podatkov CNN/Daily Mail:

1. Novičarski Članki in Povzetki: Nabor podatkov vsebuje novičarske članke iz medijskih virov, kot sta CNN (Cable News Network) in Daily Mail, skupaj s pripadajočimi povzetki. Ti članki pokrivajo različne teme in dogodke ter so različnih dolžin.
2. Avstraktivno Povzemanje: Za razliko od ekstraktivnega povzemanja, kjer se izvlečejo in združijo stavki iz izvirnega besedila, avstraktivno povzemanje vključuje ustvarjanje povzetka v povsem novih besedah. Nabor podatkov CNN/Daily Mail je priljubljen za tovrstno naloge avstraktivnega povzemanja.

CNN/Daily Mail podatkovna zbirka je običajno razdeljena na tri sklope:

1. učni (train) z 287.000 vrsticami primerov
2. validacijski (validation) z 13.400 vrsticami primerov
3. preizkusni (test) z 11.500 vrsticami primerov

7.2.5 semeval-2017 dataset

SemEval podatkovne zbirke so zbirke besedilnih podatkov, ki so anotirane za različne naloge na področju obdelave naravnega jezika.

Tukaj je nekaj ključnih značilnosti SemEval podatkovnih zbirk:

1. Anotacije: Podatki v SemEval podatkovnih zbirkah so običajno anotirani, kar pomeni, da so označeni z dodatnimi informacijami. Na primer, v podatkovni zbirki za naloge razreševanja sentimenta bi bili vzorci besedil označeni s pozitivnimi, negativnimi ali nevtralnimi sentimenti.
2. Naloge: Vsaka SemEval podatkovna zbirka je oblikovana za reševanje specifične naloge naravnega jezika. To lahko vključuje naloge, kot so analiza sentimenta, prepoznavanje imenovanih entitet, razreševanje ko-reference, klasifikacija besedil itd.
3. Raznolikost: SemEval podatkovne zbirke zajemajo širok spekter nalog, jezikov in domen. To omogoča raziskovalcem primerjavo modelov in pristopov na različnih področjih.
4. Uporaba v tekmovanjih: SemEval podatkovne zbirke se pogosto uporabljajo v tekmovanjih, imenovanih SemEval naloge. Tekmovalci tekmujejo za razvoj najboljših algoritmov za določeno nalogo in se primerjajo z drugimi udeleženci.

Raziskovalna skupnost: SemEval podatkovne zbirke so postale pomemben del naravnega jezika raziskovalne skupnosti, saj omogočajo primerjavo najnovejših pristopov in tehnologij na enotnem naboru podatkov. SemEval podatkovna zbirka je običajno razdeljena na tri sklope:

1. učni (train) z 49.547 vrsticami primerov
2. validacijski (dev) z 12.285 vrsticami primerov
3. preizkusni (test) z 12.285 vrsticami primerov

Chapter 8

Dataseti

8.1 Metrike

8.1.1 Opis spremenljivk za izračun metrik

Pravilno pozitivni (True Positive)

Je izraz, ki se uporablja v statistiki in strojnem učenju za opis primerov, kjer je model pravilno napovedal pozitiven rezultat za določen razred. To pomeni, da je model prepoznal pozitiven pojav, ko je bil dejansko prisoten.

Predpostavimo, da razvijamo model za prepoznavanje spam sporočil v elektronski pošti. Model pravilno prepozna 25 sporočil kot nezaželeno (spam), ki dejansko vsebujejo nezaželeno vsebino. To pomeni, da imamo 25 primerov "pravih pozitivnih". Te primere model pravilno prepozna kot spam, ker resnično vsebujejo neželeno vsebino.

Napačno pozitivni (False Positive)

Označuje situacijo, ko model napačno napove, da je nekaj pozitivno, medtem ko je v resnici negativno. Gre za vrsto napake, kjer model napačno identificira primer kot pripadajoč pozitivnemu razredu, čeprav dejansko pripada negativnemu razredu.

Na primer, predpostavimo, da imamo model za prepoznavanje spam sporočil v elektronski pošti. Če model označi sporočilo kot "spam", čeprav ni dejan-

sko spam, imamo situacijo lažno pozitivnega primera. Drugače povedano, model je napačno napovedal pozitiven primer (spam), ko je dejansko negativen primer (ni spam).

Napačno negativni (False Negatives)

Označuje napako, ki se pojavi v kontekstu klasifikacije ali analize besedila, ko model napačno napove, da je nekaj negativno, čeprav je v resnici pozitivno. To je vrsta napake, kjer model spregleda ali ne prepozna pozitivnih primerov.

V primeru analize besedila v naravni jezikovni obdelavi (NLP), false negative se zgodi, ko model ne uspe zaznati pozitivnega elementa v besedilu, ki bi ga moral prepoznati. Na primer, če imamo model za prepoznavanje pozitivnih izjav v komentarjih in model spregleda pozitivno izjavo, to bi bil primer false negative.

Predpostavimo, da imamo napreden sistem za filtriranje neželenih sporočil (spam), ki ga uporabljamo za preverjanje prihajajočih e-poštnih sporočil. Sistem je zasnovan tako, da prepozna in premika neželena sporočila v mapo za neželeno pošto.

Vendar pa se zgodi False Negative, ko sistem napačno presodi e-poštno sporočilo kot varno (ne-spam), čeprav vsebuje vse znake neželene vsebine. Na primer, če e-poštno sporočilo vsebuje povezave do nerealnih ponudb ali oglasov za sumljive izdelke, bi bila takšna sporočila številčno gledano ena od "False Negatives".

V tem primeru je sistem spregledal prepoznavo neželene vsebine, kar je povzročilo, da je sporočilo pristalo v glavnem predalu prejete pošte namesto v mapi za neželeno pošto. To lahko predstavlja težavo, saj se takšni neželeni vsebini lahko izognemo le, če sistem zanesljivo prepozna vse takšne primere.

8.1.2 Precision

Precision je pomembna metrika za ocenjevanje uspešnosti modelov v različnih nalogah, kot je klasifikacija, kjer se ukvarjamo z razdelitvijo podatkov v različne razrede. Poudarja natančnost pozitivnih napovedi, torej tistih primerov, ki jih model prepozna kot pozitivne. Visoka preciznost pomeni,

da so pozitivne napovedi modela zanesljive in imajo malo lažno pozitivnih napak.

V kontekstu naravne jezikovne obdelave (NLP), precision igra ključno vlogo pri razumevanju besedila. Na primer, pri analizi sentimenta želimo natančno ugotoviti, ali je izraz pozitiven ali negativen. Visoka preciznost v tem primeru pomeni, da so napovedi modela o sentimentu točne in se malo zmotijo.

Formula za izračun:

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

Povzeto po: <https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/>

8.1.3 Recall

8.1.4 F1

8.1.5 Accuracy

8.1.6 ROUGE

Chapter 9

Analiza raziskave

9.0.1 Prepoznavanje imenskih entitet (Named Entity Recognition)

Table 9.1: Prepoznavanje imenskih entitet

	Precision	Recall	F1
Transformers	0.949	0.953	0.951
Vertex AI	0.920	0.919	0.919
AWS SageMaker	0.954	0.961	0.958
Azure Cognitive Services	0.858	0.824	0.840

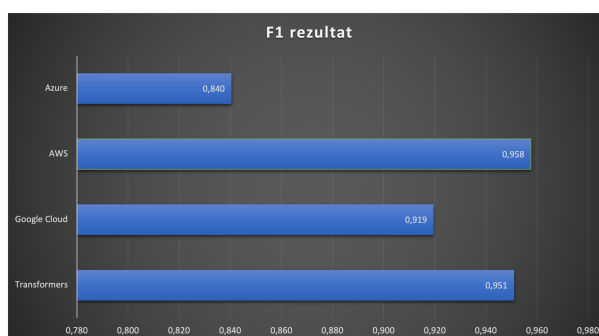


Figure 9.1: Prepoznavanje imenskih entitet F1 rezultat

Pri analizi imenskih entitet je bil uporabljen CONLL-2003 dataset.

Za prepoznavanje oseb (PER) in organizacij (ORG) se je najbolje izkazal Vertex AI storitev. Na splošno pa je bil v vseh področjih najboljši AWS SageMaker storitev.

9.0.2 Analiza sentimenta (Sentiment Analysis)

Table 9.2: Analiza sentimenta

	Precision	Recall	F1
Transformers	0.800	0.799	0.799
Vertex AI	0.924	0.924	0.924
AWS SageMaker	0.862	0.891	0.876
Azure Cognitive Services	0.862	0.891	0.877

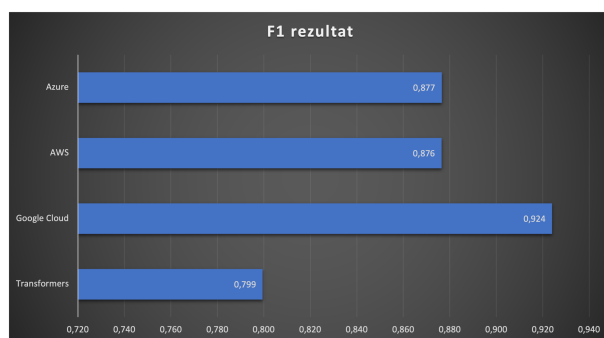


Figure 9.2: Analiza sentimenta F1 rezultat

Pri analizi imenskih entitet je bil uporabljen IMDb Reviews dataset.

Za analizo sentimenta je bila najboljša Vertex AI storitev.

9.0.3 Povzetek (Summarisation)

Table 9.3: Povzetek

	ROUGE-1	ROUGE-2	ROUGE-L
Transformers	0.209	0.190	0.192
Vertex AI	0.429	0.208	0.306
AWS SageMaker	0.226	0.021	0.201
Azure Cognitive Services	0.426	0.220	0.330

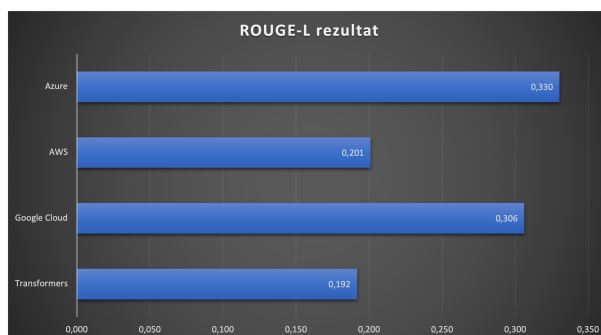


Figure 9.3: Povzetek F1 rezultat

Pri izdelavi povzetka je bil uporabljen dataset CNN/Daily Mail.

Kot vrhunska izbira za ustvarjanje povzetkov pa se je izkazala storitev Vertex AI.

9.0.4 Izvleček besedne zveze (Key Phrases)

Table 9.4: Izvleček besedne zveze

	Precision	Recall	F1
Transformers	0.475	0.573	0.519
Vertex AI	0.637	0.543	0.586
AWS SageMaker	0.513	0.622	0.562
Azure Cognitive Services	0.530	0.670	0.592

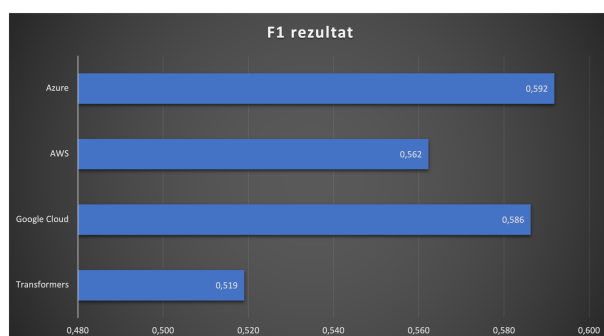


Figure 9.4: Izvleček besedne zveze F1 rezultat

Pri izvajanju naloge zvlečka besedne zveze je bil uporabljen podatkovni niz [semeval-2017](#).

Kot najboljša rešitev za naloge izvlečka besedne zveze pa se je izkazala storitev Azure Cognitive Services.

9.0.5 Klasifikacija besedila (Text Classification)

Table 9.5: Klasifikacija besedila

	Precision	Recall	F1
Transformers	0.930	0.926	0.928
Vertex AI	0.957	0.862	0.907
AWS SageMaker	0.858	0.763	0. 807
Azure Cognitive Services	0.880	0.920	0.900

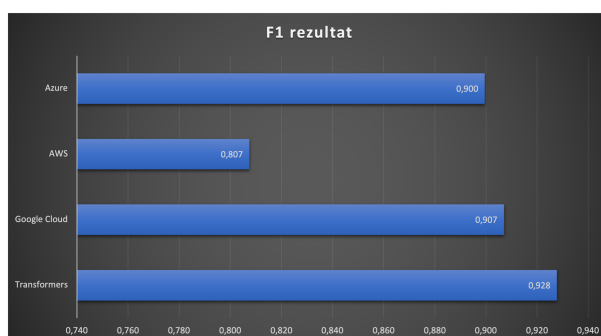


Figure 9.5: Klasifikacija besedila F1 rezultat

Pri izvajanju naloge klasifikacije besedila je bil uporabljen podatkovni niz IMDb Reviews.

Kot najboljša rešitev za naloge klasifikacije besedila pa se je izkazala storitev Transformers.

9.0.6 Zaznava objektov (Object Detection)

Table 9.6: Klasifikacija besedila

	Accuracy
Transformers	0.968
Vertex AI	0.977
AWS SageMaker	0.980
Azure Cognitive Services	0.965

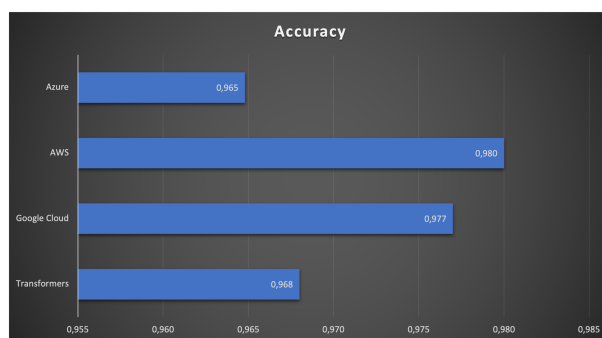


Figure 9.6: Zaznava objektov Accuracy rezultat

Pri zaznavi objektov je bil uporabljen COCO dataset.

Kot najboljša izbira za zaznavanje objektov pa se je izkazala storitev AWS SageMaker.

Chapter 10

Sklepne ugotovitve

Izbira \LaTeX ali ne \LaTeX je seveda prepuščena vam samim. Res je, da so prvi koraki v \LaTeX u težavni. Ta dokument naj vam služi kot začetna opora pri hoji.

Appendix A

Title of the appendix 1

Example of the appendix.

Bibliography

- [1] L. Fortnow, “Viewpoint: Time for computer science to grow up”, *Communications of the ACM*, št. 52, zv. 8, str. 33–35, 2009.
- [2] D. E. Knuth, P. Bendix. “Simple word problems in universal algebras”, v zborniku: Computational Problems in Abstract Algebra (ur. J. Leech), 1970, str. 263–297.
- [3] L. Lamport. *LaTEX: A Document Preparation System*. Addison-Wesley, 1986.
- [4] O. Patashnik (1998) BiBT_EXing. Dostopno na: <http://ftp.univie.ac.at/packages/tex/biblio/bibtex/contrib/doc/btxdoc.pdf>
- [5] licence-cc.pdf. Dostopno na: <https://ucilnica.fri.uni-lj.si/course/view.php?id=274>