Sveučilište u Rijeci – Odjel za informatiku

Diplomski studij informatike – modul informacijski i komunikacijski sustavi

Prikupljanje i analiza podataka s portala eZadar

Projektni zadatak

**Student:** Josip Lukin

**Mentori:** izv. prof.dr. sc. Ana Meštrović

dr. sc. Slobodan Beliga

**Kolegij**: Upravljanje znanjem

**Akademska godina**: 2021./2022.

Rijeka, 17.1.2022.

Sadržaj

[1. Uvod 3](#_Toc93353492)

[2. Web scraping 4](#_Toc93353493)

[3. Analiza prikupljenih podataka 6](#_Toc93353494)

[4. Vizualizacija podataka 17](#_Toc93353495)

[5. Analiza jezičnog diskursa 20](#_Toc93353496)

[6. Popis slika 34](#_Toc93353497)

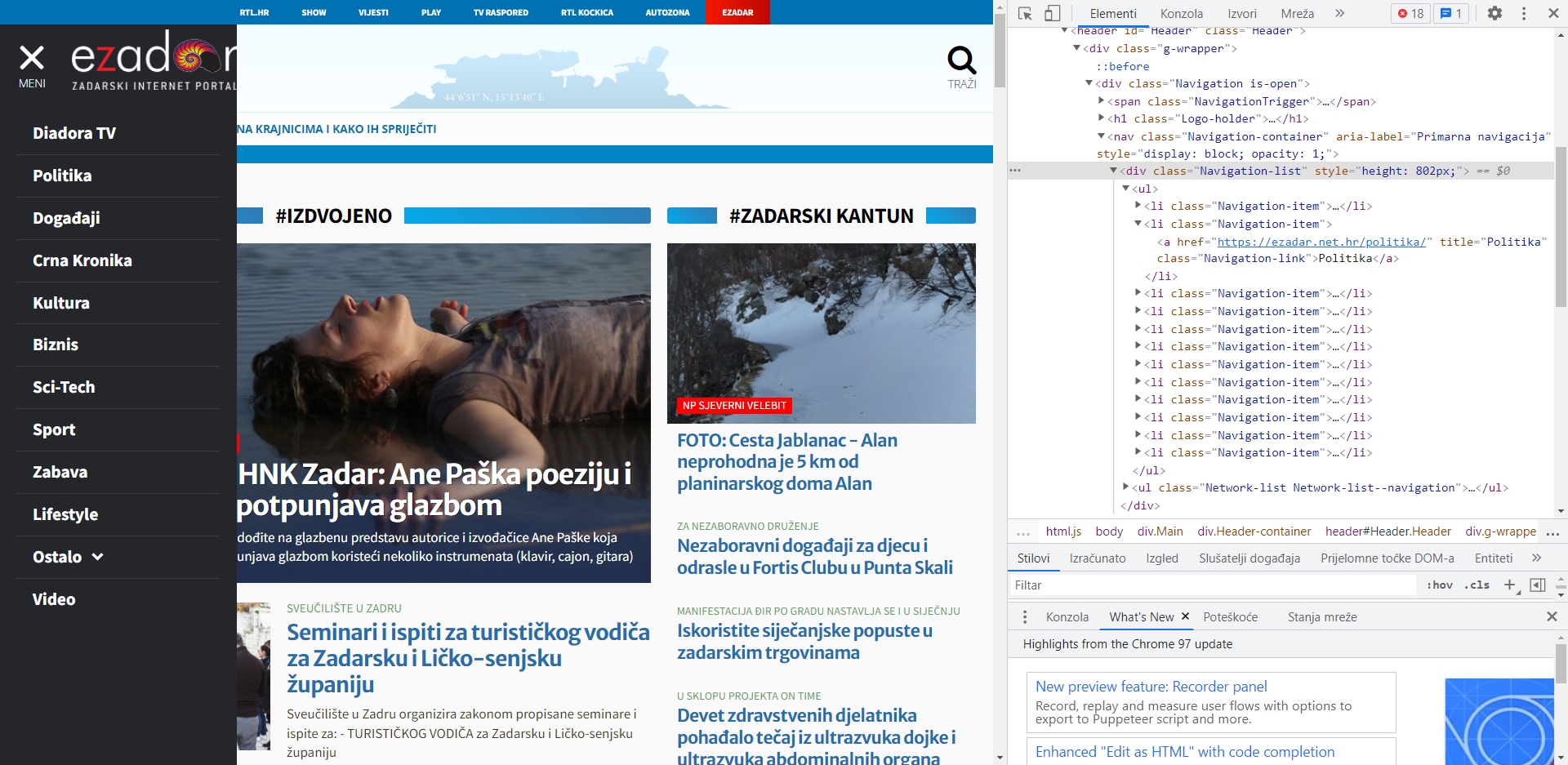
[7. Popis tablica 35](#_Toc93353498)

# Uvod

Cilj projekta je analiziranje podataka koji su prikupljeni s weba automatskim postupcima te generiranje novog znanja koje je temeljeno na otkrivenim činjenicama. Projekt je podijeljen u dva dijela. Zadatak prvog dijela projekta bio je prikupiti podatke portala eZadar.net.hr odnosno prikupiti poveznice svih članak objavljenih u 2021. godini. Nadalje je potrebno prikupiti podatke iz članka kao što su naslov, autor, datum objavljivanja, tekst članka i sl. U drugom dijelu projekta odrađena je analiza i vizualizacija dobivenih podataka te je kvantificiran broj vijesti (članaka) u medijskom prostoru portala eZadar.net.hr, koji se odnose na tematiku SARS-CoV-2 virusa.

# Web scraping

Prije pisanja programske skripte ***eZadar\_web\_scraping.py*** kojom će se odraditi web scraping, potrebno je istražiti HTML strukturu odabranog portala. Navedeno je moguće odraditi alatom za razvojne programere integriranom u većini preglednika.



Slika - Istraživanje HTML strukture

Uvidom u HTML strukturu naslovne stranice portala eZadar.net.hr primijećeno je da svaka kategorija u izborniku sadrži klasu *„Navigation-link“*. Korištenjem Python biblioteke za izvlačenje podataka iz HTML i XML datoteka, Beautiful Soup, napisana je fuknkcija *dohvati\_kategorije()* koja će upisati poveznice svih kategorija u tekstualnu datoteku *linkoviKategorije.txt*.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Slika - Funkcija za dohvaćanje kategorija

Daljnjim istraživanjem portala uočeno je kako za svaku od kategorija postoji arhiva. Na pojedinačnoj stranici arhive ukratko je prikazano 18 članaka po stranici. Kreirana je funkcija *dohvati\_linkove\_clanaka()* koja za svaku od kategorija prelistava stranice arhive „ulazeći“ u svaki članak te upisuje poveznicu članka u tekstualnu datoteku *linkoviClanaka.txt* ukoliko je datum objavljivanja članka u odgovarajućem period odnosno u 2021. godini. Ukoliko je datum objavljivanja članka nakon navedenog perioda funkcija zanemaruje takav članak i nastavlja s radom. Ukoliko datum objavljivanja članka prije navedenog perioda funkcija mijenja kategoriju te ponovno odrađuje prethodno opisan način rada.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Slika - Datumska kontrola u funkciji

Naposljetku kreirana je funkcija *eZadar\_scrap()* koja obavlja prikupljanje podataka poput poveznice, autora, potpunog teksta članka, kategorije, datuma objavljivanja, broj facebook like-ova i share-ova. Funkcija radi pomoću alata Selenium koji je inicijalno napravljen za testiranje web aplikacija, ali nije ograničen samo za takvu upotrebu. Funkcija pokreće Mozila Firefox preglednik te otvara svaku poveznicu iz datoteke *linkoviClanaka.txt* prikupljajući navedene podatke. Podatke dohvaća na osnovu CSS klase, XPath putanje ili HTML oznake i upisuje u CSV datoteku *eZadarPodaci.csv.* Iz CSV datoteke kreirana je odgovarajuća JSON datoteka.

# Analiza prikupljenih podataka

U programskoj skripti ***eZadar\_analiza.py*** odrađena je analiza prethodno dobivenih podataka. Prije analize prikupljenih podataka potrebno je transformirati određene podatke nakon učitavanja u DataFrame. Funkcijom *dataTransformation()* uklonjen je stupac ID, stupci TAG, NASLOV i POTPUNI TEKST konvertirani su u mala slova dok je iz stupca DATUM OBJAVLJIVANJA uklonjeno vrijeme objave te je podijeljen u tri nova stupca: DAN, MJESEC i GODINA.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Slika - Transformacija prikupljenih podataka

Funkcijom *covidNews()* pretraženi su stupci NASLOV, POTPUNI TEKST i TAG prethodno kreiranog DataFrame-a po odabranim ključnim riječima. Pronađeni indeksi članaka vezani uz korona tematiku pohranjeni su u listu *covid\_news\_indexes* jedino ako indeks već ne postoji u listi. Navedeno je implementirano kako bi izbjegli dupliciranje pojedinih indeksa i osigurali točnost dobivenih rezultata. Na kraju, funkcija će prikazati pronađen broj članaka. Funkcije *vaccineNews(), antimaskerNews()* i *antivakserNews()* rade po istom principu te se jedino razlikuju u listama ključnih riječi.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Slika - Dobiveni ispis funkcija

Funkcijom *newsDaily()* kreirana je excel datoteka *Objave\_po\_danima.xlsx* koja sadržava broj dnevnih objava na portalu i broj dnevnih objava vezanih uz korona tematiku. U funkciji je prvotno kreiran DataFrame sa stupcem Datum koji sadrži sve datume u 2021. godini. Za svaki datum u DataFrame-u pobrojani su članci s istim datumom objavljivanja, a zatim su pobrojani i članci vezani uz korona tematiku. Niže se nalazi ispis dobivenih podataka.

Tablica - Objave po datumima

|  |  |  |
| --- | --- | --- |
| **Datum** | **Broj svih objava na portalu** | **Broj objava vezanih za koronu** |
| 01.01.2021 | 47 | 15 |
| 02.01.2021 | 58 | 16 |
| 03.01.2021 | 61 | 19 |
| 04.01.2021 | 74 | 27 |
| 05.01.2021 | 75 | 24 |
| 06.01.2021 | 60 | 16 |
| 07.01.2021 | 80 | 20 |
| 08.01.2021 | 51 | 23 |
| 09.01.2021 | 70 | 20 |
| 10.01.2021 | 63 | 17 |
| 11.01.2021 | 67 | 21 |
| 12.01.2021 | 70 | 22 |
| 13.01.2021 | 65 | 17 |
| 14.01.2021 | 78 | 27 |
| 15.01.2021 | 70 | 20 |
| 16.01.2021 | 61 | 14 |
| 17.01.2021 | 50 | 13 |
| 18.01.2021 | 65 | 17 |
| 19.01.2021 | 74 | 23 |
| 20.01.2021 | 78 | 21 |
| 21.01.2021 | 64 | 17 |
| 22.01.2021 | 58 | 16 |
| 23.01.2021 | 57 | 17 |
| 24.01.2021 | 57 | 17 |
| 25.01.2021 | 70 | 18 |
| 26.01.2021 | 66 | 15 |
| 27.01.2021 | 61 | 22 |
| 28.01.2021 | 68 | 18 |
| 29.01.2021 | 52 | 20 |
| 30.01.2021 | 56 | 10 |
| 31.01.2021 | 59 | 16 |
| 01.02.2021 | 67 | 17 |
| 02.02.2021 | 70 | 21 |
| 03.02.2021 | 58 | 19 |
| 04.02.2021 | 60 | 17 |
| 05.02.2021 | 57 | 16 |
| 06.02.2021 | 50 | 15 |
| 07.02.2021 | 55 | 13 |
| 08.02.2021 | 68 | 19 |
| 09.02.2021 | 72 | 21 |
| 10.02.2021 | 72 | 24 |
| 11.02.2021 | 57 | 19 |
| 12.02.2021 | 62 | 17 |
| 13.02.2021 | 56 | 12 |
| 14.02.2021 | 54 | 13 |
| 15.02.2021 | 68 | 17 |
| 16.02.2021 | 58 | 16 |
| 17.02.2021 | 51 | 14 |
| 18.02.2021 | 63 | 14 |
| 19.02.2021 | 55 | 21 |
| 20.02.2021 | 50 | 10 |
| 21.02.2021 | 59 | 19 |
| 22.02.2021 | 70 | 17 |
| 23.02.2021 | 74 | 12 |
| 24.02.2021 | 54 | 17 |
| 25.02.2021 | 72 | 24 |
| 26.02.2021 | 63 | 21 |
| 27.02.2021 | 58 | 15 |
| 28.02.2021 | 70 | 8 |
| 01.03.2021 | 69 | 18 |
| 02.03.2021 | 70 | 23 |
| 03.03.2021 | 66 | 19 |
| 04.03.2021 | 84 | 22 |
| 05.03.2021 | 58 | 16 |
| 06.03.2021 | 51 | 9 |
| 07.03.2021 | 63 | 12 |
| 08.03.2021 | 74 | 19 |
| 09.03.2021 | 64 | 17 |
| 10.03.2021 | 58 | 19 |
| 11.03.2021 | 72 | 21 |
| 12.03.2021 | 56 | 16 |
| 13.03.2021 | 61 | 23 |
| 14.03.2021 | 51 | 9 |
| 15.03.2021 | 76 | 24 |
| 16.03.2021 | 81 | 22 |
| 17.03.2021 | 58 | 20 |
| 18.03.2021 | 71 | 22 |
| 19.03.2021 | 67 | 18 |
| 20.03.2021 | 59 | 13 |
| 21.03.2021 | 56 | 15 |
| 22.03.2021 | 75 | 27 |
| 23.03.2021 | 81 | 25 |
| 24.03.2021 | 63 | 22 |
| 25.03.2021 | 66 | 25 |
| 26.03.2021 | 54 | 19 |
| 27.03.2021 | 54 | 8 |
| 28.03.2021 | 44 | 14 |
| 29.03.2021 | 74 | 20 |
| 30.03.2021 | 63 | 19 |
| 31.03.2021 | 67 | 15 |
| 01.04.2021 | 69 | 21 |
| 02.04.2021 | 59 | 15 |
| 03.04.2021 | 57 | 15 |
| 04.04.2021 | 41 | 12 |
| 05.04.2021 | 50 | 13 |
| 06.04.2021 | 71 | 21 |
| 07.04.2021 | 86 | 29 |
| 08.04.2021 | 74 | 27 |
| 09.04.2021 | 48 | 16 |
| 10.04.2021 | 63 | 15 |
| 11.04.2021 | 58 | 15 |
| 12.04.2021 | 67 | 21 |
| 13.04.2021 | 68 | 19 |
| 14.04.2021 | 78 | 21 |
| 15.04.2021 | 76 | 24 |
| 16.04.2021 | 64 | 21 |
| 17.04.2021 | 41 | 9 |
| 18.04.2021 | 60 | 12 |
| 19.04.2021 | 77 | 21 |
| 20.04.2021 | 63 | 15 |
| 21.04.2021 | 76 | 16 |
| 22.04.2021 | 74 | 22 |
| 23.04.2021 | 60 | 17 |
| 24.04.2021 | 64 | 16 |
| 25.04.2021 | 65 | 24 |
| 26.04.2021 | 74 | 20 |
| 27.04.2021 | 71 | 14 |
| 28.04.2021 | 88 | 17 |
| 29.04.2021 | 88 | 26 |
| 30.04.2021 | 63 | 21 |
| 01.05.2021 | 56 | 15 |
| 02.05.2021 | 47 | 11 |
| 03.05.2021 | 76 | 18 |
| 04.05.2021 | 58 | 12 |
| 05.05.2021 | 65 | 21 |
| 06.05.2021 | 61 | 15 |
| 07.05.2021 | 69 | 22 |
| 08.05.2021 | 62 | 17 |
| 09.05.2021 | 68 | 12 |
| 10.05.2021 | 81 | 24 |
| 11.05.2021 | 64 | 14 |
| 12.05.2021 | 79 | 18 |
| 13.05.2021 | 76 | 21 |
| 14.05.2021 | 83 | 24 |
| 15.05.2021 | 39 | 10 |
| 16.05.2021 | 62 | 13 |
| 17.05.2021 | 73 | 13 |
| 18.05.2021 | 76 | 13 |
| 19.05.2021 | 68 | 11 |
| 20.05.2021 | 71 | 17 |
| 21.05.2021 | 66 | 14 |
| 22.05.2021 | 61 | 15 |
| 23.05.2021 | 52 | 10 |
| 24.05.2021 | 78 | 21 |
| 25.05.2021 | 73 | 12 |
| 26.05.2021 | 76 | 15 |
| 27.05.2021 | 76 | 12 |
| 28.05.2021 | 84 | 22 |
| 29.05.2021 | 59 | 10 |
| 30.05.2021 | 58 | 7 |
| 31.05.2021 | 78 | 15 |
| 01.06.2021 | 82 | 17 |
| 02.06.2021 | 75 | 21 |
| 03.06.2021 | 69 | 15 |
| 04.06.2021 | 49 | 12 |
| 05.06.2021 | 78 | 19 |
| 06.06.2021 | 61 | 10 |
| 07.06.2021 | 90 | 20 |
| 08.06.2021 | 80 | 12 |
| 09.06.2021 | 84 | 17 |
| 10.06.2021 | 78 | 16 |
| 11.06.2021 | 63 | 12 |
| 12.06.2021 | 59 | 8 |
| 13.06.2021 | 60 | 11 |
| 14.06.2021 | 81 | 17 |
| 15.06.2021 | 59 | 10 |
| 16.06.2021 | 85 | 18 |
| 17.06.2021 | 79 | 8 |
| 18.06.2021 | 78 | 18 |
| 19.06.2021 | 71 | 11 |
| 20.06.2021 | 57 | 12 |
| 21.06.2021 | 70 | 12 |
| 22.06.2021 | 48 | 13 |
| 23.06.2021 | 68 | 14 |
| 24.06.2021 | 82 | 14 |
| 25.06.2021 | 82 | 19 |
| 26.06.2021 | 52 | 14 |
| 27.06.2021 | 53 | 16 |
| 28.06.2021 | 78 | 19 |
| 29.06.2021 | 75 | 17 |
| 30.06.2021 | 85 | 18 |
| 01.07.2021 | 75 | 16 |
| 02.07.2021 | 83 | 17 |
| 03.07.2021 | 82 | 16 |
| 04.07.2021 | 55 | 11 |
| 05.07.2021 | 72 | 9 |
| 06.07.2021 | 85 | 23 |
| 07.07.2021 | 101 | 21 |
| 08.07.2021 | 88 | 27 |
| 09.07.2021 | 53 | 16 |
| 10.07.2021 | 82 | 25 |
| 11.07.2021 | 69 | 11 |
| 12.07.2021 | 82 | 19 |
| 13.07.2021 | 83 | 19 |
| 14.07.2021 | 79 | 15 |
| 15.07.2021 | 89 | 21 |
| 16.07.2021 | 53 | 15 |
| 17.07.2021 | 76 | 14 |
| 18.07.2021 | 59 | 8 |
| 19.07.2021 | 74 | 13 |
| 20.07.2021 | 81 | 16 |
| 21.07.2021 | 78 | 13 |
| 22.07.2021 | 89 | 22 |
| 23.07.2021 | 55 | 8 |
| 24.07.2021 | 68 | 14 |
| 25.07.2021 | 79 | 13 |
| 26.07.2021 | 83 | 12 |
| 27.07.2021 | 86 | 15 |
| 28.07.2021 | 80 | 12 |
| 29.07.2021 | 84 | 14 |
| 30.07.2021 | 54 | 6 |
| 31.07.2021 | 84 | 14 |
| 01.08.2021 | 71 | 12 |
| 02.08.2021 | 70 | 10 |
| 03.08.2021 | 78 | 10 |
| 04.08.2021 | 83 | 17 |
| 05.08.2021 | 78 | 15 |
| 06.08.2021 | 61 | 9 |
| 07.08.2021 | 64 | 10 |
| 08.08.2021 | 49 | 14 |
| 09.08.2021 | 69 | 6 |
| 10.08.2021 | 69 | 14 |
| 11.08.2021 | 56 | 8 |
| 12.08.2021 | 68 | 12 |
| 13.08.2021 | 49 | 6 |
| 14.08.2021 | 74 | 13 |
| 15.08.2021 | 77 | 12 |
| 16.08.2021 | 69 | 7 |
| 17.08.2021 | 74 | 13 |
| 18.08.2021 | 83 | 13 |
| 19.08.2021 | 82 | 22 |
| 20.08.2021 | 66 | 15 |
| 21.08.2021 | 56 | 7 |
| 22.08.2021 | 58 | 6 |
| 23.08.2021 | 41 | 10 |
| 24.08.2021 | 79 | 19 |
| 25.08.2021 | 60 | 11 |
| 26.08.2021 | 75 | 17 |
| 27.08.2021 | 64 | 17 |
| 28.08.2021 | 65 | 9 |
| 29.08.2021 | 69 | 10 |
| 30.08.2021 | 78 | 15 |
| 31.08.2021 | 75 | 8 |
| 01.09.2021 | 75 | 15 |
| 02.09.2021 | 75 | 16 |
| 03.09.2021 | 75 | 13 |
| 04.09.2021 | 59 | 7 |
| 05.09.2021 | 55 | 8 |
| 06.09.2021 | 70 | 12 |
| 07.09.2021 | 77 | 15 |
| 08.09.2021 | 61 | 10 |
| 09.09.2021 | 68 | 13 |
| 10.09.2021 | 71 | 12 |
| 11.09.2021 | 55 | 10 |
| 12.09.2021 | 77 | 9 |
| 13.09.2021 | 65 | 14 |
| 14.09.2021 | 82 | 11 |
| 15.09.2021 | 70 | 15 |
| 16.09.2021 | 67 | 17 |
| 17.09.2021 | 77 | 16 |
| 18.09.2021 | 53 | 8 |
| 19.09.2021 | 62 | 8 |
| 20.09.2021 | 69 | 13 |
| 21.09.2021 | 81 | 12 |
| 22.09.2021 | 70 | 13 |
| 23.09.2021 | 95 | 26 |
| 24.09.2021 | 58 | 9 |
| 25.09.2021 | 70 | 7 |
| 26.09.2021 | 68 | 8 |
| 27.09.2021 | 79 | 8 |
| 28.09.2021 | 91 | 20 |
| 29.09.2021 | 92 | 17 |
| 30.09.2021 | 85 | 14 |
| 01.10.2021 | 75 | 9 |
| 02.10.2021 | 59 | 5 |
| 03.10.2021 | 71 | 9 |
| 04.10.2021 | 80 | 19 |
| 05.10.2021 | 75 | 8 |
| 06.10.2021 | 78 | 13 |
| 07.10.2021 | 77 | 10 |
| 08.10.2021 | 53 | 11 |
| 09.10.2021 | 75 | 11 |
| 10.10.2021 | 68 | 12 |
| 11.10.2021 | 77 | 15 |
| 12.10.2021 | 85 | 16 |
| 13.10.2021 | 82 | 14 |
| 14.10.2021 | 78 | 13 |
| 15.10.2021 | 62 | 12 |
| 16.10.2021 | 60 | 7 |
| 17.10.2021 | 54 | 8 |
| 18.10.2021 | 85 | 11 |
| 19.10.2021 | 75 | 15 |
| 20.10.2021 | 81 | 16 |
| 21.10.2021 | 80 | 15 |
| 22.10.2021 | 78 | 8 |
| 23.10.2021 | 71 | 7 |
| 24.10.2021 | 73 | 13 |
| 25.10.2021 | 91 | 20 |
| 26.10.2021 | 84 | 15 |
| 27.10.2021 | 77 | 14 |
| 28.10.2021 | 91 | 22 |
| 29.10.2021 | 102 | 19 |
| 30.10.2021 | 53 | 12 |
| 31.10.2021 | 49 | 5 |
| 01.11.2021 | 63 | 8 |
| 02.11.2021 | 87 | 15 |
| 03.11.2021 | 73 | 16 |
| 04.11.2021 | 91 | 22 |
| 05.11.2021 | 78 | 24 |
| 06.11.2021 | 73 | 15 |
| 07.11.2021 | 76 | 15 |
| 08.11.2021 | 77 | 21 |
| 09.11.2021 | 83 | 20 |
| 10.11.2021 | 84 | 21 |
| 11.11.2021 | 114 | 21 |
| 12.11.2021 | 98 | 28 |
| 13.11.2021 | 57 | 15 |
| 14.11.2021 | 67 | 15 |
| 15.11.2021 | 71 | 25 |
| 16.11.2021 | 83 | 24 |
| 17.11.2021 | 90 | 27 |
| 18.11.2021 | 77 | 20 |
| 19.11.2021 | 80 | 17 |
| 20.11.2021 | 81 | 22 |
| 21.11.2021 | 65 | 19 |
| 22.11.2021 | 106 | 22 |
| 23.11.2021 | 98 | 28 |
| 24.11.2021 | 92 | 21 |
| 25.11.2021 | 109 | 19 |
| 26.11.2021 | 76 | 24 |
| 27.11.2021 | 65 | 21 |
| 28.11.2021 | 61 | 16 |
| 29.11.2021 | 72 | 24 |
| 30.11.2021 | 96 | 19 |
| 01.12.2021 | 78 | 18 |
| 02.12.2021 | 107 | 32 |
| 03.12.2021 | 78 | 18 |
| 04.12.2021 | 77 | 23 |
| 05.12.2021 | 66 | 9 |
| 06.12.2021 | 85 | 14 |
| 07.12.2021 | 94 | 22 |
| 08.12.2021 | 99 | 23 |
| 09.12.2021 | 86 | 15 |
| 10.12.2021 | 95 | 31 |
| 11.12.2021 | 62 | 16 |
| 12.12.2021 | 59 | 10 |
| 13.12.2021 | 74 | 21 |
| 14.12.2021 | 96 | 23 |
| 15.12.2021 | 90 | 22 |
| 16.12.2021 | 90 | 24 |
| 17.12.2021 | 67 | 17 |
| 18.12.2021 | 68 | 17 |
| 19.12.2021 | 64 | 19 |
| 20.12.2021 | 87 | 24 |
| 21.12.2021 | 78 | 20 |
| 22.12.2021 | 77 | 16 |
| 23.12.2021 | 74 | 22 |
| 24.12.2021 | 80 | 19 |
| 25.12.2021 | 61 | 15 |
| 26.12.2021 | 73 | 12 |
| 27.12.2021 | 65 | 14 |
| 28.12.2021 | 79 | 23 |
| 29.12.2021 | 70 | 17 |
| 30.12.2021 | 52 | 19 |
| 31.12.2021 | 51 | 13 |

Na sličan način radi i funkcija *newMonthly()*. Kreiran je DataFrame koji sadrži stupac MJESEC u kojem se nalaze mjeseci odnosno brojevi od 01-12. Za svaki mjesec u DataFrame-u pobrojani su članci s istim mjesecom objavljivanja, a zatim su pobrojani i članci vezani uz korona tematiku. Niže se nalazi ispis dobivenih podataka.

Tablica - Objave po mjesecima

|  |  |  |
| --- | --- | --- |
| **MJESEC** | **Ukupan broj objava** | **Broj objava vezanih za koronu** |
| 01 | 1986 | 578 |
| 02 | 1723 | 468 |
| 03 | 2006 | 571 |
| 04 | 1993 | 555 |
| 05 | 2095 | 474 |
| 06 | 2131 | 440 |
| 07 | 2361 | 475 |
| 08 | 2110 | 367 |
| 09 | 2152 | 376 |
| 10 | 2299 | 384 |
| 11 | 2443 | 604 |
| 12 | 2382 | 588 |

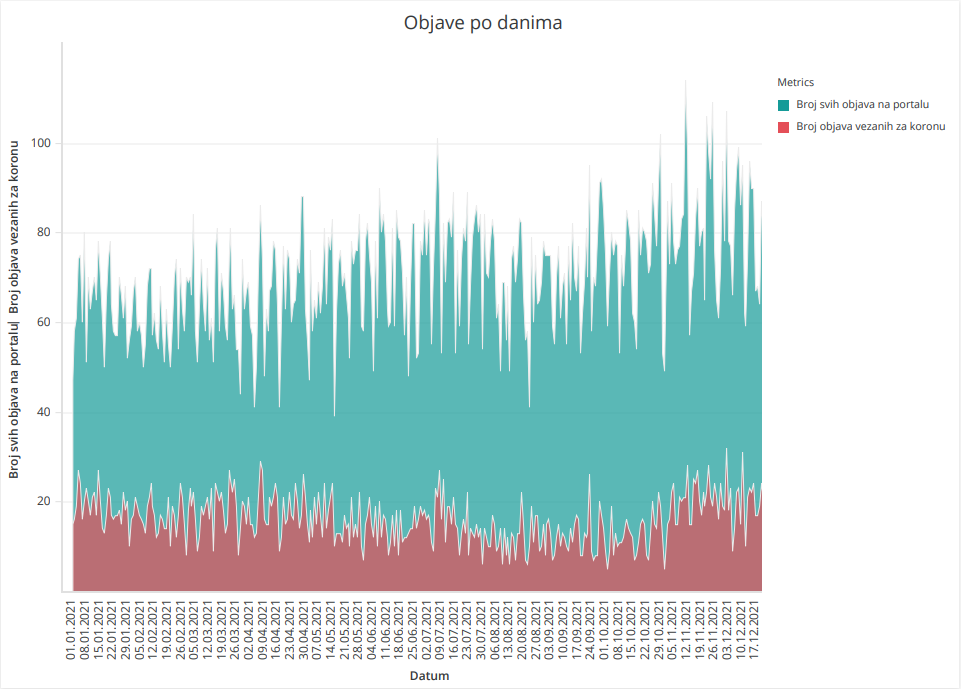
Također na sličan način radi i funkcija *newsCategory()*. Kreiran je DataFrame koji sad jedinstvene kategorije iz originalnog DataFrame-a prikupljenih podataka. Za svaku kategoriju pobrojani su članaci s istom kategorijom, a zatim i članci vezani uz korona tematiku. Niže se nalazi ispis dobivenih podataka:

Tablica - Objave po kategorijama

|  |  |  |
| --- | --- | --- |
| **KATEGORIJA** | **Broj objava** | **Broj objava vezanih za koronu** |
| Diadoratv | 225 | 35 |
| Politika | 736 | 156 |
| Dogadaji | 11093 | 4218 |
| Crna-kronika | 1262 | 39 |
| Kultura | 1597 | 157 |
| Biznis | 805 | 207 |
| Sci-tech | 791 | 297 |
| Sport | 4157 | 413 |
| Zabava | 1116 | 137 |
| Lifestyle | 3756 | 172 |
| Ostalo | 143 | 49 |

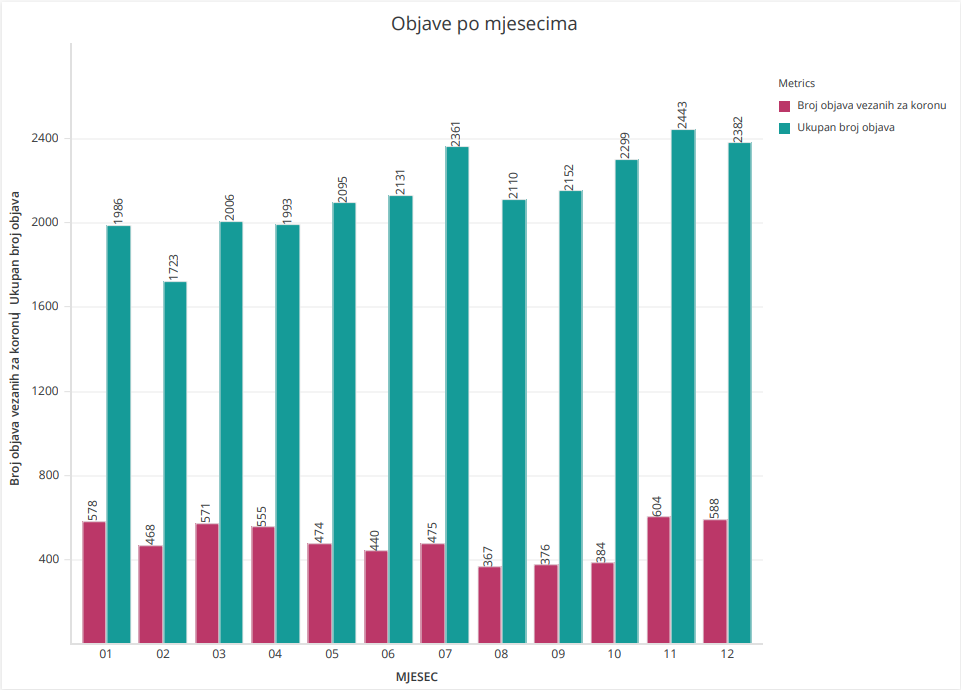
# Vizualizacija podataka

Vizualizacija podataka odrađena je u programu MicroStrategy. Prethodno ispisani podaci prikazani su u obliku grafova.



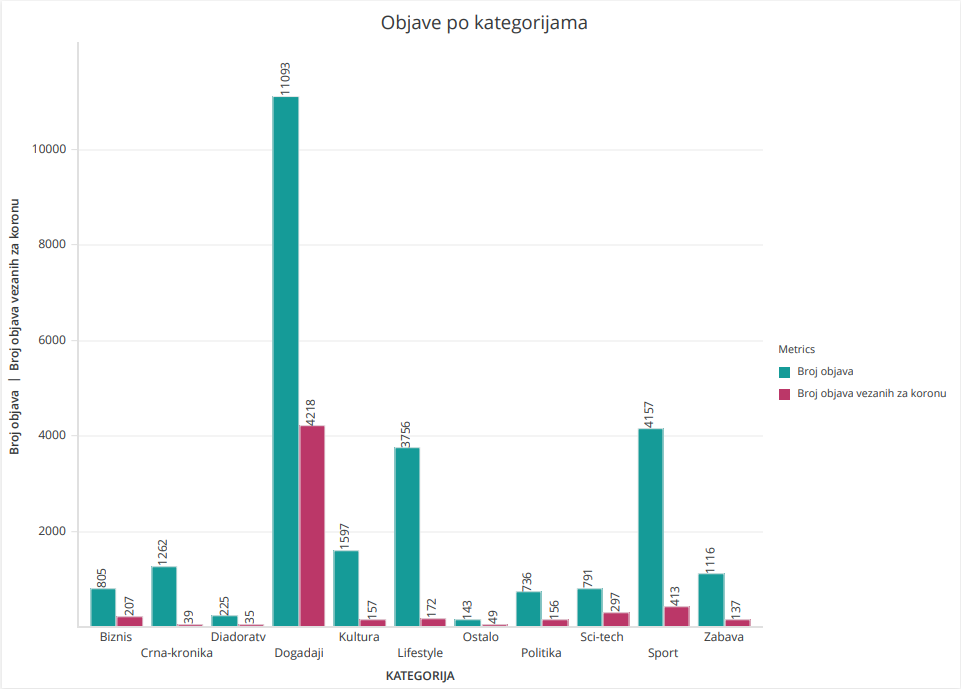
Slika - Objave po danima

Na datum 02.10.2021. godine objavljen je najmanji broj objava vezanih uz korona tematiku. Na taj dan objavljeno je 59 članak od kojih je 5 vezano za korona tematiku odnosno 8.47% ukupnih objava odnosilo se na korona tematiku. Na datum 02.12.2021. godine objavljeno je najveći broj objava vezanih uz korona tematiku. Na taj dan objavljeno je 107 članka od kojih je 32 vezano za korona tematiku odnosno 29.90% ukupnih objava odnosilo se na korona tematiku. Na datum 11.11.2021. godine objavljeno je najviše članak na portalu eZadar.net.hr odnosno objavljeno je 114 članak od kojih se 21 odnosio na korona tematiku što čini 18.42%.



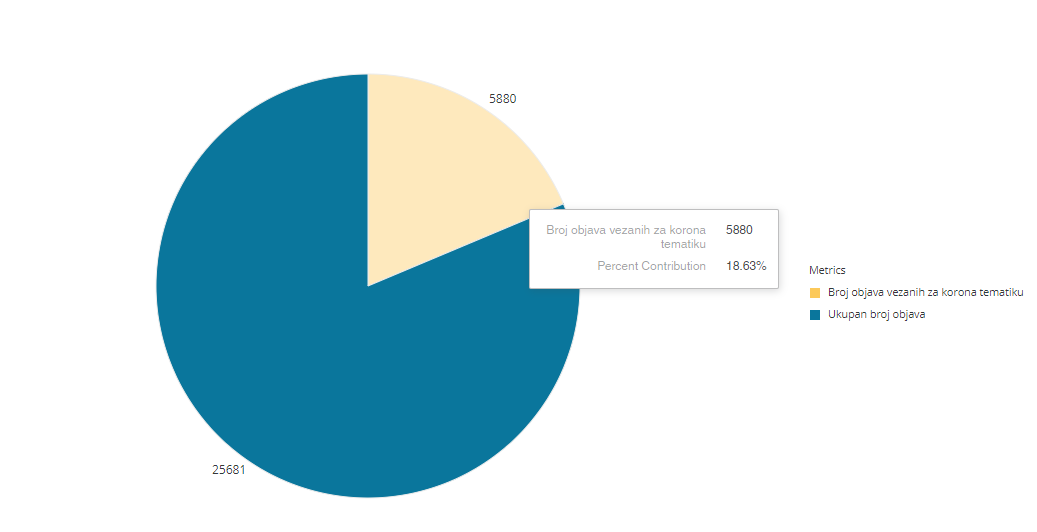
Slika - Objave po mjesecima

Iz prikazanog grafa zaključuje se da je u studenome objavljen najveći broj članaka odnosno 2443 članka od kojih se 604 odnosilo na korona tematiku što bi bilo 24.72%. To je ujedno i mjesec u kojem je objavljeno najviše članaka vezanih za korona tematiku. U veljači je objavljen najmanji broj članak u 2021. godini, čak 1723 članka od kojih se 27.16% odnosilo na korona tematiku. Najmanji broj članak vezanih za korona tematiku objavljeno je u kolovozu gdje je udio korona tematike na portalu iznosio 17.05%.



Slika - Objave po kategorijama

Najmnogobrojnija kategorija na portalu očigledno je kategorija događaji. U navedenoj kategoriji objavljeno je 11093 članaka od kojih je 38.02% vezanih uz korona tematiku. Najmanje članaka nalazi se pod kategorijom ostalo te se od 143 članka 34.27% odnosi na korona tematiku. Najmanje objavljenih članaka vezanih za korona tematiku nalazi se u kategoriji diadoratv te udio takvih članak u navedenoj kategoriji iznosi 15.55%.



Slika - Omjer ukupnih i objava vezanih uz korona tematiku

# Analiza jezičnog diskursa

Analiza jezičnog diskursa obuhvaća analizu najčešće korištenih pojmova, riječi i termina u objavama koje se analiziraju. Prije same analize potrebno je pročistiti podatke od simbola koje se mogu nalaziti unutar njih. Funkcijom *mostUsedWords()* pročišćeni su podaci te zatim pretraženi. Dobiveni rezultat bio je 150 riječi za svaki mjesec koji je pročišćen od zaustavnih riječi. Zaustavne riječi su one koje se najčešće koriste, a nisu ključne za značenje u prijenosu informacija. To su zamjenice, veznici, prilozi, prijedlozi, usklici i sl. Zaustavne riječi korištene u sklopu ovog projekta nalaze se u datoteci *stopWords.txt.* Zatim je izdvojeno 25 najčešće korištenih riječi pojedinačno za svaki mjesec. Niže su prikazani rezultati:



Slika - TOP 25 riječi za siječanj

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika - TOP 25 riječi za veljaču

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika - TOP 25 riječi za ožujak

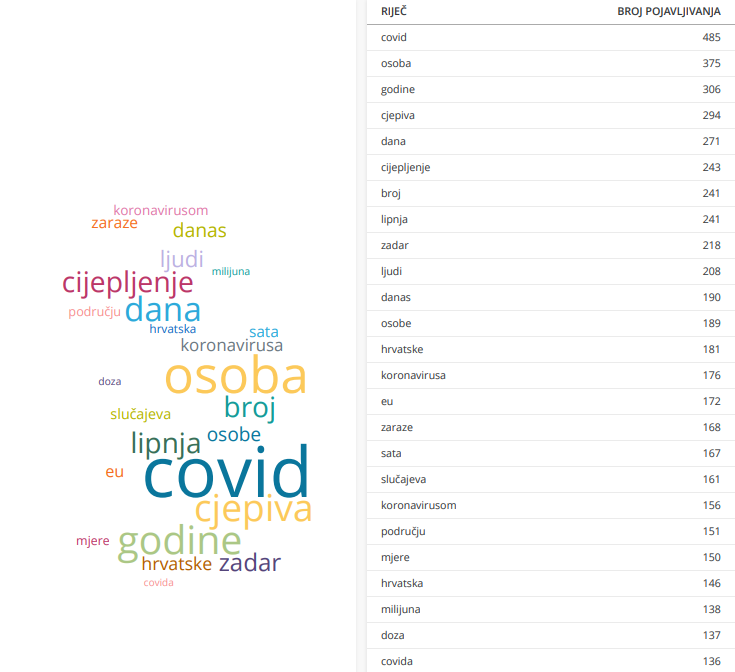
Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika - TOP 25 riječi za travanj



Slika - TOP 25 riječi za svibanj



Slika - TOP 25 riječi za lipanj

Slika na kojoj se prikazuje tekst

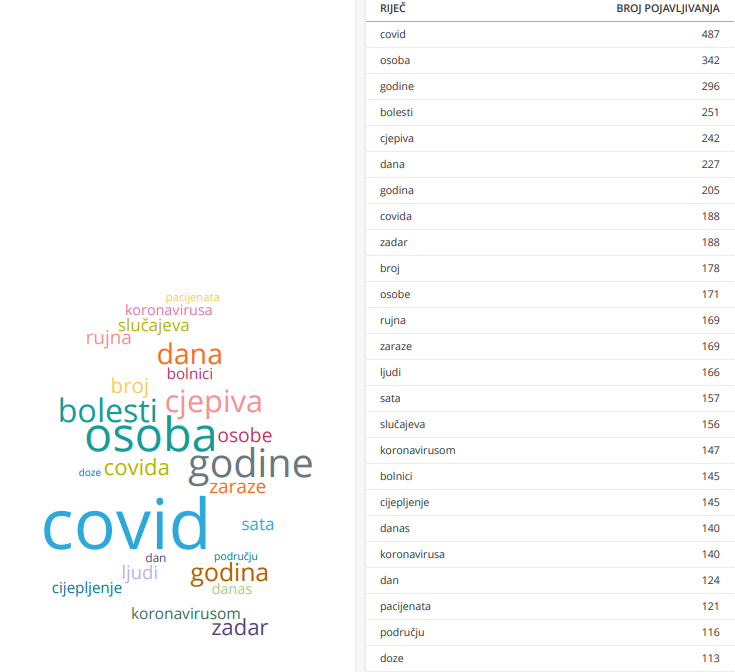
Opis je automatski generiran

Slika - TOP 25 riječi za srpanj

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika - TOP 25 riječi za kolovoz

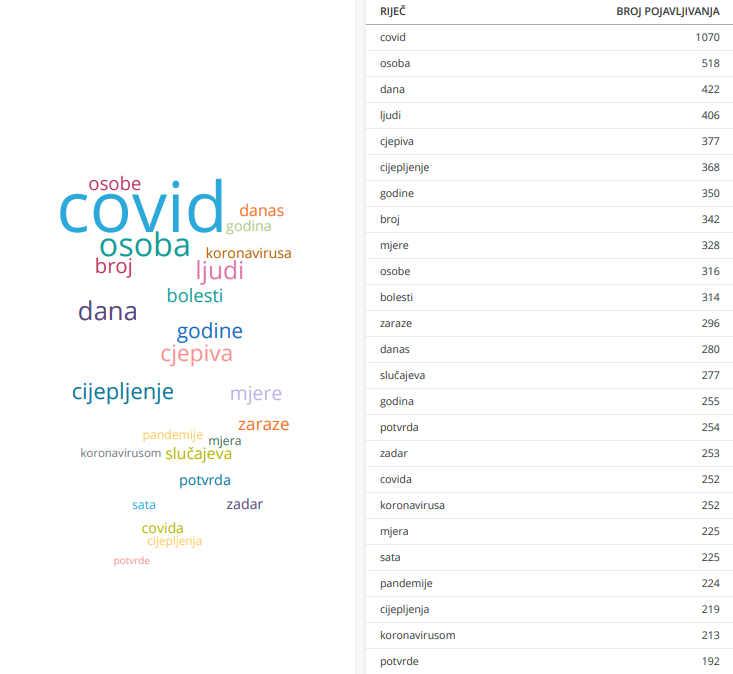


Slika - TOP 25 riječi za rujan

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika - TOP 25 riječi za listopad



Slika - TOP 25 riječi za studeni



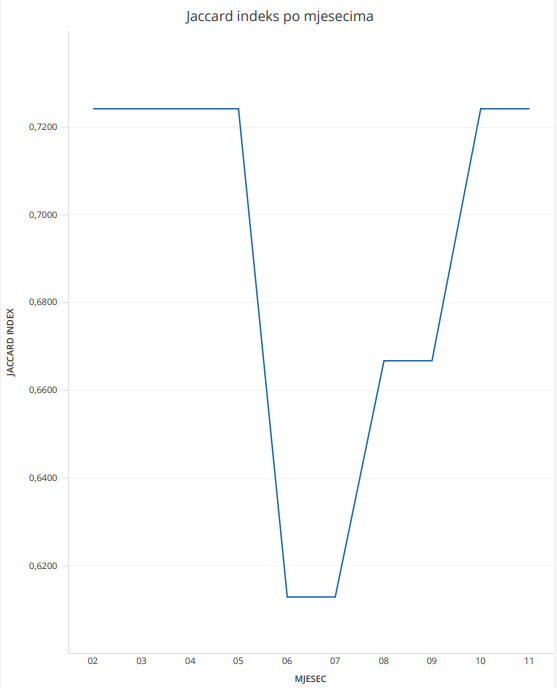
Slika - TOP 25 riječi za prosinac

Iz navedenih vizualnih prikaza da se zaključiti da se na portalu eZadar.net.hr većinom koriste iste riječi u člancima vezanim uz korona tematiku. Navedeno je potvrđeno izračunom Jaccard indeksa poznat kao Jaccardov koeficijent sličnosti. U funkciji *jaccardIndex(list1,list2)* definiran je način izračuna Jaccard indeksa te je prikazan niže. Jaccardov index mjeri sličnost između konačnih skupova uzoraka i definira se kao veličina presjeka podijeljena s veličinom unije uzoraka skupova.

Slika na kojoj se prikazuje tekst, na zatvorenom

Opis je automatski generiran

Slika - definiran Jaccard Indeks



Slika - Jaccard indeks po mjesecima

Iz prikazanog grafa moguće je zaključiti kako se promjene u jezičnom diskursu odnosno promjene u korištenim riječi nemaju prevelikih odstupanja. Vidljivo je da se događaju promjene u korištenim riječima u ljetnim mjesecima odnosno da su korištene riječi u preostalim mjesecima vrlo slične za razliku od ljetnih mjeseci kada je Jaccardov indeks niži.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Slika - Vrijednosti Jaccard indeksa

# Popis slika

[Slika 1 - Istraživanje HTML strukture 4](#_Toc93353345)

[Slika 2 - Funkcija za dohvaćanje kategorija 4](#_Toc93353346)

[Slika 3 - Datumska kontrola u funkciji 5](#_Toc93353347)

[Slika 4 - Transformacija prikupljenih podataka 6](#_Toc93353348)

[Slika 5 - Dobiveni ispis funkcija 6](#_Toc93353349)

[Slika 6 - Objave po danima 17](#_Toc93353350)

[Slika 7 - Objave po mjesecima 18](#_Toc93353351)

[Slika 8 - Objave po kategorijama 19](#_Toc93353352)

[Slika 9 - Omjer ukupnih i objava vezanih uz korona tematiku 19](#_Toc93353353)

[Slika 10 - TOP 25 riječi za siječanj 20](#_Toc93353354)

[Slika 11 - TOP 25 riječi za veljaču 21](#_Toc93353355)

[Slika 12- TOP 25 riječi za ožujak 22](#_Toc93353356)

[Slika 13 - TOP 25 riječi za travanj 23](#_Toc93353357)

[Slika 14 - TOP 25 riječi za svibanj 24](#_Toc93353358)

[Slika 15 - TOP 25 riječi za lipanj 25](#_Toc93353359)

[Slika 16 - TOP 25 riječi za srpanj 26](#_Toc93353360)

[Slika 17 - TOP 25 riječi za kolovoz 27](#_Toc93353361)

[Slika 18 - TOP 25 riječi za rujan 28](#_Toc93353362)

[Slika 19 - TOP 25 riječi za listopad 29](#_Toc93353363)

[Slika 20 - TOP 25 riječi za studeni 30](#_Toc93353364)

[Slika 21 - TOP 25 riječi za prosinac 31](#_Toc93353365)

[Slika 22 - definiran Jaccard Indeks 31](#_Toc93353366)

[Slika 23 - Jaccard indeks po mjesecima 32](#_Toc93353367)

[Slika 24 - Vrijednosti Jaccard indeksa 33](#_Toc93353368)

# Popis tablica

[Tablica 1 - Objave po datumima 7](#_Toc93353387)

[Tablica 2 - Objave po mjesecima 15](#_Toc93353388)

[Tablica 3 - Objave po kategorijama 16](#_Toc93353389)