

Elektrotehniško-računalniška strokovna šola in gimnazija Ljubljana

Analiza besedil s pomočjo podatkovnega rudarjenja

(slovenščina)

Raziskovalna naloga

Avtor: Simon Lipovšek, 2. letnik SŠ

Mentor: Jernej Golobič, prof. slovenščine

Ljubljana, april 2021

Kazalo vsebine

[Povzetek 5](#_Toc68892120)

[Abstract 6](#_Toc68892121)

[1. Uvod 7](#_Toc68892122)

[2. Teoretični del 8](#_Toc68892123)

[2.1 Podatkovno rudarjenje 8](#_Toc68892124)

[2.1.1 Lematizacija 8](#_Toc68892125)

[2.1.2 Oblak besed 9](#_Toc68892126)

[2.1.3 Sentimentalna analiza 10](#_Toc68892127)

[2.2 Python 12](#_Toc68892128)

[2.3 Programsko okolje Orange 12](#_Toc68892129)

[2.3.1 Uvozi dokument 13](#_Toc68892130)

[2.3.2 Korpus 13](#_Toc68892131)

[2.3.3 Gledalec korpusa 14](#_Toc68892132)

[2.3.4 Obdelava besedila 15](#_Toc68892133)

[2.3.5 Oblak besed 16](#_Toc68892134)

[2.3.6 Sentimentalna analiza 16](#_Toc68892135)

[2.3.7 Vreča besed 16](#_Toc68892136)

[2.3.8 Razdalje 17](#_Toc68892137)

[2.3.9 Hierarhično grozdenje 17](#_Toc68892138)

[2.3.10 Podatkovna tabela 17](#_Toc68892139)

[2.3.11 Graf raztrosa 18](#_Toc68892140)

[2.3.12 Tematsko modeliranje 18](#_Toc68892141)

[2.4 Literatura 19](#_Toc68892142)

[2.4.1 Romantika 19](#_Toc68892143)

[2.4.2 France Prešeren 20](#_Toc68892144)

[2.4.3 Ivan Cankar 21](#_Toc68892145)

[2.5 Opredelitev problema 21](#_Toc68892146)

[3. Empirični del 22](#_Toc68892147)

[3.1 Metoda 22](#_Toc68892148)

[3.1.1 Prikaz najpogostejših besed v besedilu 22](#_Toc68892149)

[3.1.2 Čustvena vrednost besedila 23](#_Toc68892150)

[*3.2* Tema besedila 24](#_Toc68892151)

[3.2.1 Podobnosti med besedili 25](#_Toc68892152)

[4. Rezultati 25](#_Toc68892153)

[4.1 Rezultati analize besedil Franceta Prešerna 26](#_Toc68892154)

[4.1.1 Rezultate analize literarnega dela *Krst pri Savici* 31](#_Toc68892155)

[4.2 Rezultati analize besedil Ivana Cankarja 36](#_Toc68892156)

[4.3 Rezultati analize besedil Antona Tomaža Linharta 38](#_Toc68892157)

[5. Razprava 39](#_Toc68892158)

[6. Zaključek 41](#_Toc68892159)

[7. Priloge 42](#_Toc68892160)

[7.1 Koda za menjavo besed oz. znakov v programskem jeziku Python 42](#_Toc68892161)

[8. Viri in literatura 42](#_Toc68892162)

[8.1 Literatura 42](#_Toc68892163)

[8.2 Viri 43](#_Toc68892164)

**Kazalo slik**

[Slika 1: Logo programskega jezika Python 12](#_Toc68892088)

[Slika 2: Orange logo 13](#_Toc68892089)

[Slika 3: Modul: Uvozi dokument 13](#_Toc68892090)

[Slika 4: Modul: Korpus 14](#_Toc68892091)

[Slika 5: Modul: Gledalec korpusa 14](#_Toc68892092)

[Slika 6: Modul: Obdelava besedila 15](#_Toc68892093)

[Slika 7: Prikaz nastavitev v modulu obdelave besedila 15](#_Toc68892094)

[Slika 8: Modul: Sentimentalna analiza 16](#_Toc68892095)

[Slika 9: Modul: Vreča besed 16](#_Toc68892096)

[Slika 10: Modul: Razdalje 17](#_Toc68892097)

[Slika 11: Modul: Hierarhično grozdenje 17](#_Toc68892098)

[Slika 12: Modul: podatkovna tabela 17](#_Toc68892099)

[Slika 13: Modul: Graf raztrosa 18](#_Toc68892100)

[Slika 14: Modul: Tematsko modeliranje 18](#_Toc68892101)

[Slika 15: Prikaz skupka modulov za iskanje najpogostejših besed 23](#_Toc68892102)

[Slika 16: Prikaz skupka modulov za sentimentalno analizo 24](#_Toc68892103)

[Slika 17: Prikaz skupka modulov za tematsko analizo besedil 24](#_Toc68892104)

[Slika 18: Prikaz skupka modulov za iskanje ponavljajočih besed 25](#_Toc68892105)

[Slika 19: Analiza Prešernovih besedil z modulom oblak besed in vključeno lematizacijo besedila 27](#_Toc68892106)

[Slika 20: Analiza Prešernovih besedil z modulom oblak besed in izključena lematizacija besedil 28](#_Toc68892107)

[Slika 21: Prikaz sentimentalne analize Prešernove literature z grafom raztrosa 29](#_Toc68892108)

[Slika 22: Prikaz sentimentalne analize ter njihove povezave Prešernovih besedil z modulom hirerhično grozdenje 30](#_Toc68892109)

[Slika 23: Prikaz ponavljajočih besed v Krstu pri Savici z vključeno lematizacijo 31](#_Toc68892110)

[Slika 24: Prikaz posamične sentimentalne analize uvoda in krsta z grafom raztrosa 32](#_Toc68892111)

[Slika 25: Prikaz pozitivnih besed v najbolj pozitivnih kiticah literarnega dela Krst pri Savici 33](#_Toc68892112)

[Slika 26: Prikaz negativnih besed v zgoraj naštetih negativnih kiticah literarnega dela Krst pri Savici 34](#_Toc68892113)

[Slika 27: Prikaz sentimentalne vrednosti posameznih kitic Krsta pri Savici 35](#_Toc68892114)

[Slika 28: Prikaz sentimentalne analize pozitivnih kitic z modulom hierarhično grozdenje 35](#_Toc68892115)

[Slika 29: Prikaz sentimentalne analize izpostavljenih negativnih kitic z modulom hierarhično grozdenje 36](#_Toc68892116)

[Slika 30: Prikaz grozdenja besedil Ivana Cankarja 36](#_Toc68892117)

[Slika 31: Prikaz sentimentalne analize del Ivana Cankarja 37](#_Toc68892118)

[Slika 32: Prikaz sentimentalne analize besedil Antona Tomaža Linharta 38](#_Toc68892119)

# Povzetek

V pričujoči nalogi smo predstavili podatkovno rudarjenje, osredotočili smo se na oblak besed, sentimentalno analizo in tematsko modeliranje. Prav tako smo predstavili programsko okolje Orange in delo v njem. Analizirali smo literarna besedila kanonskih slovenskih avtorjev, Franceta Prešerna, Ivana Cankarja ter Antona Tomaža Linharta, in preverjali uporabnost tovrstne metode. Potrdili smo uporabnost podatkovnega rudarjenja s programom Orange, vendar spoznali, da nekatere metode zahtevajo več časa pri pripravi besedil na analizo. Zaradi tega moramo biti pri rezultatih pozorni oz. moramo ponoviti postopek priprave besedila in analize.

**Ključne besede:** podatkovno rudarjenje, Orange, slovenska literatura, programska vizualizacija literature, sentimentalna analiza, oblak besed

# Abstract

In this paper, we have researched data mining. Our focus was on Orange modules word cloud, sentiment analysis and topic modulation. We represented the software environment Orange and work in it. With the help of Orange software, we analysed the canonical literature of Slovenian authors. These Slovenian authors are France Prešeren, Ivan Cankar and Anton Tomaž Linhart. We affirmed productive usage of data mining with software Orange. Learn that some texts must still be prepared by hand before processed with software for further analysis. Notice that sometimes we must recheck the results and repeat the preparation process before making conclusions about the literature.

**Keywords:** data mining, Orange, Slovenian literature, software visualization of literature, sentimental analysis, word cloud.

# Uvod

Interpretacija besedil je navadno zahteven postopek, ki se ga učimo v srednji in osnovni šoli pri pouku književnosti pri slovenščini. S tem znanjem lahko razumemo zahtevnejša besedila ter lažje razločimo njihov pomen. Eno izmed zahtevnejših besedil, ki jih obravnavamo v drugem letniku srednje šole, je *Krst pri Savici*. To besedilo smo natančneje pregledali s programom Orange. Program Orange nam pripomore k hitrejšemu iskanju tematske besede ter razbiranju vzdušja v literarnem delu. Čeprav dijaki oz. študentje le redko pomislijo na analizo besedil s pomočjo podatkovnega rudarjenja, je to izredno priporočen pripomoček za hitro in lažje razumevanje besedil.

Predmet analize pričujočega dela predstavljajo literarna dela Franceta Prešerna, Antona Tomaža Linharta in Ivana Cankarja, ki skupaj tvorijo ogromno bazo besedil oz. korpus. Večina njihovih literarnih del je prosto dostopnih v različnih oblikah besedilnih dokumentov, vendar jih je potrebno za analizo ustrezno pripraviti, kar pa je postopek, ki ga je težje avtomatizirati. Kanonska besedila so za bralce zahtevnejša od trivialne literature, zato potrebujejo dlje časa, da jih razumejo oz. dešifrirajo. Postopek bi morda olajšali rezultati podatkovnega rudarjenja, npr. oblak besed, ustvarjanje katerih nam omogoča program Orange. Takšne metode predstavljajo nov, alternativen, do zdaj še skorajda neraziskan način vstopa v svet književnosti.

# Teoretični del

## Podatkovno rudarjenje

Izraz podatkovno rudarjenje lahko zasledimo tudi pod izrazom odkrivanje podatkov ali odkrivanje znanja. Je novodobni proces analiziranja podatkov in povzemanje njih v uporabne informacije. Podatkovno rudarjenje se v večini primerov izvaja z računalnikom. Nekaj let nazaj so bili podatki ročno vneseni v računalnik, zdaj pa to ni več nujno potrebno. Večina podatkov za podatkovno rudarjenje se zdaj zbira preko interneta, in sicer s pomočjo pajkov. Pajki so programi, ki samodejno preiščejo internet ter želene besedilne dokumente in zberejo relevantne podatke. Programi, ki omogočajo podatkovno rudarjenje, so še eni izmed številnih analitičnih orodij za analizo podatkov. Pravilna podatkovna analiza lahko uporabniku predstavi podatke iz različnih pogledov, jih razvrsti v skupine oz. podskupine in prikaže razmerja med njimi. Na kratko se podatkovnemu rudarjenju lahko reče iskanje povezav ali vzorcev v nekem skupku podatkov.

### Lematizacija

Pri podatkovni analizi slovničnih besedil se velikokrat srečamo s podobno napisanimi besedami z enakim geslom. Programi ne vidijo podobnosti med takšnimi besedami, zato je potrebna lematizacija besedila. Pri lematizaciji besedila je potrebno najti geslo posameznih besed v besedilu ter zapisati vse besede, ki so bile opredeljene v isto geslo, enako. Definicija lematizacije iz SSKJ je določanje besedilnih enot za gesla ali podgesla v slovarju. Je torej proces, ki pretvori besede v njihovo slovarsko obliko, smiseln je samo pri besednih vrstah, ki so pregibne. Podobna metoda je korenjenje, ki v vsako pregibno besedo pretvori v koren. Razlika med lematizacijo in korenjenjem besede je končni rezultat. Pri lematizaciji je končna beseda vedno slovnično pravilna, medtem ko pri korenjenju to ne velja.

Kot primer lahko uporabimo samostalnik *avto*. Pri sklanjatvi besede avto se spreminjajo končnice (*avta, avtu, z avtom* itd.). Za ljudi je je razumljivo, da tudi z preurejanjem končnice še vedno govorimo o avtu. Pri podatkovni analizi program tega ne prepozna in tolmači različne oblike besede (*avta*, *avtu*, *z avtom*) kot samostojne besede.

Lematizacija ima še vedno mnogo problemov, kot je pretvarjanje besed, ki se zaradi pravopisnih pravil (npr. na začetku povedi) pišejo z veliko začetnico. Take besede program ne bo uvrstil pod enak sklop, saj je eno napisano z veliko, drugo pa z malo začetnico.

### Oblak besed

Oblak besed je grafična vizualizacija frekvence ponavljajočih besed znotraj besedila. Te besede nam lahko predstavljajo literarno osrednje motive, simbole, snov in tematiko. Pri vstavljanju besedil v modul oblak besed je treba biti pozoren, da je besedilo pravilno urejeno. Besedilo mora biti urejeno na sledeč način:

1. Razdeliti besedilo na posamezne člene oz. besede (angl. tokenize text).
2. Odstranjevanje veznikov in drugih ne čustvenih besed (angl. stop words removal).
3. Odstranitev naglasov, potrebno paziti, da se ne odstrani šumnikov (angl. punctuation removal).
4. Pretvoriti vse besedilo v velike oz. male začetnice.
5. Lematizacija besedila.

Modul oblak besed je posebno občutljiv na zapise besed, zato je potrebno besedilo izredno dobro pripraviti.

Literarna snov je snov, ki je izhodišče nastanka literarnega dela. Snov je obstajala še preden se jo je avtor odločil ubesedil iti in je zato imenovana pred jezikovna prvina književnega dela. Ko je neka snov oblikovana in ujeta v besedilu nekega avtorja postane literarna. Za literarno snov je lahko uporabljeno kakršno koli stvarno ali nestvarno dogajanje. Do 18. stoletja je bil izbor snovi ozek, saj so bili pisatelji in pesniki omejeni na plemstvo, bogove, junake idr. Danes pa temu ni tako in lahko pesniki in pisatelji pišejo o kakršnokoli snovi.

Motiv izrazimo z pomensko zvezo ali besedo. Motiv je izluščen iz splošne snovi, kot so predmet, dogodek, situacije, pojmi, čustva in pojavi. Primeri: neuresničena ljubezen, smrt, čas, obžalovanje idr.

Temo nam odgovori na vprašanje o čem govori literarno delo oz. večina tega. Temo lahko delimo na razumsko ali čustveno. Razumske teme so lahko vojna, socialna, bivanjska idr. Čustvene teme so lahko ljubezenske, prijateljske, družinske idr.

### Sentimentalna analiza

Osnovna sentimentalna analiza besedila določi odnos pisca do teme ali sama čustva napisana v besedilu. Pove nam torej, ali je odnos pisca do teme pozitiven, nevralen ali negativen. Naprednejše sentimentalne analize lahko določijo tudi čustva tega besedila, kot so jeza, žalost, sreča.

Prva obsežna sentimentalna analiza besedil je bila leta 2002 in se je imenovala Turney. Tuney je uporabljal preprosto analizo, ki je dosegla 74-odstotno natančnost. Leta 2005 so uporabljali enako metodo in primerjali filme ter primerjali njihove ocene z ocenami podeljenih zvezdic gledalcev. Prvo iskanje potencialne vrednosti sentimentalne analize v Twitterju je bilo leta 2009. Od takrat se je uporaba interneta močno povečala in je zdaj gonilna sila za sentimentalno analizo. Sentimentalna analiza je bila v preteklosti omejena na ljudi, ki znajo programirati. Zdaj pa nam program Orange omogoča preprosto sentimentalno procesiranje besedil tudi za ljudi, ki niso vešči v programiranju.

Mi smo za sentimentalno analizo besedila uporabili način analize čustev na podlagi besedil (angl. rule based sentimental analysis). Ta način se navezuje na podatke pridobljene z jezikovnimi strokovnjaki. Končano delo tega se imenuje sentimentalni oz. čustveni leksikon (angl. sentimental lexicon). V njem so besede kvalificirane z pozitivno ali negativno konstanto. Spreminja se tudi velikost oz. intenziteta besede. Preden pošljemo besedilo v sentimentalno analizo je potrebno ustrezno urediti, po sledečih postopkih:

1. Razdeliti besedilo na posamezne člene oz. besede (angl. tokenize text).
2. Odstranjevanje veznikov in drugih ne čustvenih besed (angl. stop words removal).
3. Odstranitev naglasov, potrebno paziti, da se ne odstrani šumnikov (angl. punctuation removal).

Kot primer lahko vzamemo poved *Miha je spekel dobro potico*. Besedilo je treba porazdeliti na posamezne člene in dobimo: »Miha«, »je«, »spekel«, »dobro«, »potico«.

Treba je odstraniti veznike, v tem primeru veznik »je«.

Program nato določi sentimentalno vrednost posamezni besedi.

»Miha« = nevtralno

»spekel« = nevtralno

»dobro« = pozitivno z neko amplitudno konstanto

»potico« = rahlo pozitivno

S tem program zaključi da je vrednost besedila pozitivna. Seveda je program še bolj zapleten, ampak to je nekakšna osnova, po kateri deluje.

## Python

Program Orange in tudi posebno narejen program spodaj temeljita na prosto dostopnem večnamenskem programskem jeziku python. Leta 1990 ga je ustvaril Guido van Rossum. Od različice 2.1 si avtorske pravice lasti Python Software Fondation. Organizacija je neodvisna, neprofitna in si prizadeva za razvoj odprtokodne tehnologije.

Programski jezik Python je podatkovno orientiran oz. je optimiziran za hitro ter preprosto analizo zbranih podatkov. Vključuje tudi samodejno upravljanje pomnilniškega prostora ter dinamične podatkovne tipe. Je eden izmed priljubljenejših programskih jezikov, zato ima številne skupnosti, kjer lahko zaprosite za pomoč ali pa pregledate razreševanje njihovih problemov. Programski jezik ima na voljo številne prosto dostopne knjižnice, ki jih lahko uporabnik uporablja ali preureja. Knjižnice lahko uporabnik tudi sam napiše ter objavi na splet.



Slika 1: Logo programskega jezika Python

## Programsko okolje Orange

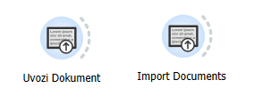
Orange je odprtokodni interaktivni program, ki omogoča rudarjenje podatkov, strojno učenje in vizualizacijo podatkov. Glavne komponente so moduli, ki jih je mogoče povezovati na programskem platnu. Program je odprtokoden zato lahko vsak uporabnik po želji dodaja oz. ureja module. Napisan je v programskem jeziku Python, tako da je za razvijanje modulov potrebno znanje v programiranju z jezikom Python. Vsak modul lahko prejme podatke, jih uredi oz. obdela ter jih vrne kot izhodne podatke. Preprost način programiranja modulov nam omogoča hitro in preprosto dodajanje modulov. Dodatne module lahko dodamo v Orange kot dodatek. Program je razvil in ga dodatno razvija Laboratorij za bioinformatiko Fakultete za računalništvo in informatiko (FRI) na Univerzi v Ljubljani.



Slika 2: Orange logo

### Uvozi dokument

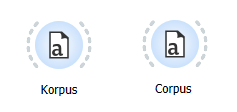
Z modulom uvozi dokument oz. angl. Import documents, se uvaža dokumente iz računalniških datotek v program. Modul prebere tudi pozicijo dokumenta v sistemu in ime njegove mape. Imena map ter dokumentov se lahko predaja na drugim modulom.



Slika 3: Modul: Uvozi dokument

### Korpus

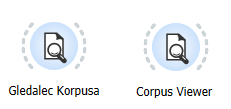
Modul korpus oz. angl. corpus, se uporaba pri ureditvi uvoženih dokumentov v besedila in prepoznavo besedil znotraj mape. Poleg tega prevede besedila v lažjo obdelujočo datoteko in ustvari navidezni prazni prostor v katerega lahko moduli dodajo prirejene vrednosti dokumentu. Vse, kar je bilo prirejeno pri modulu korpus ali kasneje, ne vpliva na originalno datoteko, ampak je shranjen v obliki kopije v korpusu ali v posameznih modulih.



Slika 4: Modul: Korpus

### Gledalec korpusa

Gledalec korpusa oz. angl. corpus viewer, nam prikaže kako je besedilo urejen v modulu korpus. V primeru da so besedila urejena v mape ali podmape se to prikaže v njem. Če je dokumentu prirejena vrednost bo videna v tem modulu.



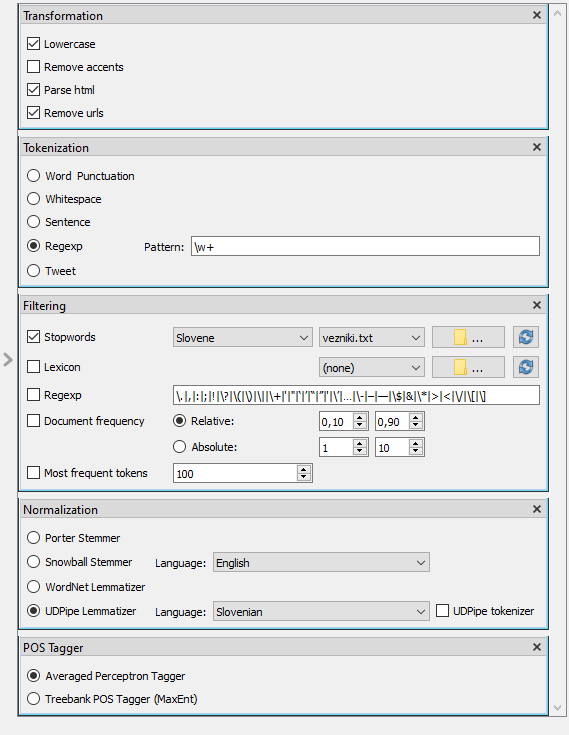
Slika 5: Modul: Gledalec korpusa

### Obdelava besedila

Obdelava besedila oz. angl. preprocess text, nam uredi besedila za bolj pravilno analizo. V njem so različne nastavitve, ki lahko avtomatsko uredijo besedilo za pravilnejšo prihodnjo analizo. Spremeni lahko velike in male začetnice, odstrani linke, odstrani naglase, odstrani veznike in druge želene besede ki jih lahko določi uporabnik, lematizira besedilo itd.



Slika 6: Modul: Obdelava besedila



Slika 7: Prikaz nastavitev v modulu obdelave besedila

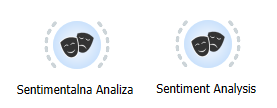
### Oblak besed

Oblak besed oz. angl. Word cloud, prikaže frekvenco ponavljanja besede v besedilu z velikostjo. Višja kot je frekvenca ponavljanja besede v besedilu, večja je prikazana beseda.



### Sentimentalna analiza

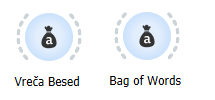
Sentimentalna analiza oz. angl. sentiment analysis poda sentimentalno vrednost posamezni besedi v besedilu oz. poda sentimentalno vrednost celotnega besedila.



Slika 8: Modul: Sentimentalna analiza

### Vreča besed

Vreča besed oz. angl. bag of words, prikaže število ponovitve besede v besedilu oz. dokumenta v datoteki.



Slika 9: Modul: Vreča besed

### Razdalje

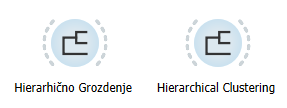
Razdalje oz. angl. distances, izračuna posamezne razdalje med stolpci in vrsticami. Razdalje so lahko normalizirane z različnimi funkcijami. Za procesiranje besedil je najbolj optimalna evklidska funkcija. Končni rezultat je matrica podatkov ter njihovi odnosi.



Slika 10: Modul: Razdalje

### Hierarhično grozdenje

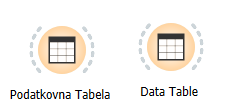
Hierarhično grozdenje oz. angl. hierarchical clustering, prikaže matrico podatkov, ki jo ustvari modul razdalje z hierarhičnim prikazom.



Slika 11: Modul: Hierarhično grozdenje

### Podatkovna tabela

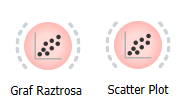
Podatkovna tabela oz. angl. data table, prikaže vse podatke usmerjene vanj v obliki tabele.



Slika 12: Modul: podatkovna tabela

### Graf raztrosa

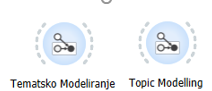
Graf raztrosa oz. angl. scatter plot prikaže podatke v obliki dvodimenzionalnega grafa. Za boljšo reprezentacijo so točke lahko obarvane ter izbrane za prikaz imen.



Slika 13: Modul: Graf raztrosa

### Tematsko modeliranje

Tematsko modeliranje oz. angl. topic modeling, išče prej izbrano število tem v besedilu. Temo predstavljajo pogoste besede oz. besedne zveze. Pri prepoznavi teme določi vsaki besedi utež glede na pomembnost v besedilu in katere vrste beseda je.



Slika 14: Modul: Tematsko modeliranje

## Literatura

### Romantika

Romantika je literarno obdobje oz. literarna smer, ki ima svoje korenine v različnih kulturnih gibanjih po Evropi sredi 18. stoletja. 50 let kasneje se pojavi novoromantična smer, ki začenja počasi nadomeščati romantiko. Poleg novoromantične smeri romantiko začenja nadomeščati realizem, najbolj zakasnel primer romantike se pojavi v drugi polovici 19. stoletja.

Temeljne značilnosti romantike:

* Razklanost med stvarnostjo in idealom (stvarnost se jim zdi slaba).
* Idealnost posameznika se skriva v čustvih (poudarjajo čustva ter njihov pomen).
* Ignoriranje klasičnih pravil o obliki in vsebini književne zvrsti.
* Umikanje iz stvarnosti v domišljijo.
* Zanimanje za ljudsko slovstvo ter njegovo preprostost.
* Nasprotovanje plemiški kulturi.
* Upiranje avtoritetam, tradiciji in nasilju vladajočega socialnega in političnega reda.

Značilne vrste proz ki so se takrat razvile so; lirika, epika in dramatika.

* Lirika

V primerjavi z klasično književnostjo je lirika izrazito čustvena ter razpoloženjska, pogosto je tudi politična in socialna. Pesniški jezik postane metaforično bogatejši, pesnik pa predstavlja posrednik med naravnim in nadnaravnim.

* Epika

Epika ima prosto obliko dolžine, saj se lahko razteza v dolge zgodbe podobne romanom ali pa je kratka po vsebin kakor pripovedka. V prozi je usoda romantičnega junak ali skupnosti prikazana s slikovitimi dogodki in napetim dogajanjem. Pod epske pesniške oblike spadajo romanca, balada in soneti.

* Dramatika

V primerjavi s klasicizmom so se v romantičnem obdobju razvile še nekatere bolj mešane vrste dramatske osnove, kot so; tragikomedija, melodrama, dramska pesnitev. Z zgodovine je črpala legende, ljubezenske zgodbe, pustolovke in psihološke igre.

### France Prešeren

France Prešeren se je rodil leta 1800 v Vrbi na Gorenjskem, umrl pa leta 1848 v Kranju. V osnovno šolo je hodil v Ribnici, kjer se je tudi zapisal v zlato knjigo. Gimnazijo in 2 letnika filozofije je opravil v Ljubljani. Študij za odvetnika pa na Dunajski pravni fakulteti. Po končanem študiju se je vrnil v Ljubljano, da bi delal kot odvetnik. Večino njegovega dela je bilo napisanih v tem času. Prešeren se je nesrečno zaljubil v Primičevo Julijo, ki ji je kasneje posvetil *Sonetni venec* in še nekaj pesmi. Leta 1835 je utonil njegov prijatelj Matija Čop. Njegovi smrti je posvetil *Krst pri Savici*. Približno dve leti kasneje je spoznal Ano Jelovšek s katero je imel 3 otroke. Leta 1839 se v Ljubljan vrne Andrej Smole in je skupaj s Prešernom izdajal knjige. Leto zatem Andrej Smole umre na svojem godu. Prešeren je zaradi izgub številnih prijateljev ter zavrnjene ljubezni začel piti. Znova se je zaljubil v Jerico Podboj, ki je bila hčer gostilničarke v katero je pogosto zahajal. Napisal ji je nekaj pesmi ampak tudi ta ljubezen je bila neuslišana. Leta 1843 so začele izhajati različne novice, ki so naznanjale prihod novih časov za Slovence. Prešeren je v njih poskušal objaviti nekaj pesmi, ki mu jih je Janez Bleiweis nekajkrat zavrnil, kasneje pa mu jih je dovolil objaviti na slabše vidnem mestu. Začel je pisati zbirko *Poezije*, ki jo je pred izdajo moral poslati v pregled cenzuri. Ena izmed takrat cenzuriranih pesmi je bila Zdravljica. Leta 1848 mu je uspelo objaviti *Zdravljico*. Leto zatem pa je umrl z vodenico.

Prešeren velja za enega izmed najkvalitetnejših ustvarjalcev romantične književnosti, vendar v času njegovega življenja njegova dela niso bila priznana.

#### **Krst pri Savici**

Delo *Krst pri Savici* je napisal France Prešeren, leta 1835 in je bilo posvečeno smrti njegovega prijatelja Matija Čopa. Je obsežna epsko-lirska pesnitev, ki je sestavljena iz dveh delov plus posvetilni del Matiju Čopu. Samo založno jo je izdal leta 1836 v 600 izvodih.

##### **Uvod**

V uvodu je tema pokristjanjevanje kristjanov, kar vsebuje veliko bojevanja. Dogajanje je postavljeno na Ajdovski gradec. Glavni osebi sta Črtomir kot predstavnik poganstva in Valjhun kot predstavnik kristjanov. Uvod spada pod junaško epiko in se lahko vsebinsko razčleni na tri dele. Prvi del (1.–10. kitica) opisuje dogajanje pred bitko, drugi del (11.–17. kitice) zajema Črtomirjev govor, tretji del (18.–26. kitica) pripoveduje o bitki.

##### **Krst**

V krstu je tema besedila posameznikova problematika. Dogaja se po izgubljeni bitki in sicer ob slapu Savica. Konča se junaška epika in začne lirska epika. Začenja se vpogled v Črtomirjevo življenje, razlog za boj in njegovo ljubezen Bogomilo.

### Ivan Cankar

Ivan Cankar se je rodil 10. 5. 1876 v Vrhniki, kot osmi otrok in umrl 1918 v Ljubljani. Po končani šoli v Vrhniki se je vpisal na ljubljansko realko. Pridružil se je društvu Zadruga, kjer se je spoznal z Murnom, Kettejem in Župančičem. Kot dijak je objavljal pesmi v različnih revijah. Šolanje je nadaljeval na Dunaju kot tehnik ampak se je kmalu prepisal na slavistiko. Po šolanju je živel pri bratu in napisal Hlapci. Leta 1909 se je preselil v Ljubljano. Začel se je ukvarjati z politiko in o predlagal združitev slovanskih narodov. Zaradi tega je bil aretiran in vpoklican v vojsko. Kmalu zatem je bil oproščen služenja v vojski zaradi bolezni. Bolan je umrl dve leti kasneje.

## Opredelitev problema

Pred raziskovanjem smo si zastavili naslednja raziskovalna vprašanja:

1. Je uporaba programa Orange smiselna za analizo slovenskih šolskih besedil?
2. Ali so podatki pridobljeni pri analizi slovenskih literarnih del z programom Orange zanesljivi?
3. Ali lahko računalnik pravilno določi pomen besednim zvezam s členkom *ne*, kot so *ni žalosten*, *ni vesel* idr.?
4. Ali lahko določimo podobnosti besedil s programom Orange?
5. Ali je analiza sentimentalnih besedil z programom Orange uporabna v vsakdanjem življenju?

# Empirični del

## Metoda

Priprava na raziskovanje se je pričela z zbiranjem literarnih besedil. Odločili smo so leposlovje slovenskih avtorjev 19. in 20. stoletja, ki se jih obravnava v srednji šoli. Besedila so prosto dostopna na Wikiviru, kar je bila dodatna prednost, saj ni bilo treba prositi nosilcev avtorskih pravic za dovoljenje, prav tako ni bilo dodatnega s pretipkavanjem oz. skeniranjem besedila, kar bi bilo izredno časovno potratno. Literarna dela smo uvrstili v skupine ter podskupine, in sicer vsakega v svojo mapo v datotečnem sistemu, kar zahteva program Orange. Nekatera besedila so vsebovala avtorje, datume ter druge nepomembne zapiske, ki jih je bilo treba pred začetkom izbrisati. Nadaljnja stopnja priprave besedil je zajemala spreminjanje pregibnih besed v njihovo slovarsko obliko. Takšne besede smo v urejevalniku besedila s pomočjo regularnih izrazov pretvorili v želeno obliko. Ponovni pregled je odkril, kar je program spregledal, to smo nato ročno popravili, in sicer s posebej napisanim programom v programskem jeziku python, ki je za delovanje potreboval tri datoteke. Prva datoteka mora vsebovati besedilo, ki ga želimo procesirati. Druga datoteka vsebuje besede oz. znake, ki jih hočemo zamenjati. Tretja datoteka vsebuje besede oz. znake, s katerimi hočemo zamenjati besede oz. znake v drugem dokumentu. Besede oz. znaki morajo biti vzporedno zapisani skozi drugi in tretji dokument. Če bi torej hoteli zamenjati črko *è* s črko e, bi morali črko *è* vstaviti v drugi dokument, črko *e* pa v tretji dokument, in sicer tako, da sta črki v isti vrstici. Nove znake oz. besede lahko samo dopišemo v novo vrstico. Če smo hoteli najti podobnosti med posameznimi kiticami ali besedili, smo jih shranili v isto datoteko in to datoteko naložili v program Orange. Vanj smo datoteke naložili z modulom uvažanje dokumentov. Znotraj programa smo modul povezali z naslednjim modulom korpus. Korpus smo povezali v modul urejanje dokumentov, ki pa ima veliko možnosti za različne nadaljnje povezave. Čeprav imajo moduli različne vrste kombinacij, ki so koristne pri različnih vrstah besedil, smo opazili, da so nam najbolj priročne informacije pridobljene na spodaj razloženih kombinacijah.

### Prikaz najpogostejših besed v besedilu

Besede smo prikazali s pomočjo modula oblak besed ter podatkovno tabelo. Preden smo besedilo vstavili v oblak besed, smo ga najprej uredili z modulom obdelava besedil. S tem smo izločili veznike ter nepotrebne veznike. Vse besede, ki so se v besedilu pojavile več kot 70-krat, so bile prikazane v modulu, in sicer večkrat, ko se beseda ponovi, večja je. Besede, ki so se ponovile manjkrat, so bile dostopne preko modula podatkovna tabela. Na sliki 14 je prikazano zaporedje dejavnosti od uvoza besedila v programsko orodje Orange do oblaka besed in tabele s vseh besed z informacijo o številu ponovitev.

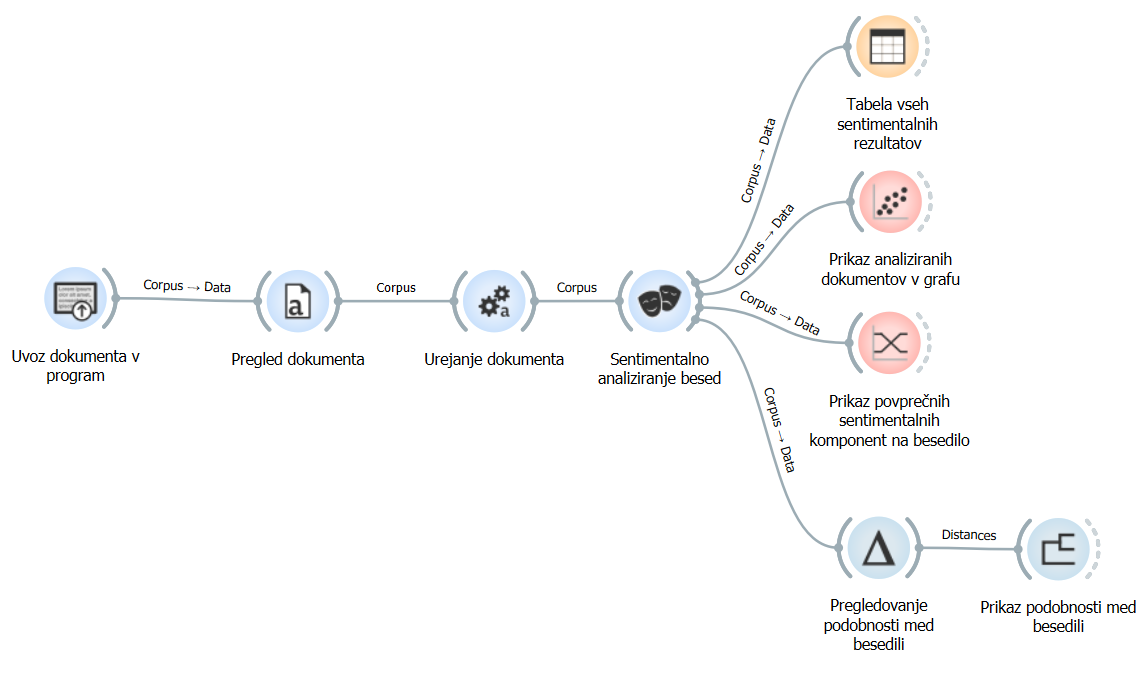


Slika 15: Prikaz skupka modulov za iskanje najpogostejših besed

### Čustvena vrednost besedila

Čustvene vrednosti smo dodali besedam s pomočjo modula sedimentarna analiza. Ta je vsaki besedi v besedilu se določi pozitivno in negativno konstanto. Z izračunanimi pozitivnimi in negativnimi vrednostmi besedil smo tvorili graf raztrosa, podatkovno tabelo, linearni graf ter hierarhično grozdenje. Pred tem smo iz besedila odstranili nepotrebne besede in veznike.

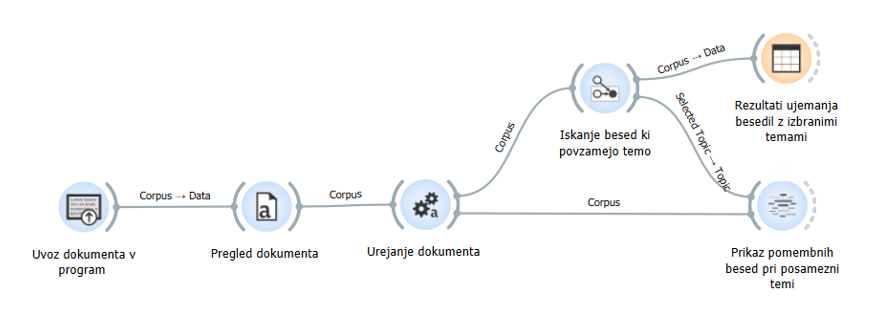
Podatkovna tabela nam je prikazala vse podatke analiziranih besedil. Graf raztrosa prikaže vsako besedilo v tabeli. V tej tabeli je bila kot horizontalna os označena skupna pozitivnost posamezne literature in vertikalna os kot skupna negativnost posamezne literature. Linearni graf nam je prikazal število nevtralnih, pozitivnih in negativnih besed vseh izbranih besedil ter razmerje med posameznimi besedami. Hierarhično grozdenje je prikazalo besedila, ki so si najbolj podobna bolj skupaj, in besedila, ki so si manj podobna, bolj narazen.



Slika 16: Prikaz skupka modulov za sentimentalno analizo

## Tema besedila

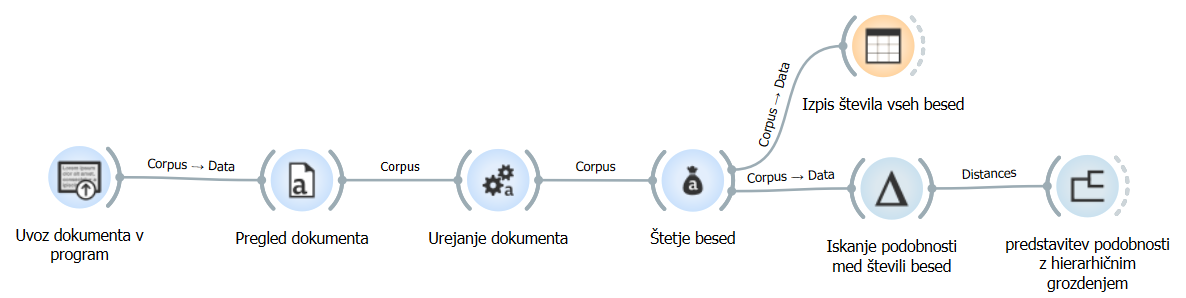
Z modulom tematsko modeliranje smo poiskali besede ki naj bi predstavljale skupno temo oz. teme v izbranih besedilih. Tematsko modeliranje je poiskalo besede, ki se pojavljajo v posameznih temah različnih besedil. Oblak besed je vizualno prikazal besede, ki so bile podane iz posameznih tem.



Slika 17: Prikaz skupka modulov za tematsko analizo besedil

### Podobnosti med besedili

Modul vreča besed je preštel število vseh besed v besedilu. Podatke je poslal naprej na modul razdalje. Ta je določil podobnost med posameznimi besedili. Podobnost smo prikazali z modulom hierarhično grozdenje.



Slika 18: Prikaz skupka modulov za iskanje ponavljajočih besed

# Rezultati

S pomočjo programa Orange smo analizirali različna besedila. Najbolj smo se osredotočili na Prešernova dela, saj so njegova besedila močno zastopana v učnem načrtu. V analizo smo vključili še nekaj besedil Ivana Cankarja, Simona Gregorčiča in Antona Tomaža Linharta. Prešernova dela smo dodatno še razporedili v podskupine balade in romance, gazele, sonetje, zbadljivi sonetje ter soneti nesreče. Skupina balade in romance je vsebovala literarna dela: *Dohtar; Hčere svet; Judovsko dekle; Lenora; Neiztrohnjeno srce; Orglar; Povodni mož; Prekop; Ribič; Turjaška Rozamunda; Učenec; Zdravilo ljubezni; Ženska zvestoba. Skupina sonetje je vsebovala literarna dela: Bilo je, Mojzes! Tebi naročeno; Dve sestri videle so zmoti vdane; Je od vesel'ga časa teklo leto; Kadar prividi učenost zdravnika; Kupido! Ti in tvoja lepa starka; Marsikateri romar gre v Rim, v Kompostelje; Na jasnem nebi mila luna sveti; Ni znal molitve žlahtič trde glave; Očetov naših imenitna dela; Oči bile pri nji v deklet so sredi; Odprlo bo nebo po sodnem dnevi; Sanjalo se mi je, de v svetem raji; Sonetni venec; Tak kakor hrepeni oko čolnarja; Velika, Togenburg! Bla je mera; Vrh sonca sije soncov cela čeda; Zgodi se včasih, de mohamedani. Skupina zbadljivi sonetje je vsebovala literarna dela: Al prav se piše kaшa ali kaſha; Apel podobo na ogled postavi; Ne bod'mo šalobrade! Moskvičanov*. V analizo so bili zajeti tudi vsi *Sonetje nesreče*: *Čez tebe več ne bo, sovražna sreča!; Hrast, ki vihar na tla ga zimski trešne; Komur je sreče dar bila klofuta; Memento mori; O Vrba! Srečna draga vas domača; Popotnik pride v Afrike puščavo; Življenje ječa, čas v nji rabelj hudi*. Za stransko oz. preventivno reprezentacijo smo razporedili tudi dela Ivana Cankarja v dramatiko, prozo in pesmi. Dramatično podskupino predstavljajo literarna dela: *Hlapci, Jakob Ruda, Kralj na Betajnovi, Lepa Vida*, *Pohujšanje v dolini Šentflorjanski, Romantične duše, Za narodov blagor*. Literarna dela Ivana Cankarja, ki smo ji vstavili v podskupino pesmi, so: *Pes. Maček in miš; Sodba; Sreča; Helena*. Literarna dela, ki smo jih vstavili v podskupino proza, so: *Aleš iz Razora, Božična zgodba, Črtice, Gospa Judit, Grešnik Lenart, Hiša Marije pomočnice, Hlapec Jernej in njegova pravica, Izpoved, Jesenske noči, Knjiga za lahkomiselne ljudi, Križ na gori, Kurent, Martin Kačur, Milan in Milena. Mimo Življenja, Moja njiva, Moje življenje, Nezbrane vinjete, Nina, Novela doktorja Grudna, Novo Življenje, Ob zori, Podobe iz sanj, Polikarp, Popotovanje Nikolaja Nikiča, Potepuh Marko in kralj Matjaž, Smrt in pogreb Jakoba Nesreče*.

## Rezultati analize besedil Franceta Prešerna

Pri lematizaciji besedil Franceta Prešerna smo opazili, da program poskuša pretvoriti imena v navadne besede oz. predvideva, da so slovnične napake, in jih pretvori v najbolj podobne besede. Kljub temu lematizacija ne imenskih besed ni povzročala nobenih vidnih problemov. Pri pregledovanju oblaka besed z lematizacijo z izjemo besede *bogomiti,* za kar smo predvidevali, da pomeni *Bogomila*, smo lahko razbrali kar nekaj podatkov, ki se ujemajo s tematiko ter pisateljevem obdobju pisanja. Opazili smo, da večino besedil vsebuje povezavo z ljubeznijo ter motivom vere, časa in smrti. Ti rezultati so bili pričakovani, saj je Prešeren pisal v času romantike. Če nebi vedeli kdo je in kdaj je živel avtor tega besedila s temi podatki še ne moremo uvrstiti dela v specifično književno obdobje, lahko pa zožimo izbor ter predvidevamo njegov slog pisanja.

Spodnja slika prikazuje besedila Franceta Prešerna, ki so bila lematizirana in prikazana z modulom oblak besed. Največje napisane besede predstavljajo besede, ki se v zbrani literaturi največkrat pojavijo. Besede, ki se največkrat pojavijo, lahko poimenujemo osrednji motivi, okoli katerih so se vrtela vsa Prešernova besedila.



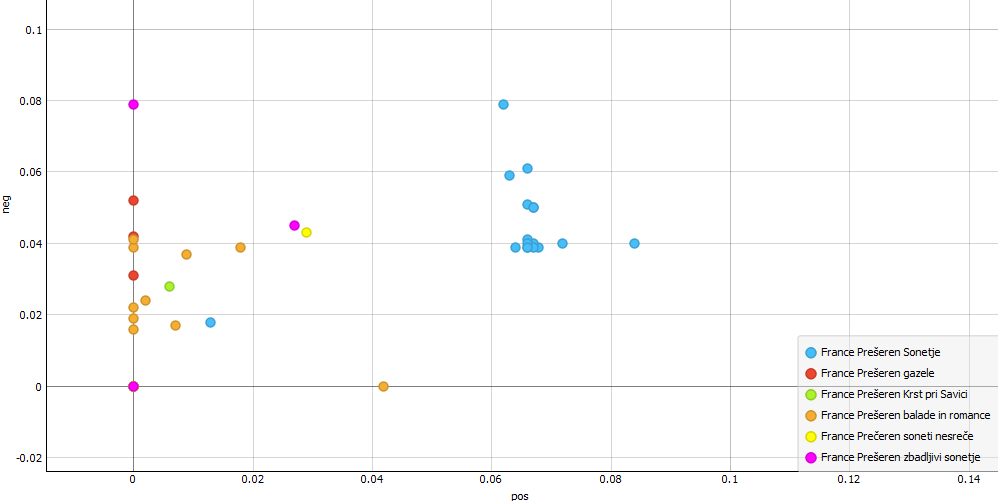
Slika 19: Analiza Prešernovih besedil z modulom oblak besed in vključeno lematizacijo besedila

Spodnja slika prikazuje besedila Franceta Prešerna, ki niso bila lematizirana pred obdelavo z modulom oblak besed. Največje napisane besede predstavljajo besede, ki se v zbrani literaturi največkrat pojavijo. Besede, ki se največkrat pojavijo lahko poimenujemo osrednji motivi okoli katerih so se vrtela vsa Prešernova besedila. Opazi se, da začenjajo prevladovati besede katerih se zapis manjkrat spremeni pri sklanjanju oz. stopnjevanju.



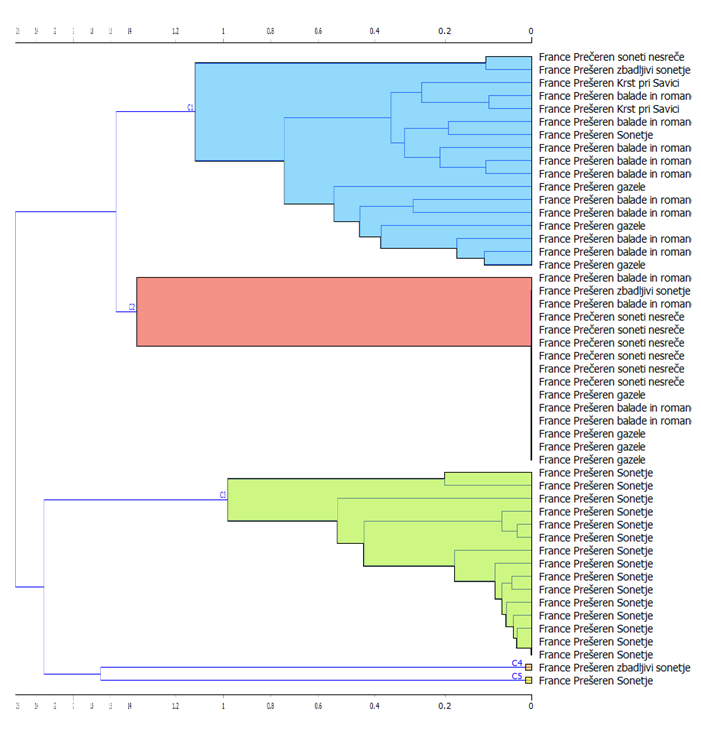
Slika 20: Analiza Prešernovih besedil z modulom oblak besed in izključena lematizacija besedil

Z grafom raztrosa lahko opazimo, kako različne vrste književnih besedil ustvarijo različno vzdušje. Razvidno je, da so Prešernovi soneti več čustveno zaznamovanih besed, ki jih je algoritem zaznal kot močno pozitivne oz. močno negativne, v primerjavi z gazelam, baladam in romancam.



Slika 21: Prikaz sentimentalne analize Prešernove literature z grafom raztrosa

Pri sentimentalni analizi Prešernovih besedil in hierarhičnim prikazom pridobljenih podatkov, opazimo, da so Prešernovi sonetje razred zase v primerjavi z drugimi njegovimi pesniškimi oblikami. Posamične podskupine bi se dalo še dodatno analizirati, njihove podobnosti, tematiko, ampak smo se odločili, da jo to izven dela te naloge. Leva stran grafa je malo bolj stisnjena za boljšo vidljivost.



Slika 22: Prikaz sentimentalne analize ter njihove povezave Prešernovih besedil z modulom hirerhično grozdenje

### Rezultate analize literarnega dela *Krst pri Savici*

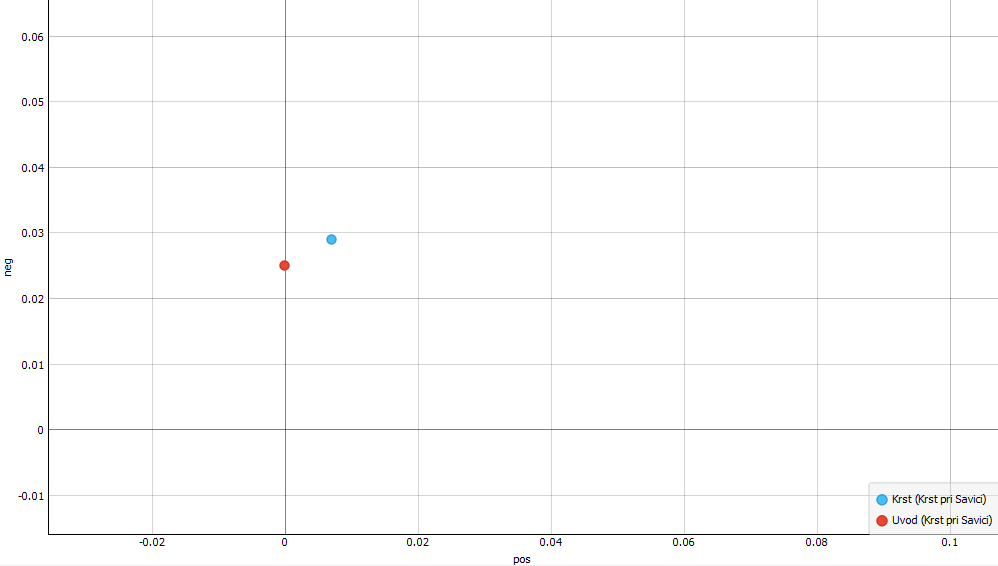
Pri pregledovanju Prešernovega *Krsta pri Savici* smo opazili, da bi lahko nekatere besede, ki se največkrat pojavijo, spadale pod ključne besede oz. osrednje motive, okoli katerih se lahko vrti dogajanje oz. zgodba. Sentimentalna analiza je pokazala vrednost besedila kot -0,17. Rahle probleme je delala prekratka dolžina besedila in različni naglasi na enakih besedah.

Spodnja slika prikazuje analizirano besedil *Krst pri Savici*, ki je bilo urejeno z lematizacijo in modulom oblak besed. Največje napisane besede predstavljajo besede, ki se v zbrani literaturi največkrat pojavijo. Besede, ki se največkrat pojavijo, lahko poimenujemo osrednji motivi, okoli katerih so se vrti *Krst pri Savici.* Lematizacija ni prepoznala imena Bogomila kot ime, zato ga je preimenovala bogomiti. Temu bi se dalo izogniti, če bi imel algoritem boljšo podporo slovenskemu jeziku ali pa z nadoknadenimi ročnimi popravki. Kljub manjši zamenjavi imena se lepo vidijo glavni motivi besedila.



Slika 23: Prikaz ponavljajočih besed v Krstu pri Savici z vključeno lematizacijo

V spodnjem grafu smo literarno delo *Krst pri Savici* razdelili na dva glavna dela in sicer krst in uvod. Horizontalna os na grafu nam prikazuje pozitivno vrednost besedila, vertikalna os pa negativno vrednost besedile. Sentimentalne vrednosti besedil so bile določene z modulom sentimentalna analiza. Sentimentalne vrednosti za besedila so prikazane v procentih. Na grafu lahko opazimo, da v uvodu kjer opisuje vojno in bitke je sentimentalna vrednost večinoma razporejena med negativnimi in nevtralnimi besedami. Krst vsebuje nekaj pozitivnih besed in tudi več negativnih besed kot uvod. V krstu se večino zgodbe odvija izven vojne oz. opisuje ob vojna dogajanja.



Slika 24: Prikaz posamične sentimentalne analize uvoda in krsta z grafom raztrosa

Opazimo tudi, da je pri literarnem delu *Krst pri Savici* program pozitivne kitice opisal samo 34., 80., 49., 58., 40. in 48., preostale pa so označene kot negativne, in sicer izstopajo 1., 27., 65., 2., 55., 24. Nekatere besede, ki naredijo zgoraj naštete kitice pozitivne, so: pride, domu, postane, junak idr. Nekatere izmed negativnih besed zgoraj naštetih negativnih kitic so: kri, prelita, življenj, končano idr. Večina pozitivnih in negativnih besed se v primerjavi z nevtralnimi besedami pojavijo malokrat, zato je priporočljivo imeti daljša besedila, ki to lažje odstranijo. V teh primerih so besedila krajša in lahko prikažejo tudi nekatere nevtralne besede.

Spodnja slika prikazuje prikaz pozitivnih besed oz. besedne zveze v 6 najbolj pozitivnih kiticah z modulom oblak besed.



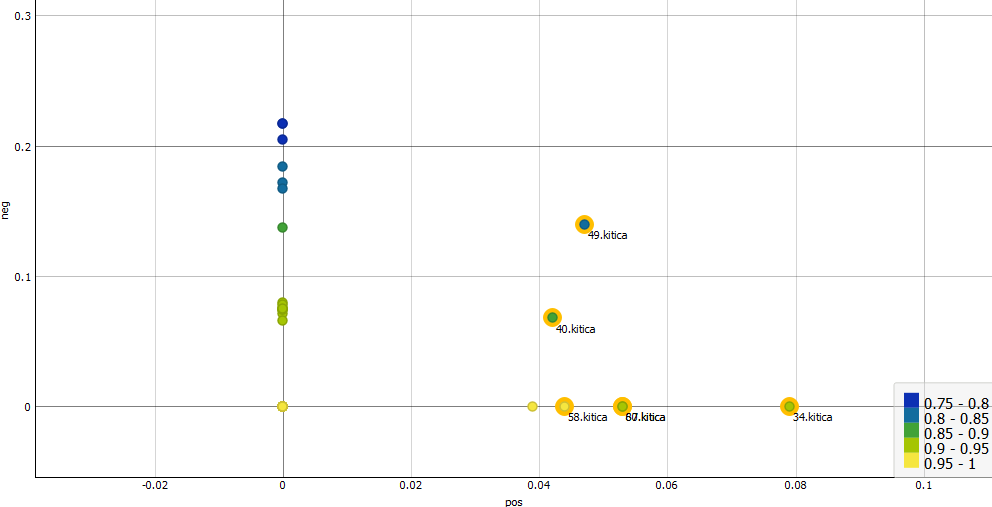
Slika 25: Prikaz pozitivnih besed v najbolj pozitivnih kiticah literarnega dela Krst pri Savici

Spodnja slika prikazuje negativne besede oz. besedne zveze zgoraj izpostavljenih negativnih kitic. Predstavljene so z modulom oblak besed.



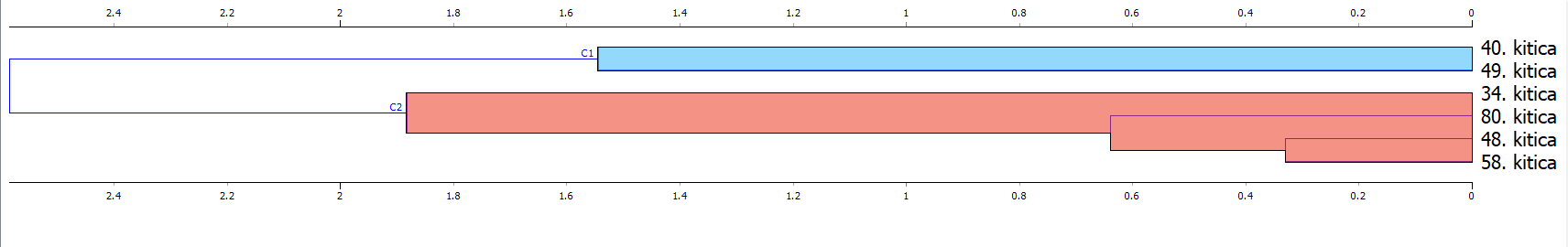
Slika 26: Prikaz negativnih besed v zgoraj naštetih negativnih kiticah literarnega dela Krst pri Savici

Barva na spodnjem grafu prikazuje nevtralno vrednost, ki jo je program dodelil posamezni kitici, horizontalna os prikazuje pozitivni koeficient kitice in vertikalna os prikazuje negativni koeficient kitice. Na desni strani grafa so označene izpostavljene pozitivne kitice, negativne izpostavljene kitice pa predstavlja najbolj negativnih šest kitic vidnih na levi strani grafa.



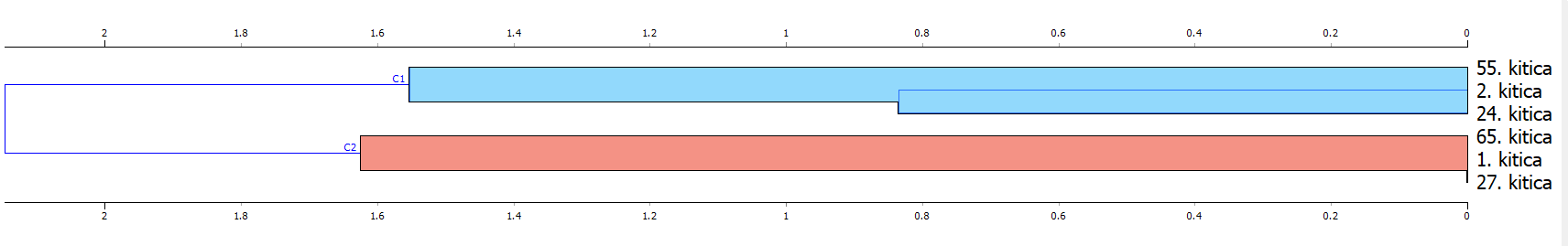
Slika 27: Prikaz sentimentalne vrednosti posameznih kitic Krsta pri Savici

Spodnja slika prikazuje sentimentalno medsebojno podobnost, njihovo urejenost ter razmerje med posameznimi zgoraj naštetimi pozitivnimi kiticami.



Slika 28: Prikaz sentimentalne analize pozitivnih kitic z modulom hierarhično grozdenje

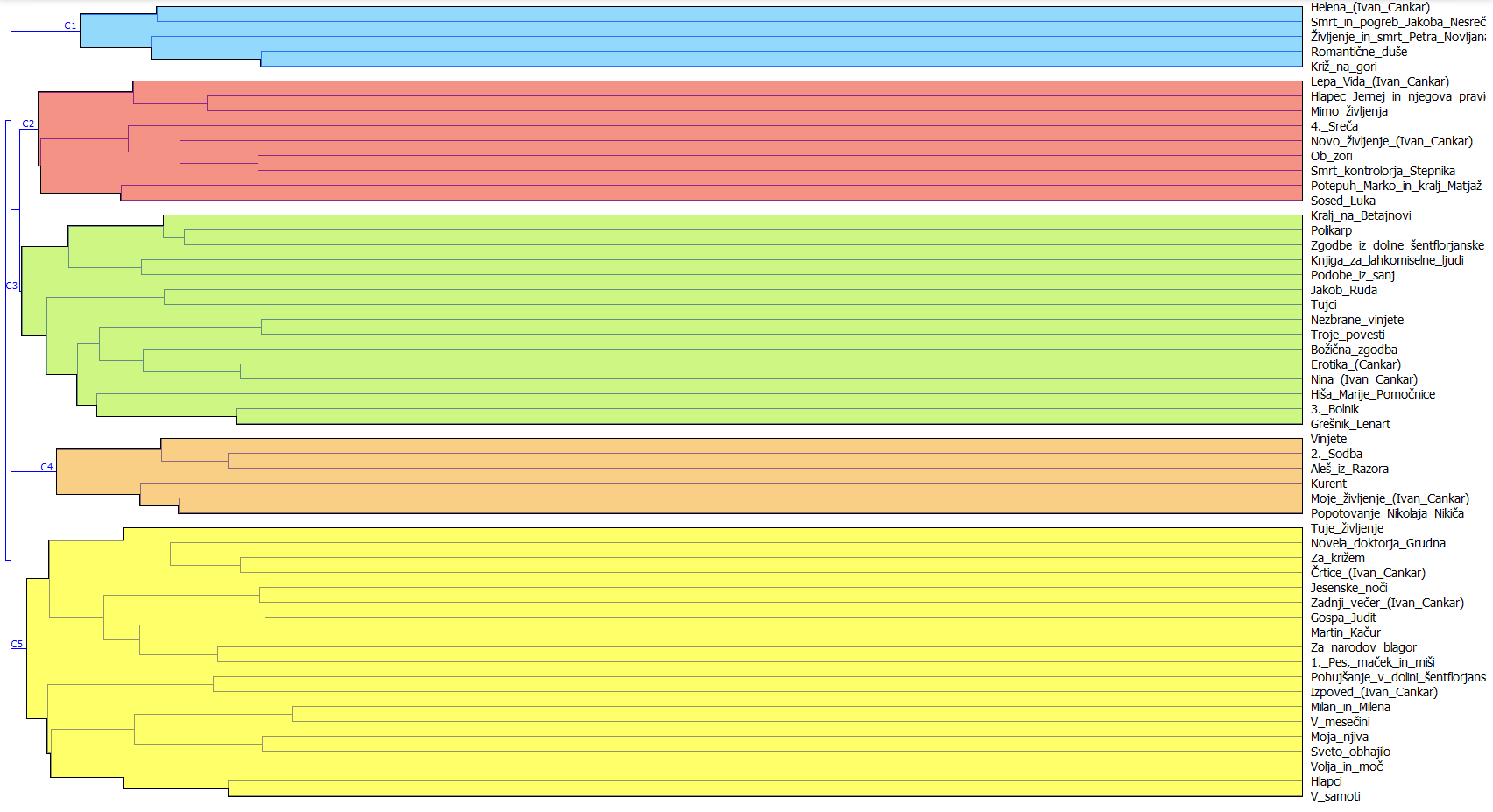
Spodnja slika prikazuje sentimentalno medsebojno podobnost, njihovo urejenost ter razmerje med posameznimi zgoraj naštetimi negativnimi kiticami.



Slika 29: Prikaz sentimentalne analize izpostavljenih negativnih kitic z modulom hierarhično grozdenje

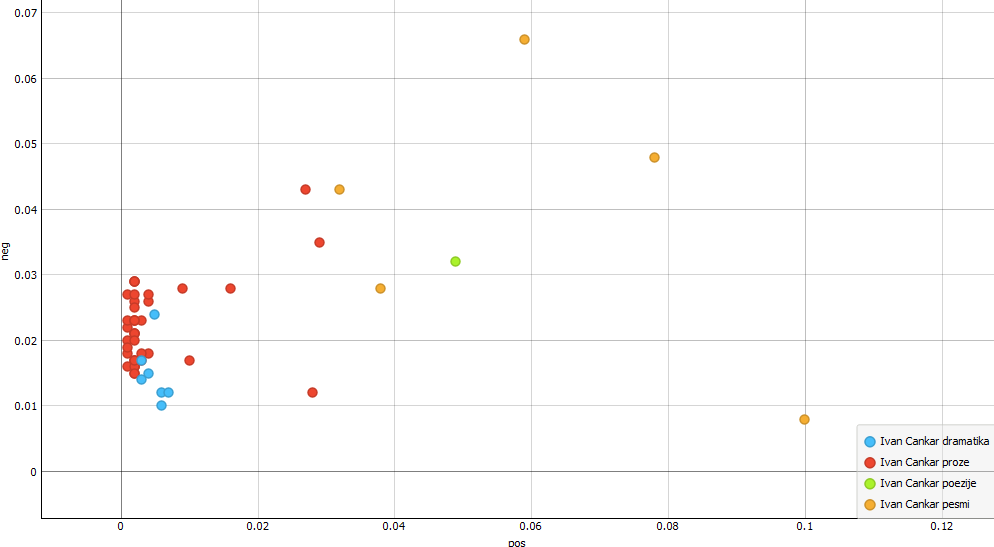
## Rezultati analize besedil Ivana Cankarja

Besedila Ivana Cankarja so lep primer daljših besedil, kar omogoča lažjo uvrščanje v podskupine.



Slika 30: Prikaz grozdenja besedil Ivana Cankarja

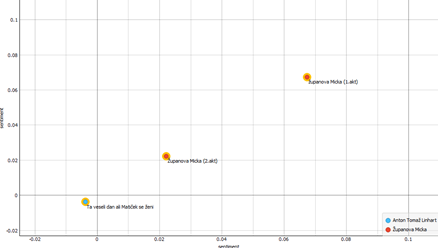
Na spodnji so prikazane različne literarne vrste besedil Ivana Cankarja. S tega grafa je razvidno, da je Cankarjeva lirika čustveno bolj razgibana kot dramatika ter preostala dela.



Slika 31: Prikaz sentimentalne analize del Ivana Cankarja

## Rezultati analize besedil Antona Tomaža Linharta

Izbrana besedila Antona Tomaža Linharta sta bili *Ta veseli dan ali Matiček se ženi* in *Županova Micka*. Slednja je razdeljena na dva akta oz. dejanji, zato smo ti dve tudi analizirali posamično. Pokazalo se je, da je čustvena podobnost med komedijo *Ta veseli dan ali Matiček se ženi* in drugim aktom *Županove Micke* večja kot med posameznima aktoma *Županove Micke*. Kljub temu pa je *Županova Micka* še vedno bolj pozitivno kot *Ta veseli dan ali Matiček se ženi*. Izbrani besedili se med zbranimi deli pisateljev najbolj ujemata z besedili Ivana Cankarja.



Slika 32: Prikaz sentimentalne analize besedil Antona Tomaža Linharta

# Razprava

V nalogi smo ugotavljali kredibilnost programa Orange pri analizi oz. podatkovnem rudarjenju različnih književnih del (angl. text mining). Medtem ko so v liriki besede, ki imajo lahko več pomenov, so v epiki in dramatiki prisotne besede, na katerih lahko izvedemo zanesljivejšo sentimentalno analizo. Daljša besedila, kot so proze omogočajo bolj natančno štetje besed ter ujemanja bistva besedila. Besedila vsebujejo tudi naglasna znamenja, ki jih je treba odstraniti, saj jih program nato smatra za različne besede. Pri prebiranju rezultatov je treba biti pozoren na napačne rezultate oz. je treba biti kritičen do pridobljenih rezultatov, saj so ti lahko zavajajoči.

1. **Je uporaba programa Orange smiselna za analizo slovenskih šolskih besedil?**

Uporaba programa Orange za analizo slovenskih šolskih besedil je smiselna. V šoli se pri slovenščini učimo razumevanja ter opisovanja prebranega. S pomočjo Orange programa lahko šolarji pred prebiranjem besedila ugotovijo, na katere stvari je dobro biti pozoren pri branju. Z rezultati podatkovnega rudarjenja si lahko pomagajo pri pisanju obnove, povzetka ali druga besedila, ki se navezujejo na procesirano besedilo. Omogoča tudi vidno reprezentacijo čustev ter uporabo besed v besedilu in s tem lahko lažje določijo književno zvrst oz. obdobje literarnega dela.

1. **Ali so podatki pridobljeni pri analizi slovenskih literarnih del z programom Orange zanesljivi?**

Program prikazuje večinoma zanesljive podatke, problemi se pojavljajo pri visoki občutljivih modulih, saj jih moti zapis besed. Pravilna predpriprava besedila je potrebna, saj nekatera besedila vsebujejo tudi nepotrebne podatke. Mi smo sledili spodaj razloženemu postopku:

* 1. Prenos dokumenta ter ureditev besedila v besedilno datoteko.
  2. Zamenjava naglasnih znamenj, kot so: è, ó, à, í idr. Znamenja smo zamenjali s pomočjo Python programa, ki podpira regularne izraze, v katerega smo napisali niz črk oz. besed, ki smo jih želeli zamenjati, in niz črk oz. besed, s katerimi smo jih zamenjali. Eden izmed načinov je tudi uporaba bližnjice ctrl + h v Wordu, kamor je sicer treba vpisati vsako obliko in jo nato zamenjati ponovno.
  3. Znotraj programa Orange lahko za dodatno preventivo analizirate besedilo z lematizacijo in brez nje. S tem bi pridobili natančnejše podatke pri štetju besed lahko pa tudi povzroči pravilnega štetja pri imenih i so podobni ne imenskim besedam.

1. **Ali lahko računalnik pravilno določi pomen zanikanim besednim zvezam, kot so; *ni žalosten, ni vesel* idr.?**

V programskem okolju Orange so nekateri moduli, kot je tematsko modeliranje lahko določi drugačno vrednost besednim zvezam, večina jih pa v tem zaostaja, saj pregledujejo samo posamične besede. Programsko okolje Orange ni prilagojeno za napredno procesiranje besednih zvez v sentimentalnih besedili. Zanesljivost programa se s tem malo zmanjša, saj večina literatur vsebuje le malo čustvenih nikalnih besednih zvez. Večina čustvenih besed imajo za nasproten pomen že opredeljeno svojo besedo.

1. **Ali lahko določimo podobnosti besedil s programom Orange?**

Podobnost besedil lahko določimo s programom Orange. Podobnost je ena izmed najbolj natančnih in zanesljivih lastnosti, ki jih lahko določimo s programom Orange. Znotraj programa lahko izbiramo med različnimi vrstami podobnosti ter različnimi vizualizacijami le teh. Glavne tri vrste za primerjavo besedil, ki smo jih mi uporabljali so: tematsko, besedno in čustveno. Let te smo pa prikazali z moduli oblak besed, graf raztrosa, podatkovna tabela in hierarhično grozdenje.

1. **Ali je analiza sentimentalnih besedil s programom Orange uporabna v vsakdanjem življenju?**

Sentimentalna analiza besedil s programom Orange je uporabna v vsakdanjem življenju. Uporablja se lahko za analizo besedil pred branjem, kar pripomore za lažje osredotočenje bralca na pomembe stvari pri branju besedila. Pri pisanju obnov ali drugih besedil, ki se navezujejo ali izhajajo iz procesiranega besedila. Program je tudi primeren za iskanje velike baze literatur za iskanje podobne npr. izbira naslednje knjige za branje če hočeš prebirati enako oz. podobno tematiko. Sentimentalna analiza besedil lahko pripomore k razvoju čustveno občutljivih umetnih inteligenc.

# Zaključek

Pri raziskovalni nalogi smo se naučili uporabe programa Orange ter ugotavljali implementacijo zbranih podatkov v pouk. Program Orange je preprost ter interaktiven načina za analizo ter vizualizacijo podatkov. Kljub njegovi preprostosti lahko postane močno orodje pri analizi daljših besedil. Program omogoča uporabnikom pestro izbiro obdelave ter vizualizacije podatkov poleg tega dopušča še razvoj dodatnih modulov, ki so lahko bolj specializirani. Program ni namenjen optimiziranem procesiranju besedil, saj je presplošen in lahko hitro preobremeni računalnik. Ima nekaj razvitih modulov za slovenščino, ampak kljub temu nekateri moduli ne zaznajo besedila. Rezultati teh modulov so zato lahko zavajajoči in jih je treba ročno obdelati ter naknadno pregledati. Ročna obdelava besedil lahko znatno upočasni in oteži proces. Za razreševanje napak ali drugih problemov je odprta skupnost na Discordu, kjer je možen pogovor z razvijalci programa in drugimi souporabniki programa. S programskim modulom oblak besed lahko hitro zberemo ključne besede, temo, literarno snov in motive besedila, ne da bi bilo potrebno prebrati celotno besedilo. Tematska analiza besedila je priročna pri pisanju vodenega spisa, saj nam pomaga pri ohranjanju spisa na izbrani temi. Sentimentalna ter besedilna analiza nam poda vpogled v pisateljevo vzdušje. Pri zbiranju ter procesiranju dodatnih dejstev avtorja, kot so vzdušja, zgodovinska dejstva ter letnice njegovih del, lahko te mapiramo in dobimo relativno natančen prevladujočih motivov, čustev, simbolov ter njihovega spreminjanja skozi ustvarjalčevo življenje in opus. Program lahko uporabimo tudi za pregled lastnega besedila in z njim lažje opazimo, če smo odtavali od glavne teme besedila. Analiza besedila s pomočjo podatkovnega rudarjenja v programu Orange se splača nadaljnjo raziskati, saj program omogoča široko izbiro načinov analize ter reprezentacije podatkov. Poleg tega uporabnik lahko razvije svoje module in se izogne ročnemu urejanju nekaterih besedil. Implementacija računalniške analize sentimentalnih besedil lahko pomaga tudi pri razvoju umetne inteligence in drugih pogovornih programov.

# Priloge

## Koda za menjavo besed oz. znakov v programskem jeziku python

Besede = []  
besede = sum(1 for line in open('Besede, ki jih hočemo zamenjati.txt', encoding="utf8"))

# Prešteje število besed v datoteki.

with open('Besede, ki jih hočemo zamenjati.txt', encoding="utf8") as f:   
 for line in f:   
 Besede.append(line.split(None, 1)[0])

# program prebere besede, ki smo jih vpisali v izbrani dokument   
  
Zamenjava = []  
zamenjava = sum(1 for line in open('Besede v katere hočemo zamenjati.txt', encoding="utf8"))  
with open('Besede, v katere hočemo zamenjati.txt', encoding="utf8") as f: for line in f:   
 Zamenjava.append(line.split(None, 1)[0])

# Program prebere besede, ki smo jih vpisali v izbrani document.  
  
  
začetno\_besedilo = open("Neprocesirano besedilo.txt",encoding="utf8")  
line = začetno\_besedilo.read().lower()   
začetno\_besedilo.close()

# Kopira vstavljeno besedilo  
  
končno\_besedilo = open("Končno besedilo.txt", "w", encoding="utf8")  
končno\_besedilo.write(line)   
končno\_besedilo.close() # Shrani kopijo vstavljenega besedila v končni dokument.  
  
for i in range(besede):  
 končno\_besedilo = open("Končno besedilo.txt", "r", encoding="utf8")  
 line = končno\_besedilo.read().lower().replace(Besede[i], Zamenjava[i]) # Odpre izbrani dokument, ga kopira in preuredi kopijo.  
 končno\_besedilo.close()  
 končno\_besedilo = open("Končno besedilo.txt", "w", encoding="utf8")   
 končno\_besedilo.write(line)   
 končno\_besedilo.close() # Odpre izbran dokument, izbriše vse, kar je notri, in prilepi preurejeno besedilo.

# Viri in literatura

## Literatura

1. Marko Zeman. ***Gradnik okolja Orange za dostop do podatkov iz baz UN Comtrade***. Diplomsko delo, Fakulteta za računalništvo in informatiko, 2017.
2. Darja Pavlič, Marjan Dović, Klemen Lah, Bernard Rovtar, Janja Preko. ***UMETNOST besede: berilo 2: učbenik za slovenščino – književnost v 2. letniku gimnazije in štiriletnih strokovnih šol.*** Mladinska knjiga Založba, d. d., Ljubljana, 2008
3. Orange, 10. feb 2021, ***Orange widget catalog,*** dostopno prek: <https://orangedatamining.com/widget-catalog/>
4. Alphabold, 10. feb 2021, ***Sentiment analysis – the lexicon based approach,*** dostopno prek: https://www.alphabold.com/sentiment-analysis-the-lexicon-based-approach/#:~:text=Rule%20based%20sentiment%20analysis%20refers,with%20their%20corresponding%20intensity%20measure.
5. Damir Kos, ***Uporaba orodja Orange za podatkovno rudarjenje.*** Diplomsko delo, Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko.
6. Matej Martinc, ***Učinkovito procesiranje naravnega jezika s Pythonom.*** Diplomsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.
7. Rok Martinc, ***Merjenje sentimenta na družabnem omrežju Twitter: izdelava orodja ter evaluacija.*** Magistrsko delo, Univerza v Ljubljani, Fakulteta za družbene vede.

## Viri

1. Cankar, Ivan. 1904. *Križ na gori.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Kri%C5%BE_na_gori> (16. 1. 2021).

1. Cankar, Ivan. 1904. *Hiša Marije pomočnice.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Hi%C5%A1a_Marije_Pomo%C4%8Dnice> (16. 1. 2021).

1. Cankar, Ivan. 1904. *Gospa Judit.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Gospa_Judit> (16. 1. 2021).

1. Cankar, Ivan. 1903/04. *Novela doktorja Grudna.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Novela_doktorja_Grudna> (16. 1. 2021).

1. Cankar, Ivan. 1903. *Življenje in smrt Petra Novljana.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/%C5%BDivljenje_in_smrt_Petra_Novljana> (16. 1. 2021).

1. Cankar, Ivan. 1903. *Ob zori.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Ob_zori> (16. 1. 2021).

1. Cankar, Ivan. 1902. *Na klancu.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Na_klancu> (16. 1. 2021).

1. Cankar, Ivan. 1901. *Tujci.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Tujci> (16. 1. 2021).

1. Cankar, Ivan. 1901. *Knjiga za lahkomiselne ljudi.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Knjiga_za_lahkomiselne_ljudi> (16. 1. 2021).

1. Cankar, Ivan. 1899. *Vinjene.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Vinjete> (16. 1. 2021).

1. Cankar, Ivan. 1922. *Romantične duše.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Romanti%C4%8Dne_du%C5%A1e> (16. 1. 2021).

1. Cankar, Ivan. 1912. *Lepa Vida.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Lepa_Vida_(Ivan_Cankar)> (16. 1. 2021).

1. Cankar, Ivan. 1910. *Hlapci.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Hlapci> (16. 1. 2021).

1. Cankar, Ivan. 1908. *Pohujšanje v dolini šentforjanski.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Pohuj%C5%A1anje_v_dolini_%C5%A1entflorjanski> (16. 1. 2021).

1. Cankar, Ivan. 1902. *Kralj na Betajnovi.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Kralj_na_Betajnovi> (16. 1. 2021).

1. Cankar, Ivan. 1901. *Za narodov blagor.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Za_narodov_blagor> (16. 1. 2021).

1. Cankar, Ivan. 1900. *Jakob Ruda.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Jakob_Ruda> (16. 1. 2021).

1. Cankar, Ivan. 1906. *Nina.* Wikivir. Dostopno prek: <https://sl.wikisource.org/wiki/Nina_(Ivan_Cankar)> (16. 1. 2021).
2. Cankar, Ivan. 1906. *Martin Kačur.* Wikivir. Dostopno prek: <https://sl.wikisource.org/wiki/Martin_Ka%C4%8Dur> (16. 1. 2021).
3. Cankar, Ivan. 1907. *V samoti.* Wikivir. Dostopno prek: <https://sl.wikisource.org/wiki/V_samoti> (16. 1. 2021).
4. Cankar, Ivan. 1905. *V mesečini.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/V_mese%C4%8Dini> (16. 1. 2021).

1. Cankar, Ivan. 1905. *Potepuh Marko in kralj Matjaž.* Wikivir. Dostopno prek:<https://sl.wikisource.org/wiki/Potepuh_Marko_in_kralj_Matja%C5%BE> (16. 1. 2021).
2. Prešeren, France. 1847. *Sanjalo se mi je, de v svetem raji.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Sanjalo_se_mi_je,_de_v_svetem_raji> (16. 1. 2021).

1. Prešeren, France. 1847. *Ni znal molitve žlahtnič trde glave.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Ni_znal_molitve_%C5%BElahtni%C4%8D_trde_glave> (16. 1. 2021).

1. Prešeren, France. 1847. *Sonetni venec.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Sonetni_venec> (16. 1. 2021).

1. Prešeren, France. 1847. *Je od veselga časa teklo leto.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Je_od_vesel%27ga_%C4%8Dasa_teklo_leto> (16. 1. 2021).

1. Prešeren, France. 1847. *Kupido! ti in tvoja lepa starka.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Kup%C3%ADdo!_t%C3%AD_in_tvoja_lepa_starka> (16. 1. 2021).

1. Prešeren, France. 1847. *Dve sestri videle so zmoti vdane.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Dve_sestri_videle_so_zmoti_vdane> (16. 1. 2021).

1. Prešeren, France. 1847. *Tak kakor hrepeni oko čolnarja.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Tak_kakor_hrepen%C3%AD_oko_%C4%8Dolnarja> (16. 1. 2021).

1. Prešeren, France. 1847. *Vrh sonca sije soncov cela čeda.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Vrh_sonca_sije_soncov_cela_%C4%8D%C3%A9da> (16. 1. 2021).

1. Prešeren, France. 1847. *Očetov naših imenitne dela.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/O%C4%8Detov_na%C5%A1ih_imenitne_dela> (16. 1. 2021).

1. Prešeren, France. 1847. *Kdor jih bere, vsak drugači pesmi moje sodi.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Kdor_jih_bere,_vsak_druga%C4%8Di_pesmi_moje_sodi> (16. 1. 2021).

1. Prešeren, France. 1847. *Al bo kal pognalo seme, kdor ga seje, sam ne ve.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Al_bo_kal_pognalo_seme,_kdor_ga_seje,_sam_ne_ve> (16. 1. 2021).

1. Prešeren, France. 1847. *Med otroci si igrala, draga! lani – čas hiti.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Med_otr%C3%B3ci_si_igrala,_draga!_lani_%E2%80%93_%C4%8Das_hiti> (16. 1. 2021).

1. Prešeren, France. 1847. *Draga! vem, kako pri tebi me opravljajo ženice.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Draga!_vem,_kako_pri_tebi_me_opravljajo_%C5%BEenice> (16. 1. 2021).

1. Prešeren, France. 1847. *žalostno komu neznana je resnica, de jo ljubim.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/%C5%BDalostna_kom%C3%BA_neznana_je_resnica,_de_jo_ljubim> (16. 1. 2021).

1. Prešeren, France. 1847. *Oči sem večkrat prašal, ali smem.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/O%C4%8Di_sem_ve%C4%8Dkrat_pra%C5%A1al,_ali_smem> (16. 1. 2021).

1. Prešeren, France. 1847. *Pesem moja je posoda tvojega imena.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Pesem_moja_je_posoda_tvojega_imena> (16. 1. 2021).

1. Prešeren, France. 1847. *Orglar.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Orglar> (16. 1. 2021).

1. Prešeren, France. 1847. *Ribič.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Ribi%C4%8D> (16. 1. 2021).

1. Prešeren, France. 1847. *Ženska zvestoba.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/%C5%BDenska_zvestoba>(16. 1. 2021).

1. Prešeren, France. 1847. *Neiztrohnjeno srce.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Neiztrohnjeno_srce>(16. 1. 2021).

1. Prešeren, France. 1847. *Prekop.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Prekop> (16. 1. 2021).

1. Prešeren, France. 1847. *Povodni mož.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Povodni_mo%C5%BE_(France_Pre%C5%A1eren)> (16. 1. 2021).

1. Prešeren, France. 1847. *Lenora.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Lenora> (16. 1. 2021).

1. Prešeren, France. 1847. *Zdravilo ljubezni.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Zdravilo_ljubezni> (16. 1. 2021).

1. Prešeren, France. 1847. *Judovsko dekle.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Judovsko_dekle> (16. 1. 2021).

1. Prešeren, France. 1847. *Turjaška Rozamunda.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Turja%C5%A1ka_Rozamunda> (16. 1. 2021).

1. Prešeren, France. 1847. *Dohtar.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Dohtar> (16. 1. 2021).

1. Prešeren, France. 1847. *Učenec.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/U%C4%8Denec> (16. 1. 2021).

1. Prešeren, France. 1847. *Hčere sveta.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/H%C4%8Dere_svet> (16. 1. 2021).

1. Prešeren, France. 1847. *Pevcu.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Pevcu> (16. 1. 2021).

1. Prešeren, France. 1847. *Nezakonska mati.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Nezakonska_mati_(France_Pre%C5%A1eren)> (16. 1. 2021).

1. Prešeren, France. 1847. *Zapuščena.* Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Zapu%C5%A1ena> (16. 1. 2021).

1. Prešeren, France. 1847. *Od železne ceste*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Od_%C5%BEelezne_ceste> (16. 1. 2021).

1. Prešeren, France. 1847. *V spomin Valentina Vodnika*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/V_spomin_Valentina_Vodnika> (16. 1. 2021).

1. Prešeren, France. 1847. *Zdravljica*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Zdravljica> (16. 1. 2021).

1. Prešeren, France. 1847. *Soldaška*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Solda%C5%A1ka> (16. 1. 2021).

1. Prešeren, France. 1847. *Mornar*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Mornar_(France_Pre%C5%A1eren)> (16. 1. 2021).

1. Prešeren, France. 1847. *Zgubljena vera*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Zgubljena_vera> (16. 1. 2021).

1. Prešeren, France. 1847. *Sila spomina*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Sila_spomina_(France_Pre%C5%A1eren)> (16. 1. 2021).

1. Prešeren, France. 1847. *K slovesu*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/K_slovesu> (16. 1. 2021).

1. Prešeren, France. 1847. *Ukazi*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Ukazi> (16. 1. 2021).

1. Prešeren, France. 1847. *Kam?*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Kam%3F_(France_Pre%C5%A1eren)> (16. 1. 2021).

1. Prešeren, France. 1847 *Prošnja*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Pro%C5%A1nja_(France_Pre%C5%A1eren)> (16. 1. 2021).

1. Prešeren, France. 1847. *Pod oknam*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Pod_oknam> (16. 1. 2021).

1. Prešeren, France. 1847. *Dekletam*. Wikivir. Dostopno prek:

<https://sl.wikisource.org/wiki/Dekletam> (16. 1. 2021).

1. Prešeren, France. 1847. *Strunam*. Wikivir. Dostopno prek:

https://sl.wikisource.org/wiki/Strunam (16. 1. 2021).

1. Prešeren, France. 1836. *Krst pri Savici*. Wikivir. Dostopno prek: <https://sl.wikisource.org/wiki/Krst_pri_Savici> (16. 1. 2021).
2. Linhart, Anton Tomaž. 1790. *Ta veseli dan ali Matiček se ženi*. Wikivir. Dostopno prek: <https://sl.wikisource.org/wiki/Ta_veseli_dan_ali_Mati%C4%8Dek_se_%C5%BEeni_(Zbrano_delo)> (25. 1. 2021).
3. Linhart, Anton Tomaž. 1790. *Županova Micka*. Wikivir. Dostopno prek: <https://sl.wikisource.org/wiki/%C5%BDupanova_Micka> (25. 1. 2021).