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

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 LaTeX. 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 LaTeX. 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

Razširjeni povzetek

To je primer razširjenega povzetka v slovenščini, ki je obvezen za naloge pisane v angleščini. Razširjeni povzetek mora vsebovati vse glavne elemente dela napisanega v angleščini skupaj s kratkim uvodom in povzetkom glavnih elementov metode, glavnih eksperimentalnih rezultatov in glavnih ugotovitev. Razširjeni povzetek naj bo strukturiran v podpoglavja (spodaj je naveden le okvirni primer in je nezavezujoč). Čez palec navadno razširjeni povzetek nanese okoli 10 odstotkov obsega celotnega dela.

I Kratek pregled sorodnih del

II Predlagana metoda

III Eksperimentalna evaluacija

IV Sklep

poljuben tekst poljuben tekst

poljuben tekst poljuben tekst

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 1 bomo na hitro spoznali besedilne konstrukte kot so izreki, enačbe in dokazi. Naučili se bomo, kako se na njih sklicujemo. V Poglavju 2 se bomo srečali s sklicevanjem na besedilne konstrukte. Poglavje 3 bo predstavilo vključevanje plovk: slik in tabel. V Poglavju 5 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 (2.1) najprej preverimo za najmanjše naravno število — 0. Res, ker je $0 < 1 = 2^0$, je neenačba (2.1) za $n = 0$ izpolnjena.

Sledi indukcijski korak. S predpostavko, da je neenakost (2.1) 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:

$$\begin{aligned} n + 1 &< 2^n + 1, && (2.2)\\ &\leq 2^n + 2^n, && (2.3)\\ &= 2^{n+1}. \end{aligned}$$

Neenakost (2.2) je posledica indukcijske predpostavke, neenakost (2.3) pa enostavno dejstvo, da je za vsako naravno število n izraz 2^n vsaj tako velik kot 1. S tem je dokaz Izreka 2.1 zaključen. □

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 LaTeXom lahko vključimo vse. Slika 3.1 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 3.1.

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`.

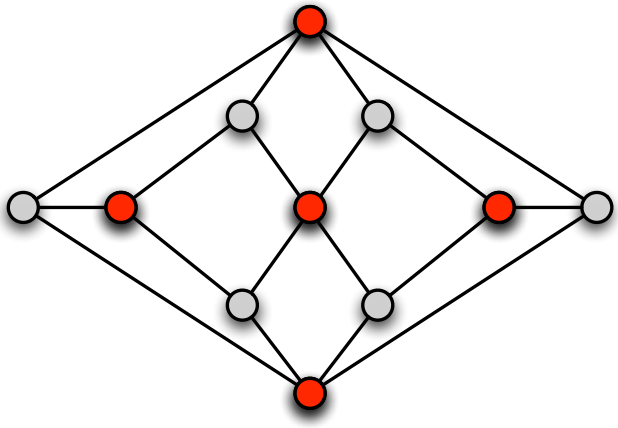

Figure 3.1: Herschelov graf, vektorska grafika.

Table 3.1

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

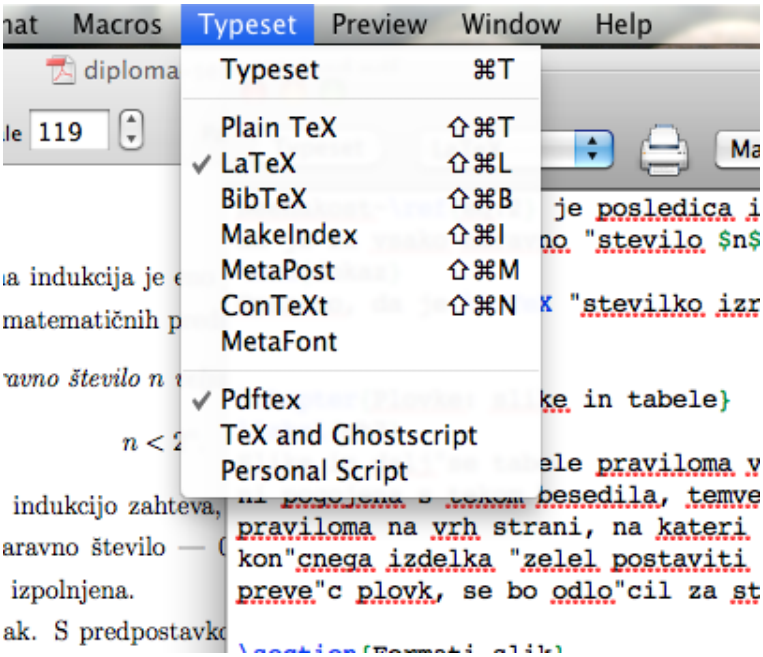

Figure 3.2: Kateri dialekt uporabljati?

Včasih se da v okolju za uporabo programskega paketa LaTeX nastaviti na kakšen način bomo prebavljali vhodne dokumente. Spustni meni na Sliki 3.2 odkriva uporabo LaTeXa v njegovi pdf inkarnaciji — `pdflatex`. Vključena Slika 3.2 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 $\boldsymbol{A}$ ter vektor $\boldsymbol{v}$,

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

4.2 Lepe tabele in psevdokoda

Psevdokoda 1 prikazuje primer delovanja genetskega algoritma, medtem ko Tabela 4.1 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: $\quad P'(t) \leftarrow Variation[P(t)] \leftarrow$ generiraj novo populacijo
6: $\quad EvalPopulation[P'(t)] \leftarrow$ evaluiraj novo populacijo
7: $\quad P(t+1) \leftarrow ApplyGeneticOperators[P'(t) \in Q]$
8: $\quad t \leftarrow t+1$
9: **until** prekinitev
10: **if** rezultat dovolj dober **then**
11: $\quad$ shrani rezultat
12: **end if**

Chapter 5

Kaj pa literatura?

Kot smo omenili že v uvodu, je pravi način za citiranje literature uporaba BibTeXa [4]. Programski paket LaTeXje prvotno predstavljen v priročniku [3] in je v resnici nadgradnja sistema TeX avtorja Donalda Knutha, znanega po denimo, če izpustim njegovo umetnost programiranja, Knuth-Bendixovem algoritmu [2].

Vsem raziskovalcem s področja računalništva pa svetujem v branje mnenje L. Fortnowa [1].

Chapter 6

Opis funkcionalnosti

6.1 Prepoznavanje imenskih entitet

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

Cilj NER 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 NER sistem prepoznal "Janez Novak" kot ime osebe, "10. avgust 1985" kot datum in "Ljubljana" kot lokacijo.

NER ima številne praktične uporabe, kot so:

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.

Metode prepoznavanja imenskih entitet se razvijajo in izpopolnjujejo s pomočjo strojnega učenja in naprednih algoritmov obdelave jezika. Ti sistemi so lahko zelo koristni pri obvladovanju in razumevanju velikih količin besedil, kar ima širok spekter uporabe v različnih industrijskih panogah.

6.2 Analiza sentimenta

Analiza sentimenta se nanaša na proces določanja čustvenega odziva, nagnjenosti ali stališča zapisanega besedila. Cilj analize sentimenta je ugotoviti, ali je določeno besedilo pozitivno, negativno ali nevtralno. To je lahko koristno pri analizi mnenj strank, razumevanju čustvenega odziva na izdelke, blagovne znamke, dogodke itd. 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 (NLP) 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 Povzetek

Pri povzetku gre za postopek avtomatskega ustvarjanja krajšega in jedrnatega povzetka izdaljšega besedila, kot je članek ali dokument. Namen povzemanja je izluščiti ključne informacije in ideje iz izvornega 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 izvornem 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 postopek, pri katerem 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 "Bag-of-Words" (vreča besed), kjer se besedilo pretvori v nabor besed, ki se pojavljajo v njem, in število pojavitev teh besed. Ta postopek lahko ponazorimo s pomočjo vektorja.

Nato sledi faza učenja, kjer uporabimo različne metode strojnega učenja, kot so Naivni Bayes, Logistična regresija, podporne vektorje, globoke nevronske mreže itd. Za učenje uporabimo skupino že klasificiranih besedil, imenovano učna množica ali učni nabor. Model se nauči prepoznati vzorce in odnose med besedili ter pripadajočimi kategorijami.

Ko je model uspešno naučen, ga preizkusimo na novih, neznanih besedilih, ki niso bila uporabljena med učenjem. S pomočjo naučenih vzorcev in odnosov model oceni, v katero kategorijo spada posamezno neznano besedilo.

Chapter 7

Sklepne ugotovitve

Izbira LaTeX ali ne LaTeX je seveda prepuščena vam samim. Res je, da so prvi koraki v LaTeXu 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) BibTEXing. 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`