ZVEZA ZA TEHNIČNO KULTURO SLOVENIJE
53. SREČANJE MLADIH RAZISKOVALCEV SLOVENIJE 2021

RAZISKOVALNA NALOGA

PODATKOVNO RUDARJENJE INSTAGRAM OBJAV O MARIBORU

Računalništvo in informatika

Raziskovalno področje: računalništvo ali telekomunikacije
Avtorica: Gaja Đukanović Babič
Mentorja: Sašo Karakatič in Mitja Osojnik
Šola: II. gimnazija Maribor

Maribor, maj 2021

KAZALO

KAZALO TABEL

KAZALO GRAFOV

1 POVZETEK

V svoji raziskovalni nalogi bom s pomočjo podatkovnega rudarjenja pregledovala objave z značko *#maribor* na družbenem omrežju Instagram. Pri tem si bom pomagala z uporabo programske opreme Python, Jupyter Notebook, Seaborn, Pandas in Scikit-learn. Razvila bom sistem za branje in analizo Instagramovih objav. Želim ugotoviti, kateri dan je največ objav, ob kateri uri je največ objav, ali je več objav med delavnikom ali med vikendom, ob katerem delu dneva je največ objav. Prav tako me zanima, katere značke se pojavijo najpogosteje in koliko ima objava v povprečju všečkov in komentarjev. Iz pridobljenih podatkov bom izrisala različne grafe. Pripravila bom oblak besed in klasifikacijska odločitvena drevesa.

Ključne besede: podatkovno rudarjenje, Instagram, *#maribor*

In my research assignment, I will use data mining to look through posts with the hashtag #maribor on the social media Instagram. I will use the software Python, Jupyter Notebook, Seaborn, Pandas, and Scikit-learn. I will develop a system for reading and analysis of Instagram posts. I want to find out if more posts are posted during the weekdays or the weekend, at which time of the day and during which day and hour specifically. I am also interested in which hashtags are most often used and the average number of likes and comments on a post. I will make different graphs from the acquired data, prepare a word cloud and a classification decision tree.

Key words: data mining, Instagram, *#maribor*

2 ZAHVALA

Najlepše se zahvaljujem mentorjema za pomoč, nasvete in spodbudo pri pisanju raziskovalne naloge. Prav tako se zahvaljujem staršem in prijateljem za podporo med pisanjem naloge. Zahvaljujem se tudi profesorici slovenščine, ki je mojo nalogo lektorirala.

3 UVOD

V raziskovalni nalogi se bomo ukvarjali s podatkovnim rudarjenjem, a najprej moramo razumeti, kaj to sploh je. Podatkovno rudarjenje je proces analiziranja velikih zbirk podatkov z namenom pridobivanja novih informacij (Oxford Languages, b. d.).

Pri podatkovnem rudarjenju uporabljamo veliko različnih metod, da lahko obdelamo svoje podatke. Naštete so najpomembnejše:

1. uporaba statističnih tehnik,
2. nevronske mreže,
3. odločitvena drevesa,
4. strojno učenje,
5. uporaba umetne inteligence (16 Data Mining Techniques , b. d.).

Da pa lahko podatke res dobro obdelamo, moramo preko določenih korakov podatkovnega rudarjenja:

1. Predobdelava podatkov – izboljšanje kakovosti podatkov:
    - čiščenje podatkov – odstranjevanje nepopolnih, nepravilnih, netočnih ali nepomembnih podatkov,
    - integracija podatkov – postopek kombiniranja več podatkovnih virov (napisanih v različnih oblikah, formatih),
    - preoblikovanje podatkov – pretvarjanje podatkov v ustrezno obliko,
    - izbira podatkov – izbira podatke, ki jih želimo uporabiti pri raziskavi.
2. Podatkovno rudarjenje – odkriti vzorce iz velike količine podatkov.
3. Vrednotenje vzorcev – izbiranje zanimivih in pomembnih vzorcev.
4. Predstavitev znanja – pridobljeno znanje in ugotovite predstavimo in vizualiziramo (7 Stages of Data Mining, 2020).

V poplavi informacij nam podatkovno rudarjenje lahko zelo pomaga. Sama sem z njim pregledala Instagram objave, ki vsebujejo značko *#maribor*. Pri tem sem si pomagala z uporabo strojnega učenja, odločitvenimi drevesi in s statističnimi tehnikami.

4 HIPOTEZE

V raziskovalni nalogi sem postavila naslednje hipoteze:

1. Največ objav med delavnikom je med 16.00 in 24.00.
2. Največ objav med vikendom je med 8.00 in 16.00.
3. Posamezni dan v vikendu ima več objav kot posamezni dan v tednu.
4. Največ objav med delavnimi tedni je v petek.

4 TEORIJA – OBČANI NA DRUŽBENIH OMREŽJIH

4.1 Družbena omrežja

Družbena omrežja so katerokoli digitalno orodje, ki uporabnikom omogoča hitro ustvarjanje in skupno rabo vsebine z javnostjo. Temeljijo na svetovnem spletu in uporabnikom omogočajo hitro elektronsko komunikacijo vsebin. Vsebina vključuje osebne podatke, dokumente, slike, videoposnetke in zvočna sporočila ter zajema ideje, mnenja in informacije. Zajemajo široko paleto spletnih mest in aplikacij (ki jih pogosto uporabljamo tudi za pošiljanje sporočil). Uporabniki na njih sodelujejo preko računalnika, tablice, pametnega telefona ali druge elektronske naprave. Večinoma so ta specializirana za deljenje določene vrste vsebine (Dollarhide, b. d. in Hudson, b. d.).

Nekatera najbolj znana omrežja so Instagram (kratki videoposnetki in slike), YouTube (videoposnetki), Facebook, Twitter (kratka sporočila), TikTok (kratki videoposnetki), Pinterest, Snapchat in drugi (Dollarhide, b. d. in Hudson, b. d.).

Ko uporabniki sodelujejo na družbenih omrežjih, ustvarjajo zelo interaktivne platforme, preko katerih lahko posamezniki, skupnosti in organizacije delijo, soustvarjajo, razpravljajo, sodelujejo, spreminjajo vsebine, ki jih ustvarijo uporabniki in objavijo na spletu. Poleg tega se družbena omrežja uporabljajo za dokumentiranje spominov, raziskovanje, oglaševanje, spoznavanje ljudi in celo ustvarjanje novih prijateljstev. Prav tako omogoča rast idej pri ustvarjanju blogov, podkastov, videoposnetkov in iger (“Social Media”, b. d.).

Družbena omrežja se od tradicionalnih medijev (časopisov, televizije, radijskih oddaj ...) razlikujejo v mnogo pogledih, in sicer v kakovosti, dosegu, frekvenci, uporabnosti, neposrednosti in trajnosti. Poleg tega delujejo v dialoškem sistemu, za razliko od tradicionalnih medijev, ki delujejo po monološkem modelu – en vir, številni prejemniki (“Social Media”, b. d.).

Glede družbenih omrežij je moč opaziti širok nabor pozitivnih in negativnih učinkov. Družbena omrežja lahko pomagajo izboljšati občutek povezanosti posameznika z resničnimi ali s spletnimi skupnostmi in so lahko učinkovito komunikacijsko ali tržno orodje za korporacije,

podjetnike, neprofitne organizacije, zagovorniške skupine, politične stranke in vlade (“Social Media”, b. d.).

V okviru naloge sem se osredotočila na omrežje Instagram. To je družbeno omrežje v lasti Facebooka. Uporabnikom omogoča deljenje fotografij in videoposnetkov, dodajanje napisov, urejanje fotografij s filtri … Z drugimi uporabniki lahko vzpostavimo stike tako, da jim sledimo, všečkamo (angl. like), komentiramo, jih označimo ali pa se z njimi zasebno pogovarjamo (Safe.Si., b. d.).

Ustvarila sta ga Kevin Systrom in Mike Krieger. Javna uporaba se je začela leta 2010. Aplikacija uporabnikom omogoča nalaganje medijev, ki jih je možno urediti s filtri, organizirati z značkami (angl. hashtag) in dodati lokacijo. Objave se lahko delijo z javnostjo ali le s svojimi sledilci, ki smo jim omogočili dostop. Uporabniki lahko po Instagramu iščejo s pomočjo značk in lokacije in tako spoznajo vsebino, ki je v trendu. Uporabniki lahko všečkajo objave in sledijo drugim uporabnikom. S sledenjem drugim uporabnikom se njihove nove objave pojavljajo na zidu. Prav tako ima Instagram možnost dodajanja svojih slik na zgodbo (angl. Story), kjer slika ostane le 24 ur. Kritiki večkrat napadejo Instagram predvsem zaradi sprememb pravilnikov in vmesnikov, veliko je tudi očitkov o cenzuri ter o nezakonitih ali neprimernih vsebinah, ki jih nalagajo uporabniki. Instagram je postala četrta najbolj naložena mobilna aplikacija drugega desetletja enaindvajsetega stoletja (“Instagram”, b. d.).

4.2 Vpliv družbenih omrežij na upravljanje (občine oziroma države)

Državljani in občani vse pogosteje uporabljajo družbena omrežja, ker želijo preko njih izvedeti čim več informacij. Organizacije javne uprave pa vedno pogosteje uporabljajo družbena omrežja, čeprav je sprejemanje teh v javni upravi še vedno v zgodnji fazi. Javne uprave želijo povečati vključenost in zanimanje državljanov, ob tem pa uporabljajo dvosmerno komunikacijo. Po svetu ugotavljajo, da je uporaba družbenih omrežij v javni upravi pozitivna zaradi preglednosti in sodelovanja državljanov z javno upravo. Predvsem v smislu, da lahko državljani izražajo svoje mnenje, modrosti in izkušnje. S tem javna uprava pridobi povratne informacije državljanov. Prav tako lahko državljani preko družbenih omrežjih prispevajo svoje mnenje in razmišljajo o določenih temah, ki jih na družbenih omrežjih objavlja javna uprava.

Pozitivno je, da imajo državljani vse informacije zbrane na eni platformi, saj tako prihranijo čas z iskanjem informacij drugje. Ne samo za ljudi, tudi za javno upravo imajo družbena omrežja veliko prednosti. Z uporabo Facebooka in Twitterja lahko javna uprava hitro in poceni nagovarja svoje državljane. YouTube ali Pinterest pa sta zelo uporabna za nagovarjanje turistov. Vse več družbenih omrežij uporabljajo tudi politiki za doseganje ciljev, predvsem za pridobivanje podpore sledilcev in prenosa informacij ter komunikacije s svojim pripadniki (Saje, 2017).

Uporaba družbenih omrežij je najbolj popularna internetna aktivnost. V letu 2020 je družbena omrežja uporabljalo kar 3,6 milijarde ljudi. Ob tem se število uporabnikov hitro povečuje (Number of Social Media Users 2025, b. d.).

Takšna številčnost daje družbenim omrežjem izjemen pomen. In če bi javna uprava družbena omrežja ignorirala oz. ne bi objavljala, bi velik del državljanov ostal brez pomembnih informacij. Javna uprava uporablja družbena omrežja predvsem za spodbujanje udeležbe državljanov. Uporaba družbenih omrežij v javni upravi povečuje stopnjo zadovoljstva in zaupanje v upravljanje (Pavlič, 2019).

Družbena omrežja so za javno upravo zelo uporabna. Zanjo niso potrebni visoki investicijski stroški, prav tako pa je vzdrževanje profilov relativno enostavno. Da je profil na družbenih omrežjih uspešen, mora biti z državljani omogočena interakcija in komunikacija ter odzivnost na njihova stališča in mnenja. Javna uprava uporablja predvsem tista družbena omrežja, ki imajo največ uporabnikov. A uporabljajo tudi manj znana družbena omrežja, da dosežejo čim več državljanov. Uporaba družbenih omrežij v javni upravi zelo hitro narašča in večinoma jih javna uprava uporablja dnevno (Pavlič, 2019).

Družbenih omrežij ne uporabljajo le organi javne uprave, ampak so zelo priljubljena tudi med političnimi osebnostmi. Tako se večkrat zgodi, da ima politična osebnost več sledilcev kot vladni organi, ki jo ta zastopa (Pavlič, 2019). Kot primer lahko navedemo Boruta Pahorja, ki ima na Instagramu kar 128 tisoč sledilcev, medtem ko jih ima gov.si (Vlada Republike Slovenije) le 15 tisoč (Instagram, 18. 2. 2021). Politične osebnosti so bile med prvimi organi, ki so začeli uporabljati družbena omrežja. Družbena omrežja so zelo pomembna pri kampanjah,

saj z njimi pridobivajo glasove, podporo in sredstva. Glavne institucije držav sledijo političnim voditeljem in povečujejo uporabo družbenih omrežij. Na njih so aktivna ministrstva, javne agencije in institucije na lokalni in regionalni ravni države (Pavlič, 2019).

Družbena omrežja so zelo uporabna pri povezovanju javne uprave z mlajšimi državljani. Različni organi bi lahko uporabljali družbena omrežja za povezovanje z mlado populacijo še bolj, kot jo že. Ugotovitve so pokazale tudi, da več kot ljudje uporabljajo družbena omrežja, večja bo možnost, da bodo obiskali politična srečanja, koga poskušali prepričati o glasovanju in da bodo sami glasovali (Pavlič, 2019).

Kot je bilo že povedano, javna uprava uporablja omrežja za sporočanje informacij in za pridobitev povratnih informacij. Pri mreženju z državljani gre za uporabo družbenih omrežij za poslušanje državljanov. To omogoča javni upravi, da s pomočjo komentarjev in pogovorov z državljani pridobi vpogled v mnenja javnosti o pomembnih in relevantnih zadevah. Uporaba družbenih omrežij lahko prinese razvoj novih rešitev in izboljšanje ponudbe storitev. Z njimi je omogočeno, da javnost ni le stranka, temveč partner pri oblikovanju politike. Vloga državljanov vodi iz pasivne vloge do njegove aktivne vključenosti v skupno reševanje problemov, saj le ti prispevajo svoj čas, znanje in trud v zameno za večjo kontrolo nad končnimi odločitvami. Državljani lahko sodelujejo v različnih fazah zagotavljanja storitve: v fazi oblikovanja, fazi izvedbe in fazi spremljanja ter evalvacije. Pri tem pa je sodelovanje pri oblikovanju in zagotavljanju storitev med javno upravo in javnostjo uspešna le, če je vključeno večje število državljanov (Merlak, 2015).

5 METODOLOGIJA DELA

V naslednjem poglavju sledi opis metodologije, ki sem jo uporabila med raziskovanjem izbrane teme. Za namen raziskovanja objav na družbenih omrežjih sem razvila sistem za branje in analizo Instagram objav. Naslednje sekcije opisujejo uporabljene programske jezike, knjižnice in pristope v procesu razvoja.

5.1 Python, Jupyter Notebook, Seaborn, Pandas in Scikit-learn

Python je interpretni večnamenski programski jezik. Jezik je enostavno učljiv, je preprost in poudarja berljivost. Uporaben je za hiter razvoj aplikacij, spodbuja modularnost in podpira knjižnice. Ker programa ni treba prevajati, je razvojni cikel (pisanje programa, testiranje, razhroščevanje) zelo hiter. Že interpreter prestreže več napak kot nekateri drugi priljubljeni programski jeziki. Prav tako omogoča razhroščevanje programske kode vrstico po vrstico. Dostopen je na vseh večjih operacijskih sistemih (What Is Python? Executive Summary, b. d.). Python je ustvaril Guido van Rossum leta 1991, naprej pa ga je razvijal Python Software Foundation (History of Python, 2019).

Jupyter Notebook je odprtokodna spletna aplikacija, ki omogoča ustvarjanje in deljenje dokumentov, kot so navadno besedilo, vizualizacije, enačbe in programska koda, ki jo hkrati lahko spreminja več avtorjev. Uporaba zajema: čiščenje in transformacijo podatkov, numerične simulacije, statistično modeliranje, vizualizacijo podatkov, strojno učenje in veliko drugega. Podpira več kot 40 različnih programskih jezikov, predvsem Python in R (Project Jupyter, b. d.).

Seaborn je knjižnica za vizualizacijo podatkov Python, ki temelji na matplotlib in se tesno povezuje s podatkovnimi strukturami pandas. Ponuja vmesnik na visoki ravni za risanje privlačnih in informativnih statističnih grafov. Pomaga nam pri raziskovanju in razumevanju podatkov (An Introduction to Seaborn – Seaborn 0.11.1 Documentation, b. d.).

Pandas je hitro, zmogljivo, prilagodljivo in enostavno odprtokodno orodje za analizo in manipulacijo podatkov, narejeno na podlagi jezika Python. Razvoj pandasa se je začel leta 2008

pri AQR Capital Management, 2009 pa se je začela tudi javno uporabljati (Pandas – Python Data Analysis Library, b. d.). Pandas se uporablja predvsem za analizo podatkov. Omogoča uvoz podatkov iz datotek različnih formatov, kot so csv, JSON, SQL, Microsoft Excel. Pandas omogoča različne operacije oz. manipulacije s podatki, kot so združevanje, izbiranje, preoblikovanje, tudi čiščenje podatkov in funkcije premeščanja podatkov (“Pandas (Software)”, 2020).

Scikit-learn je knjižnica za strojno učenje v Pythonu. Ponuja izbor orodij za strojno učenje in statistično modeliranje, vključno s klasifikacijo, z regresijo, združevanjem in zmanjševanjem dimenzij (faktorsko analizo) prek vmesnika konsistence v Pythonu. Scikit-learn temelji na NumPy, SciPy in Matplotlib. Scikit-learn je v sklopu Googlovega projekta leta 2007 razvil David Cournapeau. V letu 2010 so F. Pedregosa, G. Varoquaux, A. Gramfort in V. Michael iz FIRCA ta projekt dvignili na novo raven in objavili prvo javno različico (Scikit Learn – Intorduction, b. d.).

5.2 Pridobivanje podatkov iz Instagrama

V tem podpoglavju je predstavljena in opisana programska koda, ki je bila razvita za namen pridobivanja objav iz omrežja Instagram.

```
import time
from instagramtagcrawler import InstagramTagCrawler
```

Uvozimo knjižnico `time` za delo s časom ter iz datoteke `instagramtagcrawler` uvozimo `InstagramTagCrawler`.

```
pavza = 3600
```

Za spremenljivko `pavza` smo določili vrednost 3600 (3600 sekund = 1 ura).

```
while True:
     print('Pridobivam objave iz Instagrama.')

     crawler = InstagramTagCrawler('maribor')
     medias = crawler.get_posts()
     datoteka = open("instagram_maribor.csv", "a")

     for media in medias:
               datoteka.write(str(media) + '\n')
     datoteka.close()
     time.sleep(pavza)
```

Naredimo neskončno zanko. Uporabimo že razvito aplikacijo `InstagramTagCrawler`, ki iz Instagrama pridobi vse nove objave, ki so označene s *#maribor* in objave shrani v `medias`. Objave z značko *#maribor* smo ob vsakem ponovnem zagonu dodali na konec datoteke instagram_maribor.csv. Za vsako v medij shranjeno objavo s `str()` pretvorimo celotno objavo v niz znakov. Na koncu dodamo `'\n'` ter s tem vsilimo, da bo vsaka objava zapisana v svoji

vrstici. Tako narejeno datoteko zapremo. Po opravljenem delu program zaspi za 3600 sekund (oz. eno uro).

5.3 Analiza podatkov v Pythonu

V tej sekciji sledi pregled programske kode, ki je bila uporabljena za namen analize pridobljenih Instagramovih objav.

```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
from datetime import datetime, time
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from PIL import Image
```

Uvozimo knjižnico pandas (kot `pd`), numpy (kot `np`), seaborn (kot `sns`), plyplot (kot `plt`), `datetime`, `time`, `WordCloud`, `STOPWORDS`, `ImageColorGenerator` in `Image`.

```
from google.colab import drive
drive.mount('/content/drive')
```

Google Colabu omogočimo dostop do našega Google Driva in mu dovolimo , da dostopa do datotek iz našega Driva.

```
maribor = pd.read_csv("drive/My Drive/Podatkovno
rudarjenje/instagram_maribor.csv")
```

Podatke iz datoteke drive/My Drive/Podatkovno rudarjenje/instagram_maribor.csv shranimo v spremenljivko `maribor`.

```
time = maribor["timestamp"]
maribor["time"] = time
```

```
maribor['time'] = pd.to_datetime(maribor ['time'], unit='s')
```

Naredimo nov stolpec `time`, v katerem so shranjeni podatki iz timestampa, le-te pa spremenimo v ljudem lažje berljivo obliko. Za računalniški čas velja dogovor, da ga štejemo od prvega januarja 1970 po časovnem standardu UTC (Greenwitch mean time). Ta dogovor je zelo koristen, saj omogoča komunikacijo med različnimi računalniškimi sistemi po celem svetu. Ni pa berljiv ljudem. Zato sem ga spremenila v nam berljivo obliko (Epoch & Unix Timestamp Conversion Tools, b. d.).

```
day_of_week = maribor["time"]
maribor["day_of_week"] = day_of_week
maribor['day_of_week'] = maribor['day_of_week'].dt.strftime('%A')
```

Naredimo nov stolpec `day_of_week`, v katerega shranimo podatke iz stolpca `time`. Iz zapisa prepišemo le, kateri dan v tednu je.

```
maribor["weekend"] = np.where((maribor["time"].dt.dayofweek) <
  5,0,1)
```

Dodamo stolpec, ki ga poimenujemo `weekend`. V njega shranimo podatke iz `time`, ki so shranjeni v `maribor`, nato si pomagamo s spremenljivko `dayofweek,` ki ima dve vrednosti: 1 za vikend in 2 za delovnik.

```
maribor["daytime"] = 1 + np.where ((maribor["time"].dt.hour) <
  8,0,1) + np.where ((maribor["time"].dt.hour) < 16,0,1) #1 -
  00:00-08:00, 2 - 08:00-16:00, 3 - 16:00-24:00
```

Dodamo stolpec, ki ga poimenujemo `daytime`, v njega shranimo podatke iz stolpca `time`, ki so shranjeni v spremenljivki `maribor`, nato si pomagamo s spremenljivko `hour`, ki lahko zavzame naslednje vrednosti: 1, če je ura med 00.00 in 8.00, 2 če je ura med 8.00 in 16.00 in 3 če je ura med 16.00 in 24.00.

```
month = maribor["time"]
```

```
maribor["month"] = month
maribor["month"] = maribor["month"].dt.strftime("%m")
```

Naredimo nov stolpec `month`, v katerega shranimo podatke iz spremenljivke time. Iz zapisa vzamemo in prepišemo le, kateri mesec je.

```
hour = maribor["time"]
maribor["hour"] = hour
maribor["hour"] = maribor["hour"].dt.strftime("%H")
```

Naredimo nov stolpec `hour` in vanj shranimo podatke iz spremenljivke time. Iz zapisa povzamemo le, koliko je ura.

```
maribor['tags'] = maribor['tags'].str.decode('unicode_escape')
maribor['caption'] =
  maribor['caption'].str.decode('unicode_escape')
```

Neznane znake v stolpcu značke in `caption` spremenimo v šumnike in emotikone.

```
maribor.head()
```

Programu ukažemo, da izpiše prvih par vrstic.

```
c = maribor.tags.str.split(expand=True).stack().value_counts()
print(c)
```

Pod spremenljivko `c` shranimo podatke iz stolpca `tags` (značke), ki jemlje podatke iz spremenljivke `maribor` ter nato prešteje, v koliko objavah je napisana določena značka. Nato `c` izpišemo.

```
e = c.drop(labels["kranj","novomesto","europe","domzale","EU",
  "ifeelslovenia","ormož","streetsofljubljana","europetrip",
```

"mycountry","moskva","cyprus","slovenski","ifeelslovenia",
"europeanstyle","VisitLjubljana","BLED","slovakia","brezice","lena
rt","domžale","Ljubljanatimes","ptuj","bled",
"novagorica","sarajevo","beograd","igmaribor","mojmaribor",
"mariborinfo","visitmaribor#maribor","igslovenija",
"lovemaribor","theslovenia","ifeelslovenia","Slovenia",
"serbien","slovenialovers","MARIBOR","pula","portugal",
"rogaškaslatina","sloveniagram#instagramiesslovenia",
"bratislava","liverpool","novisad#Slovenia","maribor gram",
"ljubljanatimes","#ljubljana","europetravel","lesce",
"albanien","gorica","kocevje","maribor#celje","radgona",
"belakrajina","maribor.","poljcane","albania","gorenjska",
"slovenskekonjice","ifeelslo","fallinginslovenia","kosovo",
"usa","kroatien","slovenia_nature",".","igslo","ljubljana",
"ljubljama","macedonia","istra","eurotrip","slovenjgradec",
"slovenia360","goriskabrda","LJUBLJANA","dubrovnik","vipava","slov
eniawonders","Maribormojemesto","sloveniangirl",
"visitslovenija","slovenian","murskasobota#belakrajina",
"radenci","visit_maribor","visiteurope","kamnica",
"sloveniaoutdoors","slovenia","štajerska","Maribor",
"sloveniatravel","Slovenija","Stajerska","zurich","abudhabi","slow
enia","jesenice","skofjaloka","european","iloveslovenia","posavje"
,"croatia","#ifeelslovenia","bosna","banjaluka",
"barcelona","koreja","monaco","primorska","budva","kamnik",
"prekmurje","duplek","slo","austria","slovenska","italia",
"Koper","kras","radovljica","istra","pragersko","#maribor",
"zadar","tuzla","krsko","serbia","postojna","vienna",
"pobrežje","madrid","london","Ljubljana","varsava",
"TastingMaribor","graz","mostar","viena","lakebled","dubai",
"ljubljanainyourpocket","streetsofmaribor","velenje",
"hrvatska","berlin","bohinj","slovensko","igljubljana",
"visitmaribor","zagreb","kranjskagora","igslovenia",
"visitslovenia","ifeelslovenia","slovenija","feelslovenija",
"slovenia_ig","celje","ljubljana","maribor",
"ljubljanamoments","slovenia","ifeelsLOVEnia",

```
"mariboristhefuture","mariborbychoice","mariborslovenia",
"ljubljanacity","visitljubljana","feelslovenia",
"murskasobota","celjeslovenia","sentjur","loveslovenia",
"travelslovenia","koper", "kopertravel", "maribor_si",
"mariborcity", "stajerska", "tastingmaribor", "mojaslovenija"])
```

Pod spremenljivko `e` shranimo spremenljivko `c`, iz katere odstranimo vse zgoraj naštete značke, saj nimajo semantičnega pomena.

```
podatki_grafa = e[0:20]
plt.figure(figsize=(10,5))
graf = sns.barplot(x = podatki_grafa.index, y = podatki_grafa)
plt.ylabel('number_of_tags')
plt.xlabel("tags")
plt.title("#maribor")
plt.xticks(
    rotation=45,
    horizontalalignment='right'
)
```

Naredimo graf, na katerem je izpisanih 20 najpogostejših značk iz spremenljivke `e`. Graf uredimo tako, da ga povečamo, poimenujemo osi x in y ter graf in napise na osi x obrnemo za 45° v desno.

```
graph_hour = maribor["time"].dt.hour
sns.histplot(graph_hour)
```

V spremenljivko `graph_hour` shranimo podatke iz spremenljivke `hour` ter naredimo graf, ki nam pokaže, koliko objav je bilo objavljenih ob določeni uri.

```
graph_day_of_week = maribor["day_of_week"]
plt.figure(figsize=(7,4))
day_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday',
'Friday', 'Saturday', 'Sunday']
```

```
sns.countplot(graph_day_of_week, order=day_order)
```

V spremenljivki `graph_day_of_week` shranimo podatke iz `day_of_week` in naredimo graf, ki prikazuje, koliko objav je bilo objavljenih ob določenih dnevih. Grafu še pred tem spremenimo velikost in dneve razporedimo v pravilni vrstni red (od ponedeljka do nedelje).

```
graph_daytime = maribor["daytime"]
sns.histplot(maribor,
             x = maribor["day_of_week"],
             y = maribor["time"].dt.hour)
```

V spremenljivko `graph_daytime` shranimo podatke iz `daytime` in naredimo graf, ki prikazuje pogostost objav na določen dan ob določeni uri.

```
ax = sns.countplot(x = maribor["weekend"])
```

Naredimo graf, ki izriše, koliko objav je bilo objavljenih med vikendom in koliko med tednom. Podatke vzamemo iz `weekend`.

```
sns.histplot(maribor,
             x = maribor["time"].dt.hour,
             hue = maribor["weekend"])
```

Naredimo graf, ki izriše, koliko objav je bilo objavljenih ob določeni uri med vikendom in koliko med tednom.

```
plt.figure(figsize=(15,15))
sns.histplot(maribor,
             x = maribor["time"].dt.hour,
             hue = maribor["day_of_week"])
```

Naredimo graf, ki izriše, koliko objav je bilo objavljenih ob določeni uri glede na dan v tednu.

```
text = maribor.tags[30]

wordcloud = WordCloud(max_font_size=40,
                      max_words=100,
                      background_color="white").generate(text)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Naredimo oblak besed (angl. wordcloud), ki izpiše imena značk. Graf uredimo tako, da določimo maksimalno velikost pisave, največje število besed in barvo ozadja.

```
stopwords = set(STOPWORDS)
stopwords.update(["ljubljana", "koper", "novomesto"," celje",
  "bled", "logatec", "maribor", "bohinj", "murskasobota",
  "kranjskagora"])

mask = np.array(Image.open("drive/MyDrive/Podatkovno
  rudarjenje/mask2 (1).png"))

def transform_format(val):
    if val == 0:
        return 255
    else:
        return val

transformed_mask = np.ndarray((mask.shape[0],mask.shape[1]),
                              np.int32)
for i in range(len(mask)):
   transformed_mask[i] = list(map(transform_format, mask[i]))

wc = WordCloud(background_color="white",
               max_words=1000,
```

```
                mask=transformed_mask,
                stopwords=stopwords,
                contour_width=3,
                contour_color='firebrick')
wc.generate(text)
wc.to_file("drive/MyDrive/Podatkovno rudarjenje/srček.png")
plt.figure(figsize=[20,10])
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Najprej med `stopwords` dodamo besede, ki nimajo semantičnega pomena in jih ne bomo vključili. Naložimo sliko srca in jo oblikujemo v pravo velikost. Naredimo oblak besed, v katerem so značke napisane tako, da so omejene z obliko srca. Oblak besed še oblikujemo in shranimo.

```
mask2 = np.array(Image.open("/content/drive/MyDrive/Podatkovno
  rudarjenje/instagram 9.png"))

def transform_format(val):
    if val == 0:
        return 255
    else:
        return val

transformed_mask2 = np.ndarray((mask2.shape[0],mask2.shape[1]),
                               np.int32)

for i in range(len(mask2)):
    transformed_mask2[i] = list(map(transform_format, mask2[i]))

wc = WordCloud(background_color="white",
               max_words=1000,
               mask=transformed_mask2)
```

```
wc.generate(text)
wc.to_file("/content/drive/MyDrive/Podatkovno
  rudarjenje/instagram11.png")

plt.figure(figsize=[10,20])
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Naredimo zelo podoben oblak besed kot prej, le da ne odstranimo nobenih besed, za obliko vzamemo logotip Instagrama in obliko loga popolnoma skrijemo, tako da se vidijo le besede.

```
mb_brez_manjkajocih = maribor.dropna()
mb_brez_manjkajocih.head()
```

Sedaj bomo naredili klasifikacijska odločitvena drevesa, v ta namen pa moramo nazaj dodati značke, ki smo jih v prejšnjih korakih odstranili.

```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer()
pojav_tagov = vectorizer.fit_transform(mb_brez_manjkajocih.tags)

vectorizer.get_feature_names()[100:120]
```

Uvozimo knjižnico `CountVectorizer`. Nato izračunamo pojavnost značk in natisnemo njihova imena, da pregledamo, katere so bile zajete.

```
plt.rcParams["figure.figsize"] = (20,10)

from sklearn import tree
clf = tree.DecisionTreeClassifier(max_depth=5)
clf = clf.fit(tf_idf, mb_brez_manjkajocih.weekend)
```

```
tree.plot_tree(clf,
               feature_names=vectorizer.get_feature_names(),
               class_names=['Delovnik', 'Vikend'],
               filled=True)
```

Naredimo klasifikacijsko odločitveno drevo, ki ga zgradimo na podlagi svojih podatkov. Omejimo ga na globino 5. Uredimo ga tako, da sliko povečamo. Povemo, kako naj poimenujemo besede, kako naj poimenujemo razrede in ali naj določeno polje pobarva ali ne. To klasifikacijsko drevo odloča, ali je vikend ali delavnik.

```
clf = clf.fit(pojav_tagov, mb_brez_manjkajocih.daytime)

tree.plot_tree(clf,
               feature_names=vectorizer.get_feature_names(),
               class_names=['00.00-08.00',
                            '08.00-16.00',
                            '16.00-24.00'],
               filled=True)
```

Naredimo klasifikacijsko odločitveno drevo, ki ga zgradimo na podlagi svojih podatkov. To klasifikacijsko drevo odloča, kateri del dneva je.

```
clf = clf.fit(pojav_tagov, mb_brez_manjkajocih.day_of_week)

tree.plot_tree(clf,
               feature_names=vectorizer.get_feature_names(),
               class_names=['Ponedeljek', 'Torek', 'Sreda',
                            'Četrtek', 'Petek', 'Sobota',
                            'Nedelja'],
               filled=True)
```

Naredimo klasifikacijsko odločitveno drevo, ki ga zgradimo na podlagi svojih podatkov. To klasifikacijsko drevo odloča, kateri dan v tednu je.

```
maribor.liked.mean(axis = 0)
```

```
maribor.liked.min(axis = 0)
maribor.liked.max(axis = 0)
maribor.liked.std(axis = 0)
```

Izračunamo povprečje, minimalno število, maksimalno število in standardni odklon glede na število všečkov vseh objav.

```
maribor.commented.mean(axis = 0)
maribor.commented.min(axis = 0)
maribor.commented.max(axis = 0)
maribor.commented.std(axis = 0)
```

Izračunamo povprečje, minimalno število, maksimalno število in standardni odklon glede na število komentarjev vseh objav.

6 REZULTATI

Tabela 1 – prvih 5 objav (celotna tabela)

	id	code	timestamp	owner_id	liked	commented	display_url	caption	tags	time	day_of_week	weekend	daytime	month	hour
0	2434850552469936181	CHKU61VIlg1	1604476826	221696631	8	0	https://scontent-ham3-1.cdninstagram.com/v/t51...	No tudi že google ve kje so najboljše pizze, b...	rondosevnica sevnica posavje hrana okusno slov...	2020-11-04 08:00:26	Wednesday	0	2	11	08
1	2434850338223964771	CHKU3tzFEJj	1604476800	21446211288	4	0	https://scontent-ham3-1.cdninstagram.com/v/t51...	🌞\n.\n.\n.\n.\n.\n.\n#osebnarast #afirmacije #...	osebnarast afirmacije celje ljubljana maribor ...	2020-11-04 08:00:00	Wednesday	0	2	11	08
2	2434848813829434750	CHKUhiGFa1-	1604476618	32113954843	2	0	https://scontent-ham3-1.cdninstagram.com/v/t51...	Veseli nas, da smo podpisali Koncesijsko pogod...	komunala komunalaodtok ciscenje Ob maribor vis...	2020-11-04 07:56:58	Wednesday	0	1	11	07
3	2434848489358060716	CHKUcz6FWys	1604476580	7725223234	9	1	https://scontent-ham3-1.cdninstagram.com/v/t51...	Vegan lovers quickly found this gem and are no...	NaN	2020-11-04 07:56:20	Wednesday	0	1	11	07
4	2434836602139600139	CHKRv1EpmUL	1604475473	52350628	103	0	https://scontent-ham3-1.cdninstagram.com/v/t51...	Skrij me v svojo dlan, svojo mehko dlan, svojo...	oneofmyfavorites alyamusic alya alyalive conce...	2020-11-04 07:37:53	Wednesday	0	1	11	07

Ko v program napišemo `maribor.head()`, dobimo zgornjo tabelo. Dobimo prvih 5 objav. Ker je tabela zelo velika, sem jo razdelila na več manjših tabel.

Tabela 2 – prvih 5 objav (id, code, tiemstamp, owner_id)

	id	code	timestamp	Owner_id
0	2434850552469936181	CHKU61VIlg1	1604476826	221696631
1	2434850338223964771	CHKU3tzFEJj	1604476800	21446211288
2	2434848813829434750	CHKUhiGFa1-	1604476618	32113954843
3	2434848489358060716	CHKUcz6FWys	1604476580	7725223234
4	2434836602139600139	CHKRv1EpmUL	1604475473	52350628

Kot smo že ugotovili, je `timestamp` čas, ki predstavlja vsoto sekund od 1. 1. 1970 po času UTC. `Owner_id` (osebna identifikacija) je številka, pod katero je shranjen profil osebe.

Tabela 3 – prvih 5 objav (liked, commented, display_url)

	liked	commented	display_url
0	8	0	https://scontent-ham3-1.cdninstagram.com/v/t51...
1	4	0	https://scontent-ham3-1.cdninstagram.com/v/t51...
2	2	0	https://scontent-ham3-1.cdninstagram.com/v/t51..
3	9	1	https://scontent-ham3-1.cdninstagram.com/v/t51...
4	103	0	https://scontent-ham3-1.cdninstagram.com/v/t51...

Pod `liked` najdemo, koliko ljudi je všečkalo objavo. Pod `commented` najdemo, koliko ljudi je komentiralo objavo. Pod `display_url` najdemo povezavo do objave. V preglednici je napisan le del povezave.

Tabela 4 – prvih 5 objav (besedilo pod objavo, značke)

	caption	tags
0	No tudi že google ve, kje so najboljše pizze, b...	rondosevnica sevnica posavje hrana okusno slov ...
1	🏵\n.\n.\n.\n.\n.\n.\n#osebnarast #afirmacije #...	osebnarast afirmacije celje ljubljana maribor ...
2	Veseli nas, da smo podpisali Koncesijsko pogod...	komunala komunalaodtok ciscenje Ob maribor vis ...
3	Vegan lovers quickly found this gem and are no...	NaN
4	Skrij me v svojo dlan, svojo mehko dlan, svojo...	oneofmyfavorites alyamusic alya alyalive conce ...

Pod `caption` (besedilo pod objavo) najdemo besedilo, ki ga je oseba napisala pod svojo objavo. Če pod objavo piše `NaN`, pomeni, da oseba ni napisala besedila pod objavo. V preglednici je napisan le del besedila pod objavo. `Tags` so značke, ki jih je oseba napisala pod objavo. Če je oseba napisala *#hrana*, se bo v tabelo pod `tags` (značke) izpisala `hrana`. Če pod objavo piše `NaN`, pomeni, da oseba ni napisala nobene značke pod besedilo. V preglednici je napisan le del značk.

Tabela 5 – prvih 5 objav (time, day_of_week, weekend, daytime, month, hour)

	time	day_of_week	weekend	daytime	month	hour
0	2020-11-04 08:00:26	Wednesday	0	2	11	8
1	2020-11-04 08:00:00	Wednesday	0	2	11	8
2	2020-11-04 07:56:58	Wednesday	0	1	11	8
3	2020-11-04 07:56:20	Wednesday	0	1	11	7
4	2020-11-04 07:37:53	Wednesday	0	1	11	7

Sedaj pa prehajamo do dela tabele, za katero je bilo potrebno še dodatno programiranje. Pod time je shranjen čas, ob katerem je bila objava objavljena. Napisana je po vzorcu leto-mesec-dan, ura:minute:sekunde. Pod `day_of_week` je napisano, kateri dan v tednu je bila objava objavljena. Pod `weekend` (vikend) je napisano, ali je vikend ali ne. Številka 1 pomeni, da je vikend, 0 pa, da je delavnik. Pod `month` je napisano, kateri mesec je. Mesec je napisan s številko. Pod `daytime` je napisano, kateri del dneva je, pri čemer 1 pomeni od polnoči do osmih, 2 pomeni od osmih do šestnajstih in 3 od šestnajstih do polnoči. Pod `hour` je napisano, koliko je ura. Pri tem gre ura od 00.00 (polnoč) do 23.00 (enajstih zvečer).

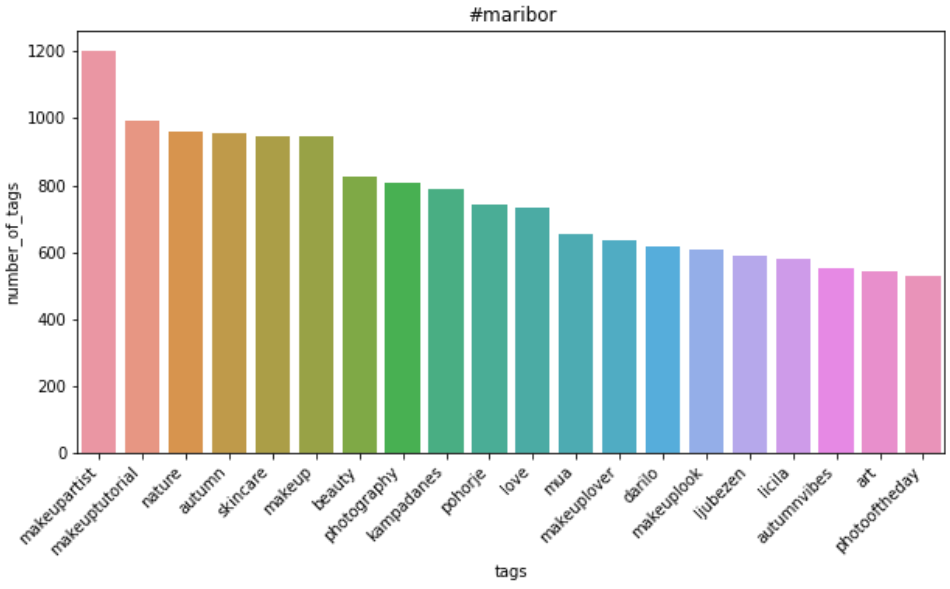
Tabela 6 – pogostost značk

maribor	15479
ljubljana	6428
slovenia	6348
slovenija	4211
celje	3628
...	
kupujmoslovensko	1
kapljicaolja	1
quattroworld	1
fotodestages	1
club	1

Length: 11998, dtype: int64

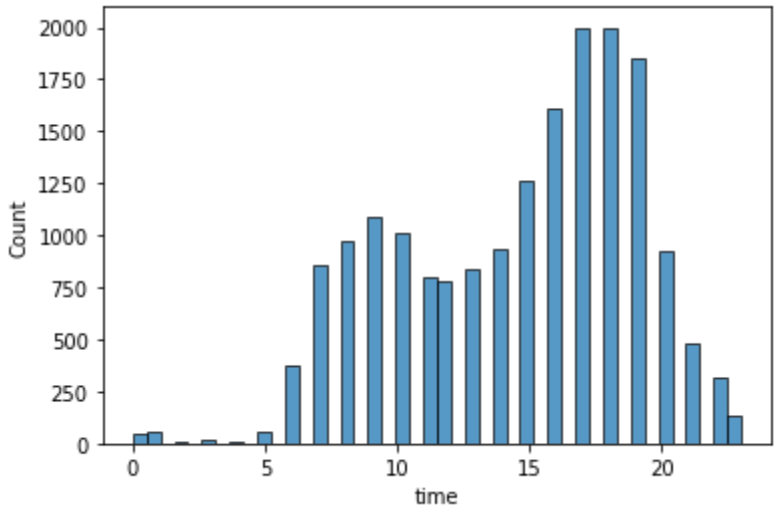
V tabeli je izpisanih pet najpogostejših značk, ki se pojavijo skupaj z značko *#maribor*, in pet značk, ki se pojavijo le enkrat. Med 18426 objavami, ki so vključene v raziskavo, najdemo kar 11998 različnih značk.

Graf 1 – najpogostejše značke skupaj z značko *#maribor*



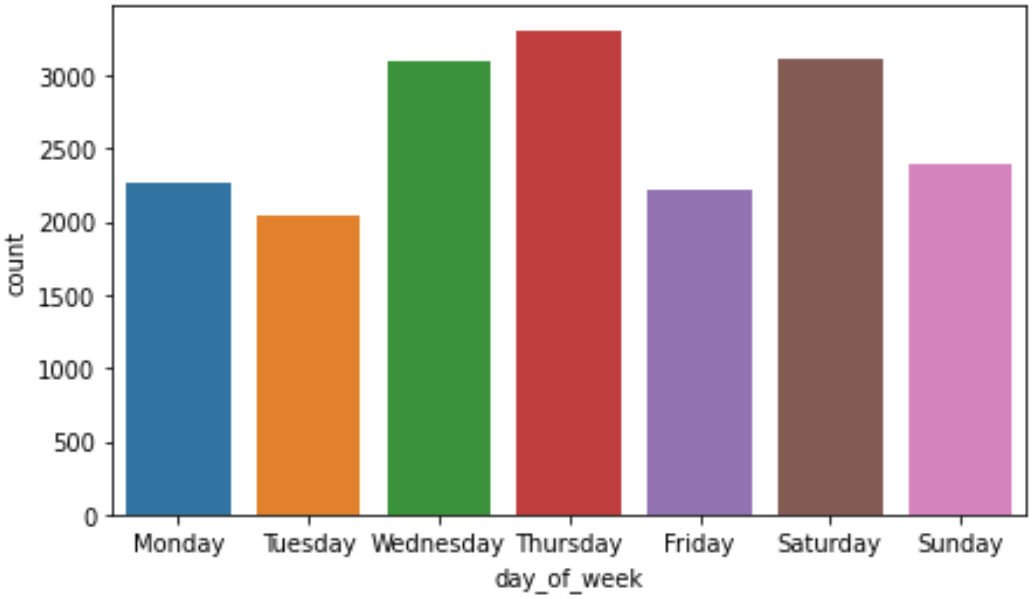
Na grafu je prikazanih dvajset najpogostejših značk, ki so bile uporabljene skupaj z značko *#maribor*. Med temi je najpogostejša *#makeupartist*, ki se pojavi skoraj tisoč dvesto krat. Opazimo lahko tudi, da se najpogosteje pojavijo značke z ličili, in sicer *#makeupartist*, *#makeuptutorial*, *#skincare*, *#mua*, *#makeuplover*, *#makeuplook* in *#licila*. Opazimo lahko, da se pojavljajo še značke, povezane z naravo, izleti, lepoto, ljubeznijo, jesenjo in umetnostjo.

Graf 2 – pogostost pojavljanja značke *#maribor* glede na uro



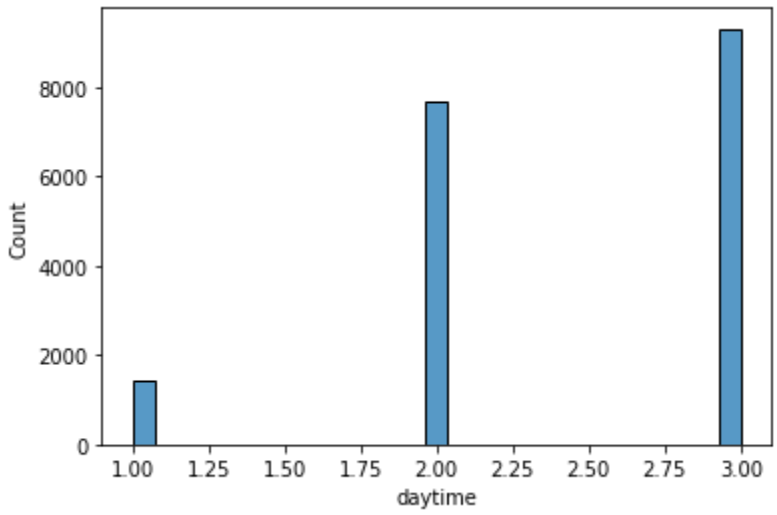
Na sliki vidimo graf, ki prikazuje, ob kateri uri je objav največ. Iz grafa lahko razberemo, da je najmanj objav od 00.00 do 5.00. Od 5.00 naprej se število objav veča, vse do 10.00, ko se začne manjšati. Manjšanje poteka do 12.00. Od 13.00 do 17.00 močno zraste, in tako imamo med 17.00 in 18.00 največ objav. Ob 19.00 število objav upade, nato pa močno pada vse do polnoči.

Graf 3 – pogostost pojavljanja značke *#maribor* glede na dan



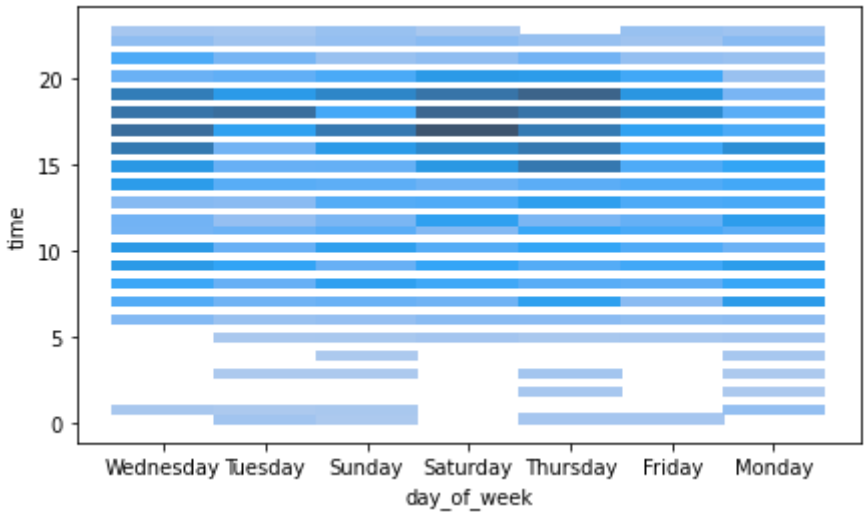
Iz grafa lahko vidimo, da je najmanj objav v torek in največ v četrtek. Če jih razporedimo od najmanjšega do največjega, je vrsti red takšen: torek, petek, ponedeljek, nedelja, sreda, sobota, četrtek.

Graf 4 – pogostost pojavljanja značke *#maribor* glede na čas v dnevu



Iz grafa lahko vidimo, da je najmanj objav med 00.00 in 8.00. Največ objav pa je med 16.00 in 24.00.

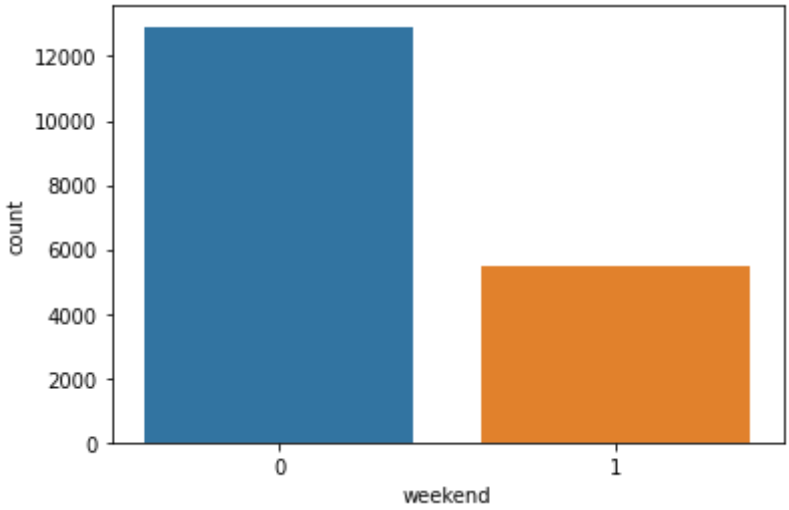
Graf 5 – pogostost pojavljanja značke *#maribor* glede na dan in uro objave



Na grafu je prikazana pogostost objav glede na dan in uro. Močneje je modra barva poudarjena, več je objav. Če je barva bela, objav takrat ni bilo. Na grafu lahko vidimo, da je bilo največ

objav v soboto, ob 17.00. Iz grafa lahko prav tako opazimo, da je med vsemi dnevi največ objav nekje med 16.00 in 17.00, najmanj pa med 00.00 in 5.00.

Graf 6 – pogostost pojavljanja značke *#maribor* glede na to, ali je vikend ali delavnik



Na grafu lahko vidimo, da je več objav med delavnikom (0) kot pa med vikendom (1), kar ne preseneča, saj je med delavnikom več dni kot med vikendom.

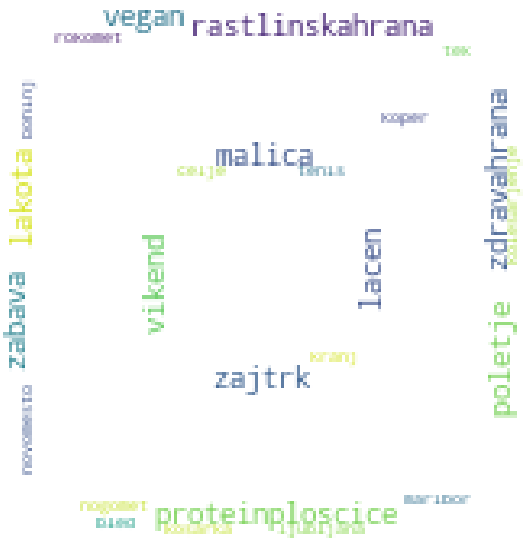
Graf 7 – oblak besed (oblika pravokotnika)

Ta graf se imenuje oblak besed. V grafu je izpisanih nekaj značk.

Graf 8 – oblak besed (oblika srca)

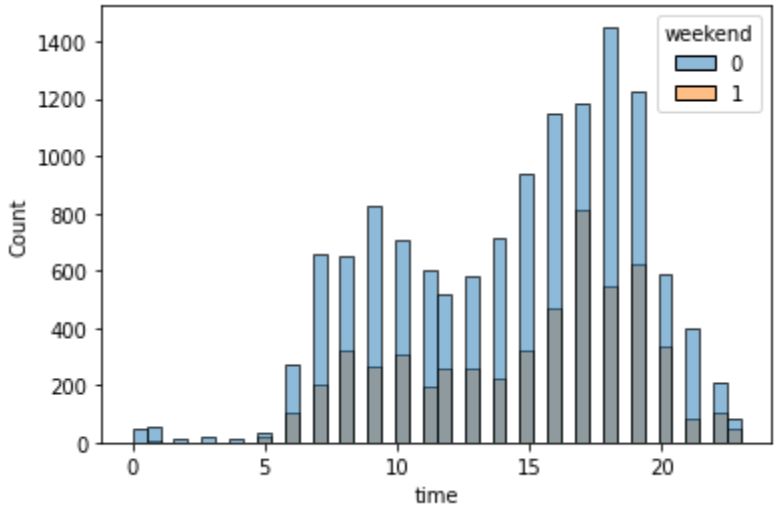
To je oblak besed, v katerem so značke zbrane v srcu.

Graf 9 – oblak besed (oblika Instagramovega logotipa)



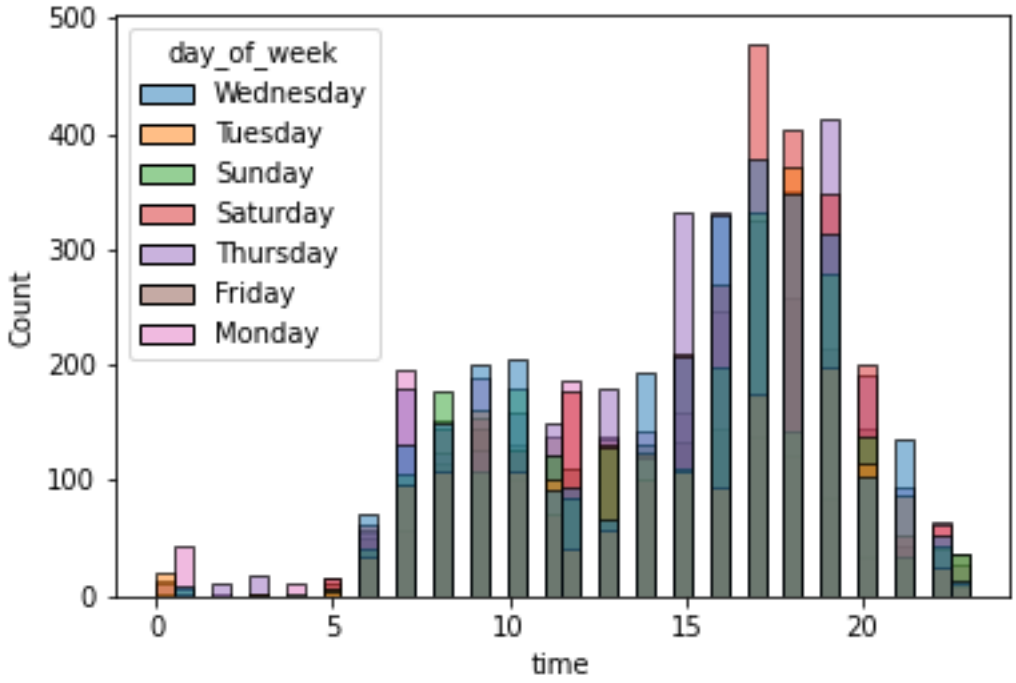
To je oblak besed, v katerem so značke zbrane v obliki Instagramovega logotipa.

Graf 10 – pogostost pojavljanja značke *#maribor* glede na uro in ali je vikend oz. delavnik



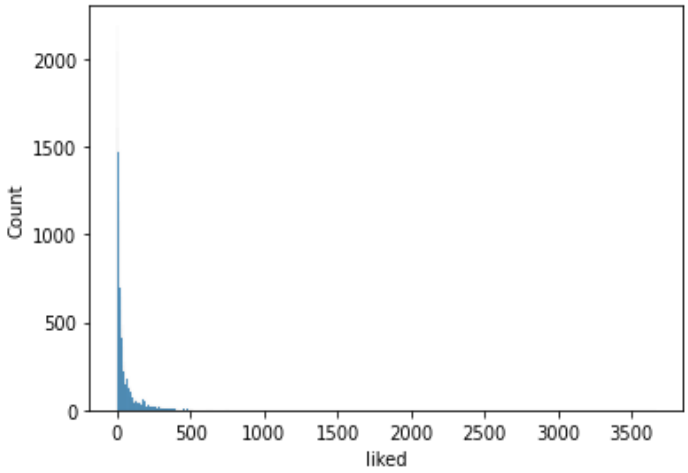
To je graf, ki prikazuje, da je ne glede na uro bilo med delavnikom več objav kot med vikendom. Takšen rezultat je logičen, saj imamo pet delovnih dni, med vikendom pa sta dneva le dva. Objave med vikendom so prikazane v sivem območju, objave med delavnikom pa čez sivo in modro območje.

Graf 11 – pogostost pojavljanja značke *#maribor* glede na uro in dan v tednu



Iz grafa lahko vidimo, ob kateri uri je na posamezen dan bilo največ objav. Tako lahko na primer vidimo, da je bilo ob 17.00 največ objav v soboto, ob 19.00 pa največ objav v četrtek.

Graf 12 – število všečkov



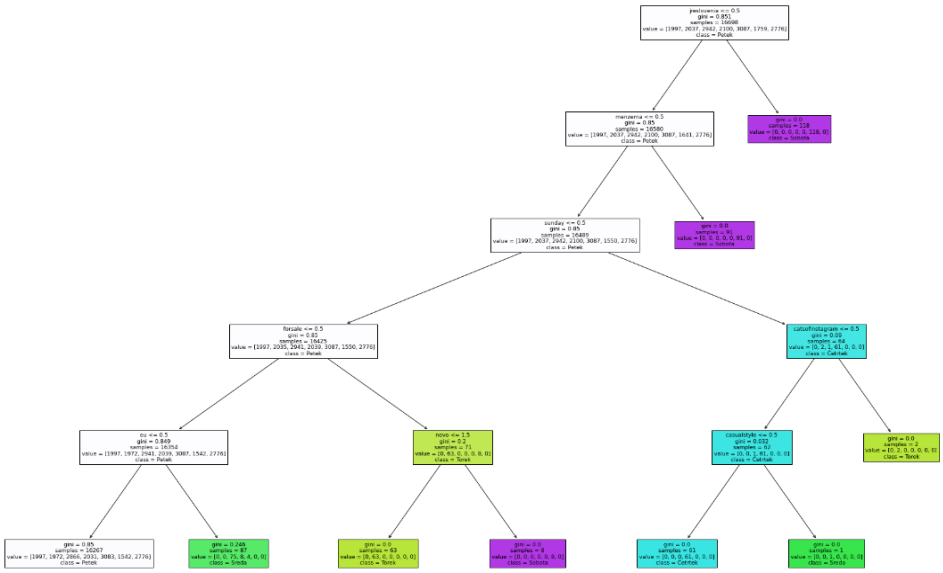
Iz grafa lahko razberemo, da ima večina objav med 0 in 10 všečkov. Objave z več kot 500 všečki so zelo redke.

Tabela 7 – všečki in komentarji objav

	povprečje	**minimum**	**maksimum**	**std**
Všečkane objave	60.397123473541384	0	3667	165.0166561626
Komentirane objave	2.299158751696065	0	318	9.439671521520

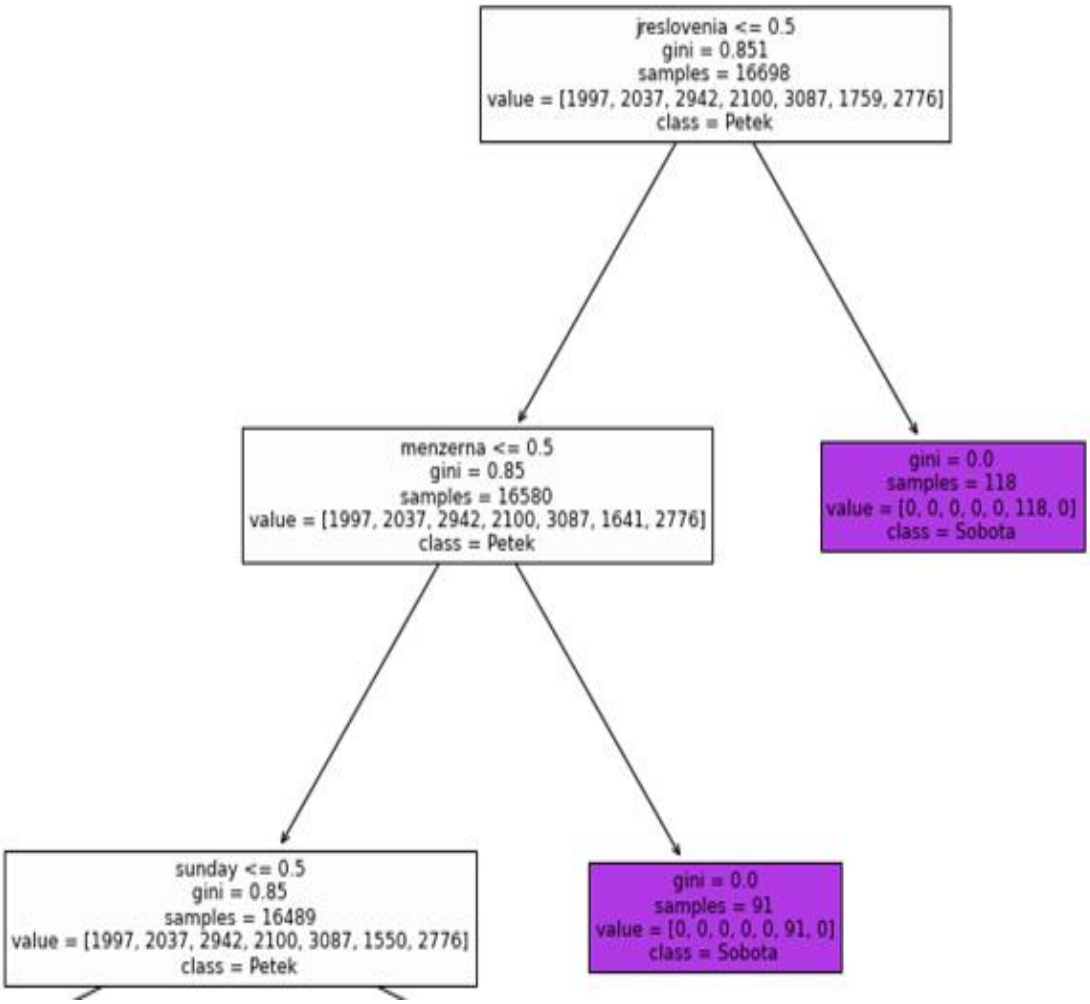
V povprečju je objave všečkalo 60,4 ljudi na posamezno objavo, medtem ko jih je v povprečju komentiralo le 2,3. Nismo presenečeni, da so obstajale objave brez všečkov oz. brez komentarjev. Všečkov je bilo maksimalno 3667, komentarjev pa 318. Standardni odklon (std) pri všečkih je 165, pri komentarjih pa 9,4.

Graf 13 – klasifikacijsko odločitveno drevo glede na dan v tednu (celotni graf)



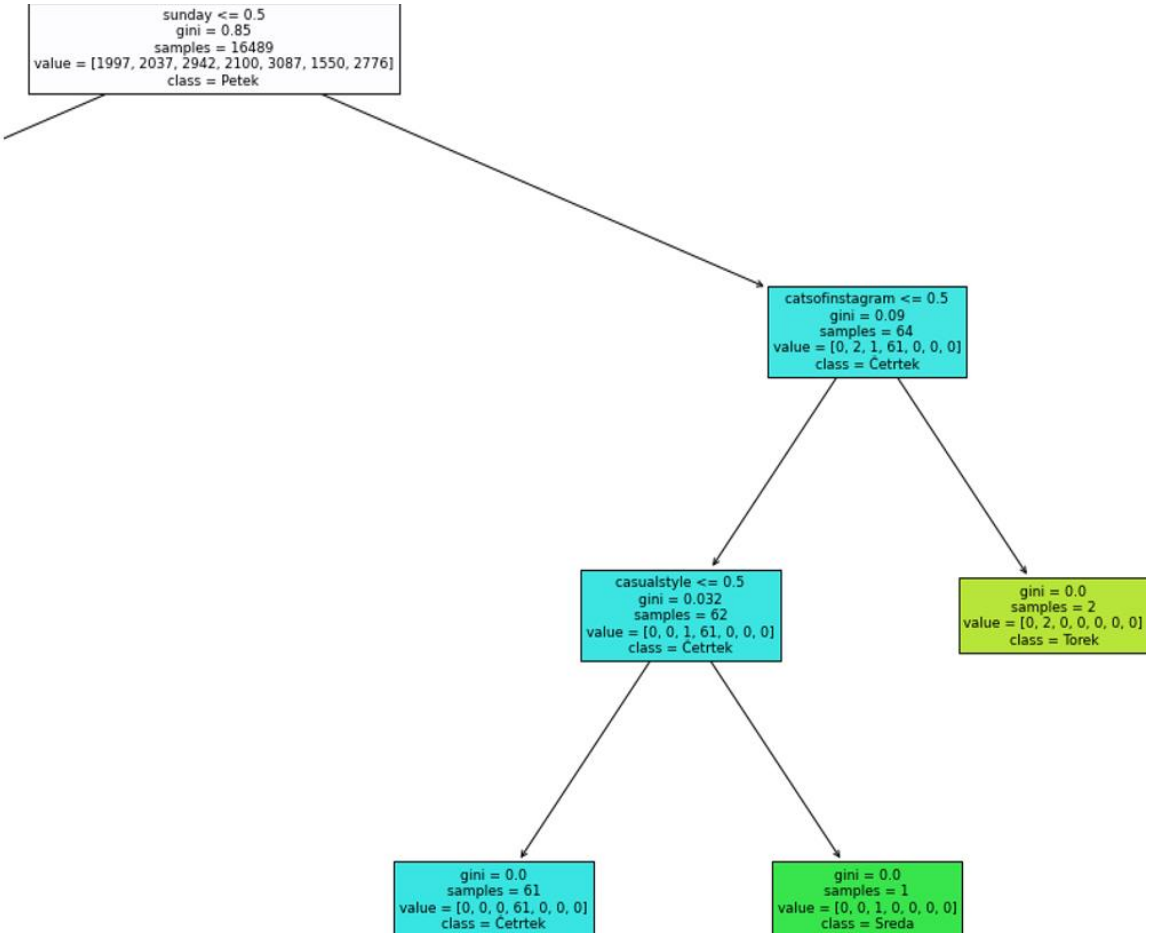
Na zgornjem grafu je prikazano prvo klasifikacijsko odločitveno drevo. To prikazuje, kateri dan v tednu bo najverjetneje objavljena določena značka skupaj z značko *#maribor*. Drevesa interpretiramo tako da začnemo na vrhu. Pogledamo prvi pravokotnik in pogledamo, če ima le ta dodano značko. Če ta obstaja, nadaljujemo v levo vejo, če značke ni pa nadaljujemo po desni veji. Ko pridemo do pravokotnika (oz. lista) iz katerega več veje ne gredo, pa pogledamo kateri dan je napisan v zadnji vrstici kvadratka in takrat je bila po vsej verjetnosti objavljena objava. Ker je slika grafa zelo velika, sem ga razdelila na manjše dele.

Graf 14 – klasifikacijsko odločitveno drevo glede na dan v tednu (prvi del)



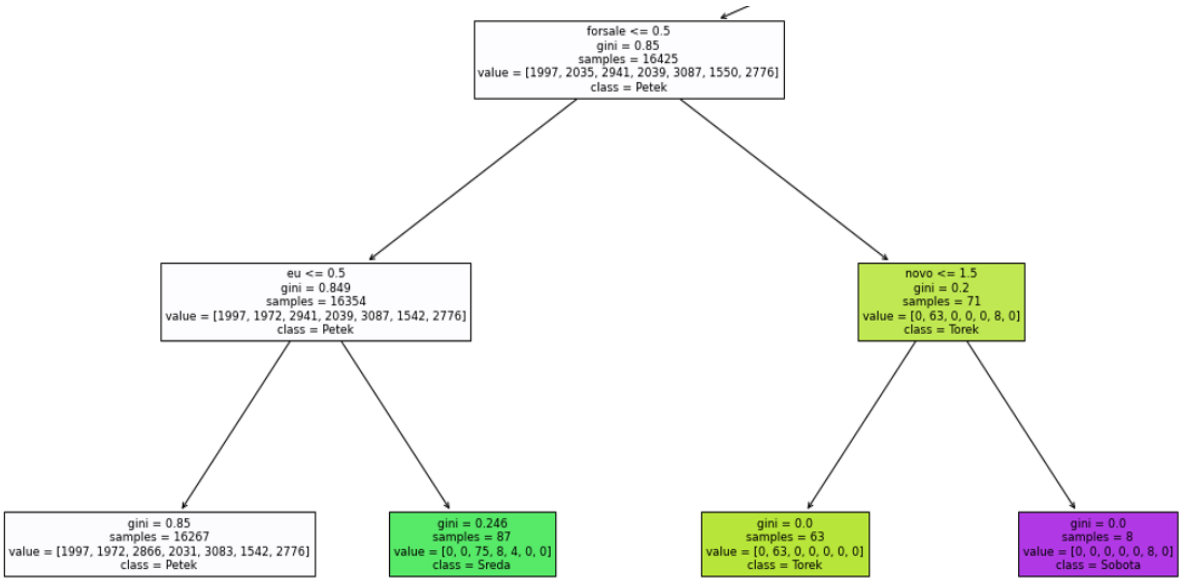
Če se značka *#jreslovenia* ne bo pojavila, se bo objava najverjetneje pojavila v soboto. Če se značka *#menzerna* ne bo pojavila, *#jreslovenia* pa se bo pojavila, se bo objava najverjetneje pojavila v soboto.

Graf 15 – klasifikacijsko odločitveno drevo glede na dan v tednu (drugi del)



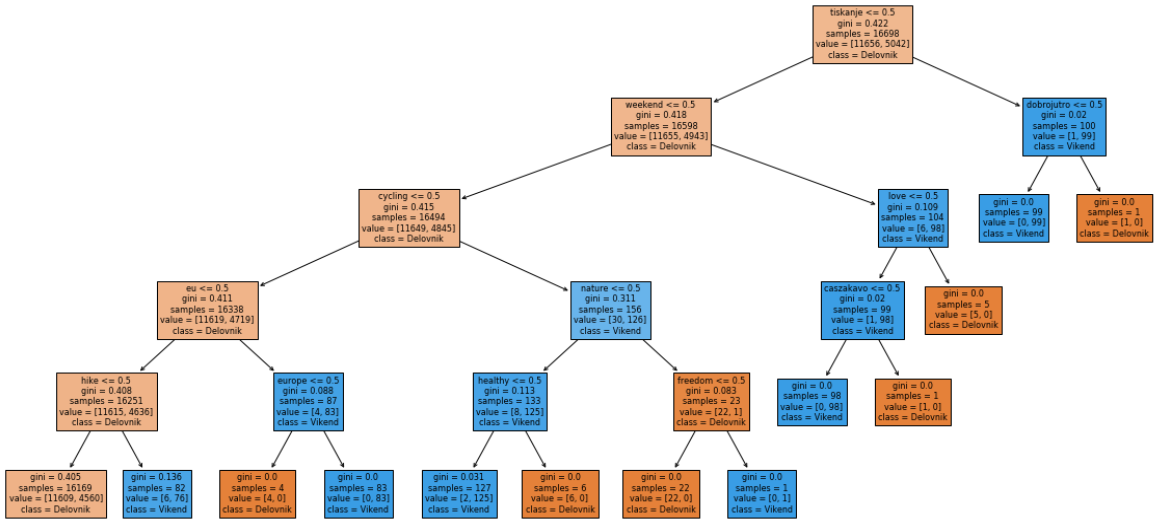
Če se pojavijo značka *#jreslovenia* in *#menzerna* brez značke *#sunday* in #catsofinstagram, se bo objava najverjetneje pojavila v torek. Če se pojavijo značke *#jreslovenia*, *#menzerna*, *#catsofinstagram* in *#casualstyle* brez značke *#Sunday*, se bo objava najverjetneje pojavila v četrtek. Če se pojavijo značke *#jreslovenia*, *#menzerna* in *#catsofinstagram* brez značk *#sunday* in *#casualstyle*, se bo objava najverjetneje pojavila v sredo.

Graf 16 – klasifikacijsko odločitveno drevo glede na dan v tednu (tretji del)



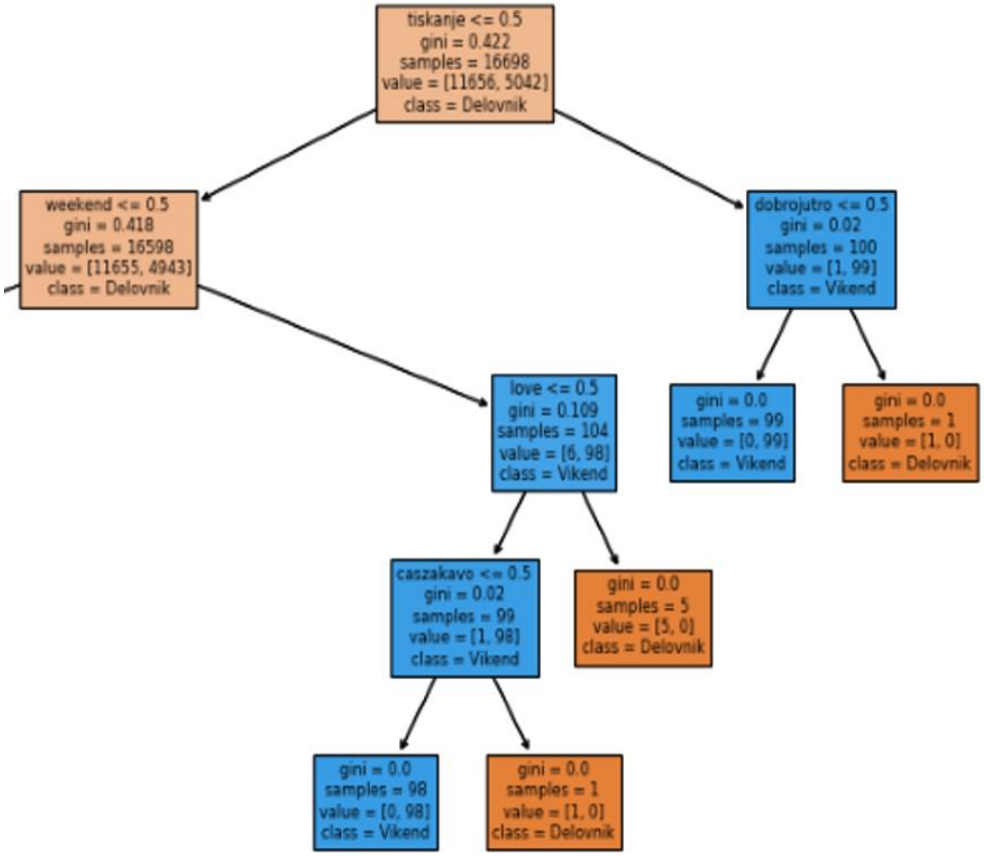
Če se pojavijo značke *#jreslovenia*, *#menzerna*, *#sunday* brez značk *#forsale* in *#novo*, se bo objava najverjetneje pojavila v soboto. Če se pojavijo značke *#jreslovenia*, *#menzerna*, *#sunday* in *#novo* brez značke *#forsale*, se bo objava najverjetneje pojavila v torek. Če pa se pojavijo značke *#jreslovenia*, *#menzerna*, *#sunday*, *#forsale* brez značke *#eu*, se bo objava najverjetneje pojavila v sredo. Če se pojavijo značke *#jreslovenia*, *#menzerna*, *#sunday*, *#forsale* in *#eu*, se bo objava najverjetneje pojavila v petek.

Graf 17 – klasifikacijsko odločitveno drevo glede na to, ali je delavnik ali vikend (celotni graf)



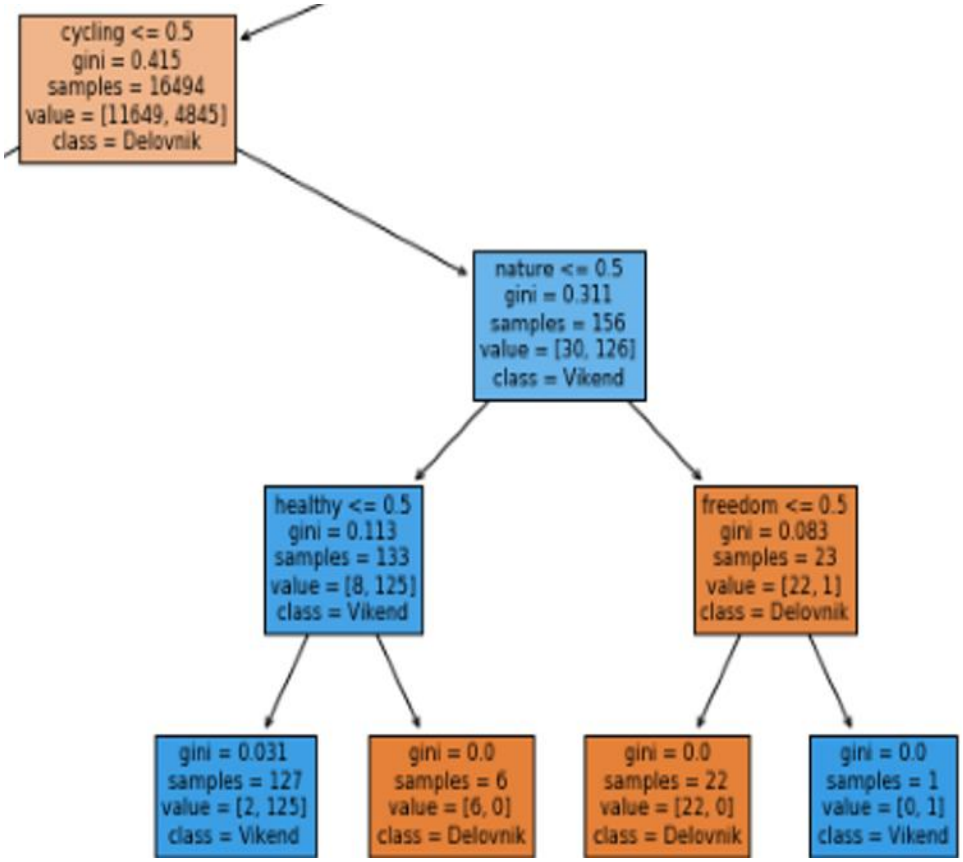
Na zgornjem grafu je prikazano drugo klasifikacijsko odločitveno drevo. Le to prikazuje, ali je bolj verjetno, da bo značka objavljena med vikendom ali med delavnikom. Ker je slika grafa zelo velika, sem ga razdelila na manjše dele.

Graf 18 – klasifikacijsko odločitveno drevo glede na to, ali je delavnik ali vikend (prvi del)



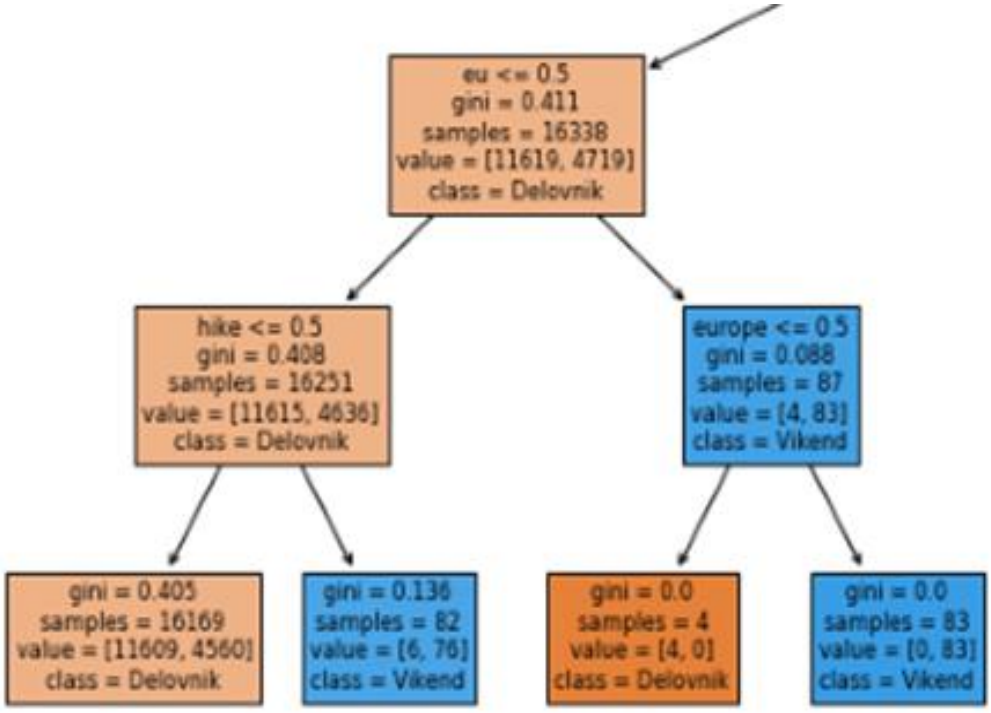
Če se značke *#tiskanje* in *#dobrojutro* ne pojavita, bo objava najverjetneje objavljena med delavnikom. Če se značka *#dobrojutro* pojavi brez *#tiskanje*, bo najverjetneje objavljena med vikendom. Če se značka *#tiskanje* pojavi brez značke *#weekend* in *#love,* se bo najverjetneje pojavila med delovnikom. Če se pojavijo značke *#tiskanje, #love* in *#caszakavo* brez *#weekend,* se bo objava najverjetneje pojavila med vikendom. Če se pojavita znački *#tiskanje* in *#love* brez značk *#weekend* in *#caszakavo,* se bo objava najverjetneje pojavila med delovnikom.

Graf 19 – klasifikacijsko odločitveno drevo glede na to, ali je delavnik ali vikend (drugi del)



Če se pojavita znački *#tiskanje* in *#weekend* brez značk *#cycling*, *#nature* in *#freedom*, se bo objava najverjetneje pojavila čez vikend. Če se pojavijo značke *#tiskanje*, *#weekend* in *#freedom* brez značk *#cycling* in *#nature*, se bo objava najverjetneje pojavila med delavnikom. Če se pojavijo značke *#tiskanje*, *#weekend*, *#nature* in *#healthy* brez *#weekend*, se bo objava najverjetneje pojavila čez vikend. Če se pojavijo značke *#tiskanje*, *#weekend*, *#nature* brez značk *#cycling* in *#healthy*, se bo objava najverjetneje pojavila med delavnikom.

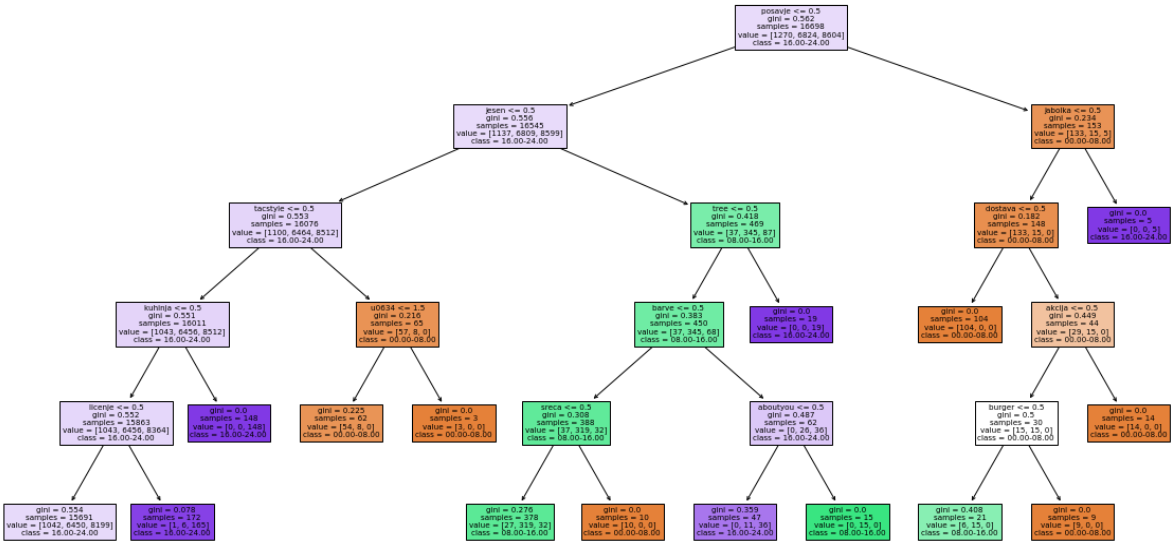
Graf 20 – klasifikacijsko odločitveno drevo glede na to, ali je delovnik ali vikend (tretji del)



Če se pojavijo značke *#tiskanje*, *#weekend* in *#cycling* brez značk *#eu* in *#europe,* se bo objava najverjetneje pojavila čez vikend. Če se pojavijo značke *#tiskanje*, *#weekend*, *#cycling* in *#europe* brez značke *#eu,* se bo objava najverjetneje pojavila med delavnikom.

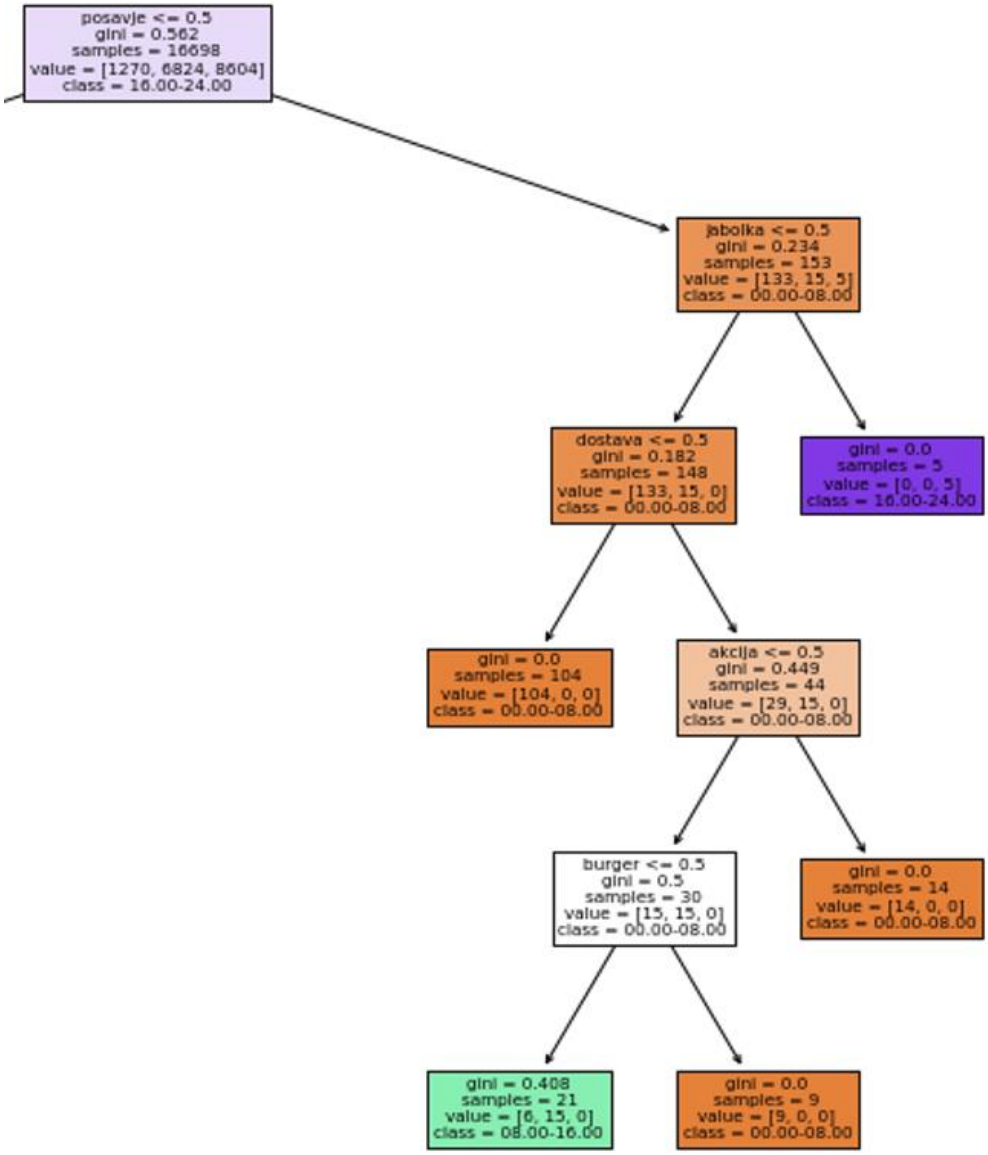
Če se pojavijo značke *#tiskanje*, *#weekend*, *#cycling* in *#eu* brez *#hike,* se bo objava najverjetneje pojavila čez vikend. Če se pojavijo značke *#tiskanje*, *#weekend*, *#cycling*, *#eu* in *#hike,* se bo objava najverjetneje pojavila med delavnikom.

Graf 21 – klasifikacijsko odločitveno drevo glede na del dneva (celotni graf)



Na zgornjem grafu imamo prikazano tretje klasifikacijsko odločitveno drevo. Le to prikazuje, v katerem času dneva je najverjetneje, da se bo značka pojavila skupaj z značko *#maribor*. Ali je to od 00.00 do 8.00, ali je to od 8.00 do 16.00 ali pa od 16.00 do 24.00. Tudi ta graf sem razdelila na manjše dele.

Graf 22 – klasifikacijsko odločitveno drevo glede na del dneva (prvi del)

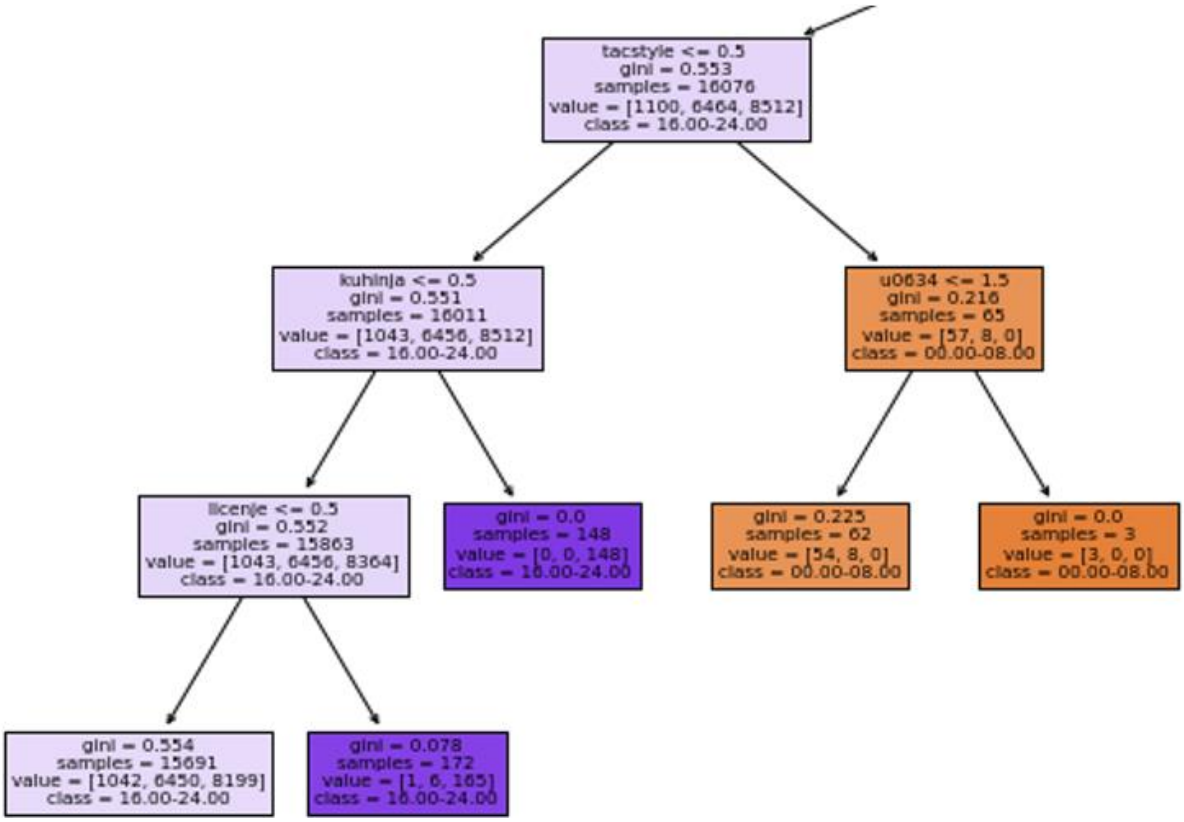


Če se pojavi objava brez značk *#posavje* in *#jabolka,* bo najverjetneje objavljena med 16.00 in 24.00. Če se pojavita znački *#jabolka* brez *#posavje*, *#dostava* in *#akcija*, se bo objava najverjetneje pojavila med 00.00 in 8.00. Če se pojavita znački *#jabolka* in *#dostava* brez značke *#posavje,* se bo objava najverjetneje pojavila med 00.00 in 8.00. Če se pojavijo značke *#jabolka*, *#akcija* in *#burger* brez značk *#posavje* in *#dostava,* se bo objava najverjetneje pojavila med 8.00 in 16.00. Če se pojavita znački *#jabolka* in *#akcija* brez značk *#posavje*, *#dostava* in *#burger,* se bo objava najverjetneje pojavila med 0.00 in 8.00.

Graf 23 – klasifikacijsko odločitveno drevo glede na del dneva (drugi del)

Če se pojavi značka #*posavje* brez značk #*jesen* in #*tree*, se bo objava najverjetneje pojavila med 16.00 in 24.00. Če se pojavita znački #*posavje* in #*tree* brez značk #*jesen*, #*barve* in #*aboutyou*, se bo objava najverjetneje pojavila med 8.00 in 16.00. Če se pojavijo značke #*posavje*, #*tree* in #*aboutyou* brez značk #*jesen* in #*barve*, se bo objava najverjetneje pojavila med 16.00 in 24.00. Če se pojavijo značke #*posavje*, #*tree*, #*barve* in #*srca* brez #*jesen*, se bo objava najverjetneje pojavila med 8.00 in 16.00. Če se pojavijo značke #*posavje*, #*tree* in #*barve* brez značk #*jesen* in #*srca*, se bo objava najverjetneje pojavila med 00.00 in 8.00.

Graf 24 – klasifikacijsko odločitveno drevo glede na del dneva (tretji del)



Če se pojavita znački *#posavje* in *#jesen* brez značk *#tacstyle* in *#u0634*, se bo objava najverjetneje pojavila med 00.00 in 8.00. Če se pojavijo značke *#posavje*, *#jesen* in *#u0634* brez *#tacstyle*, se bo objava najverjetneje pojavila med 00.00 in 8.00. Če se pojavijo značke *#posavje*, *#jesen*, *#tacstyle* brez značke *#kuhinja*, se bo objava najverjetneje pojavila med 16.00 in 24.00. Če se pojavijo značke *#posavje*, *#jesen*, *#tacstyle* in *#kuhinja* brez značke *#licenje*, se bo objava najverjetneje pojavila med 16.00 in 24.00. Če pa se pojavijo značke *#posavje*, *#jesen*, *#tacstyle*, *#kuhinja* in *#licenje*, se bo objava najverjetneje pojavila med 16.00 in 24.00.

7 RAZPRAVA, INTERPRETACIJA REZULTATOV

Interpretacija rezultatov:

1. Med dvajsetimi najpogostejšimi značkami, ki nimajo semantičnega pomena, se največkrat pojavi tema, povezana z ličenjem.

2. Največ objav z značko *#maribor* se pojavi med 17.00 in 19.00. Najmanj objav z značko *#maribor* se pojavijo med 2.00 in 4.00.

3. Največ objav z značko *#maribor* se pojavi v četrtek. Najmanj objav z značko *#maribor* se pojavi v torek.

4. Največ objav z značko *#maribor* je med 8.00 in 16.00. Najmanj objav z značko *#maribor* je med 00.00 in 8.00.

5. Med delavnikom je več objav kot med vikendom.

6. Povprečje všečkov pri objavah z značko *#maribor* je 60,4. Objava z značko *#maribor* z največ všečki ima 3667 všečkov.

7. Povprečje komentarjev pri objavah z značko *#maribor* je 2,3. Objava z značko *#maribor* z največ komentarji ima 318 komentarjev.

Hipoteze:

1. Največ objav med delavnikom je med 16.00 in 24.00.
   Hipotezo lahko potrdim, saj je bilo največ objav, ki vsebujejo *#maribor,* objavljenih med delavnikom med 16.00 in 24.00. Takšen rezultat sem predvidevala, saj takrat ljudje večinoma niso v službi oziroma šoli in imajo več časa.

2. Največ objav med vikendom je med 8.00 in 16.00.
   Hipotezo morem zavrniti, saj je bilo največ objav, ki vsebujejo *#maribor,* objavljenih med 16.00 in 24.00. Predvidevala sem, da bo največ objav med 8.00 in 16.00, saj so takrat ljudje velikokrat na izletih, imajo prosti čas in se družijo.

3. Posamezni dan v vikendu ima več objav kot posamezni dan v tednu.
   Hipotezo moram zavrniti, saj le ta ni nujno resnična. Ne drži, da je v soboto ali nedeljo več objav kot v četrtek. Prav tako ne drži, da je v nedeljo več objav kot v sredo.

4. Največ objav med delavnimi tedni je v petek.
   Hipotezo moram zavrniti, saj je bilo največ objav med delavnikom v četrtek, ne v petek. Rezultat me je presenetil, saj je ravno petek čas, ko se ljudje po navadi sprostijo in zabavajo. Morda bi se rezultat razlikoval, če ne bi bilo omejitev zaradi koronavirusa.

Zanimivosti:

1. Med 18426 objav, ki so vključene v raziskavo, najdemo kar 11998 različnih značk.

2. Največkrat uporabljena značka poleg #*maribor* je #*ljubljana*.

3. Največkrat uporabljena značka poleg #*maribor*, ki nima semantičnega pomena, je #*makeupartist*.

4. Največ objav z značko #*maribor* je v soboto ob 17.00.

8 ZAKLJUČEK/SKLEPI

V raziskovalni nalogi sem spoznala podatkovno rudarjenje. Raziskala sem, kako ga lahko uporabimo na konkretnem primeru. Pri tem sem se ogromno naučila z vidika programiranja in odkrila veliko zanimivosti o sami znački *#maribor*. Ugotovila sem, da je največ objav ob četrtkih, v časovnem pasu med 16.00 in 24.00, če smo natančnejši, med 17.00 in 18.00. Najpogostejša tema se navezuje na ličenje. Povprečje všečkov je 60,4, povprečje komentarjev pa 2,3. V raziskovalni nalogi sem prav tako spoznala uporabo oblaka besed in klasifikacijskega odločitvenega drevesa ter narisala veliko različnih grafov.

Raziskovalno nalogo bi zagotovo lahko nadaljevala. Lahko bi raziskala, kdaj se v Mariboru pogosto objavlja o različnih temah, kakšne slike se najpogosteje objavljajo med tednom in kakšne med vikendom in ali se vrste slik razlikujejo glede na lokacijo objave. Prav tako bi lahko raziskala lokacijo Maribor. Lahko bi uporabila še različne tehnike inteligentne obdelave podatkov, kot so štetje besed in besednih zvez, v kakšnih kombinacijah se značke pojavljajo in uporabo lokacij objav. Prav tako bi lahko s pomočjo tehnike umetne inteligence in prepoznavanja slik analizirala vsebino objav, iz katerih bi naredila analizo gručenja in časovno analizo.

9 DRUŽBENA ODGOVORNOST

Podatkovno rudarjenje je v sodobnem času zelo pomembno, saj smo res preplavljeni z ogromno količino podatkov. Omogoča nam, da iz velike količine dobimo le najpomembnejše. S tem prihranimo veliko časa, ki ga lahko uporabimo za druge koristne stvari. Z njim dobimo bolj razvidne ugotovitve, ki ji lahko hitro razumemo in učinkovito uporabimo.

Podatki, ki sem jih dobila z raziskovalno nalogo, lahko pomagajo mariborskim podjetjem, mariborski občini, mariborskim Instagram vplivnežem in turističnim agencijam. Predvsem glede tega, kdaj je najbolje objavljati, prav tako pa o tem, kaj Mariborčane najbolj zanima. Turistične agencije bi se lahko po epidemiji prilagodile uporabnikom in bi s spremljanjem objav prilagajale svoje delovanje. Občina lahko glede na podatke ugotovi, kdaj so občani v mestu najbolj aktivni. Glede na to lahko prilagaja vozne rede javnega prometa in cenovne načrte javnih parkirišč. Ugotovi lahko, v katerih delih mesta in ob katerem delu tedna so občani najbolj aktivni, in takrat še posebej poskrbi za red in privlačnost. Po drugi strani lahko ugotovi, v katerih dnevih je občane potrebno prepričati v dodatno uživanje v mestu, s tem pa spodbuja mestno življenje. Ugotovi lahko, o katerih delih mesta se najmanj objavlja, saj so to najverjetneje najmanj privlačni kraji, zato lahko glede teh krajev kaj spremeni in izboljša. Glede na najpogostejše značke v kombinaciji z značko *#maribor* lahko ugotovi zanimanje mladih občanov in s tem prilagodi ponudbo mesta. Določenim podjetjem lahko ponudi nižje najemnine poslovnih prostorov ali se aktivno zavzema za organizacijo povezanih dogodkov (v tem primeru bi se lahko to nanašalo na kozmetična podjetja, pa tudi na pohodniška in fotografska društva).

Značko *#maribor* sem uporabila predvsem zaradi njene atraktivnosti. Upam, da ta tehnologija vsak dan postaja vse dostopnejša, tako da si jo bodo kmalu lahko privoščili ne le naša velika mestna občina, ampak tudi manjši ponudniki, ki bodo na Instagramu preučili zanimanje strank za njihove specifične (predvsem nove) storitve. Pri tem jim utegne biti v pomoč, da sem svojo kodo pustila odprto in jo lahko samo prilagodijo za svoje potrebe.

10 VIRI IN LITERATURA

10.1 Knjižni viri

1. Merlak, M. (2015). Analiza trendov na področju družbenih medijev v javni upravi. (Diplomsko delo). Univerza v Ljubljani, Fakulteta za upravo.
2. Pavlič, K. (2019). Ovrednotenje uporabe družbenih omrežij na slovenskih ministrstvih in v vladi. (Diplomska delo). Univerza v Ljubljani, Fakulteta za upravo.
3. Saje, J. (2017). Oglaševanje na družbenih omrežjih in pomen uporabe družbenih omrežij v Občinski upravi Novo mesto. (Diplomska delo). FIŠ – Fakulteta za informacijske študije v Novem mestu.

10.2 Internetni viri

1. 7 Stages of Data Mining. (18.8.2020). Data Science & AI. https://medium.com/@datascienceandai101/7-stages-of-data-mining-process-262f48ec88ca
2. 16 Data Mining Techniques: The Complete List. (b. d.). Talend. https://www.talend.com/resources/data-mining-techniques/
3. About Pandas. (b. d.). Pandas. https://pandas.pydata.org/about/
4. An Introduction to Seaborn. (b. d.). Seaborn. https://seaborn.pydata.org/introduction.html
5. Dollarhide, Maya E. (b. d.). Social Media. Investopedia. https://www.investopedia.com/terms/s/social-media.asp
6. Epoch & Unix Timestamp Conversion Tools. (b. d.). Unixtimestamp. https://www.unixtimestamp.com/
7. History of Python. (b. d.). GeeksforGeeks. https://www.geeksforgeeks.org/history-of-python/
8. Instagram. (12. 1. 2021). Wikipedia. https://en.wikipedia.org/w/index.php?title=Instagram&oldid=999986167
9. Instagram. (b. d.). Safe.Si. https://safe.si/nasveti/druzabna-omrezja/instagram
10. Number of Social Media Users 2025. (b. d.). Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
11. Pandas (Software. (2020). Wikipedia. https://en.wikipedia.org/wiki/Pandas_(software

12. Project Jupyter. (b. d.). Jupyter.org. https://jupyter.org/
13. Scikit Learn – Introduction. (b. d.). Tutorialspoint. https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm
14. Social media. (2021). Wikipedia. https://en.wikipedia.org/w/index.php?title=Social_media&oldid=1000507616
15. What Is Python? Executive Summary. (b. d.). Python.Org. https://www.python.org/doc/essays/blurb/
16. What is social media? (b. d.). The Balance Small Business. https://www.thebalancesmb.com/what-is-social-media-2890301#:~:text=Social%20media%20is%20any%20digital,links%20and%20short%20written%20messages