Projektni zadatak - Upravljanje znanjem

Odjel za informatiku, Sveučilište u Rijeci
Akademska godina 2021/2022

Odabrani portal: https://hr.n1info.com/ (https://hr.n1info.com/)
Autor: Andrea Hrelja
Nositelji kolegija:

- izv. prof. dr. sc. Ana Meštrović
- dr. sc. Slobodan Beliga

Dokumentacija prikupljanja podataka dostupna je u korijenu ovog repozitorija u datoteci README.md (https://github.com/andhrelja/UZ-projekt/blob/master/README.md). U nastavku je implementiran i dokumentiran proces Analize podataka.

In [ ]:

```
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import pandas as pd
import os

ARTICLES_CLEAN_PATH = 'C:\\Users\\AndreaHrelja\\Documents\\Faks\\5. godina\\3. semestar
\\UPZ\\Scrapeinar\\scrape\\n1info\\csv\\clean-n1info.csv'
MONTH_WORDCOUNT_PATH = 'C:\\Users\\AndreaHrelja\\Documents\\Faks\\5. godina\\3. semesta
r\\UPZ\\Scrapeinar\\scrape\\n1info\\csv\\wordcount-n1info.csv'

df = pd.read_csv(ARTICLES_CLEAN_PATH, encoding='utf-8')
df['tags'] = df['tags'].map(eval)
df['categories'] = df['categories'].map(eval)

df = df[[
    'id', 'datetime', 'date', 'month', 'category_name',
    'title', 'author', 'text', 'categories', 'tags',
    'covid_related', 'vaccine_related', 'anti_related', 'soj_related'
]]

rename_map = {
    'covid_related': "Objave vezane uz korona tematiku",
    'vaccine_related': "Objave vezane uz tematiku cijepljenja",
    'anti_related': "Objave vezane uz tematiku antimaskera/antivaksera",
    'soj_related': "Objave vezane uz tematiku sojeva",
    'total': "Ukupno objava"
}

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50616 entries, 0 to 50615
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   id              50616 non-null  int64 
 1   datetime        50616 non-null  object
 2   date            50616 non-null  object
 3   month           50616 non-null  object
 4   category_name   50616 non-null  object
 5   title           50616 non-null  object
 6   author          50611 non-null  object
 7   text            50603 non-null  object
 8   categories      50616 non-null  object
 9   tags            50616 non-null  object
 10  covid_related   50189 non-null  object
 11  vaccine_related 50106 non-null  object
 12  anti_related    50027 non-null  object
 13  soj_related     50032 non-null  object
dtypes: int64(1), object(13)
memory usage: 5.4+ MB
```

U sklopu analize koriste se dvije CSV datoteke:

- clean-n1info.csv
- wordcount-n1info.csv

clean-n1info.csv

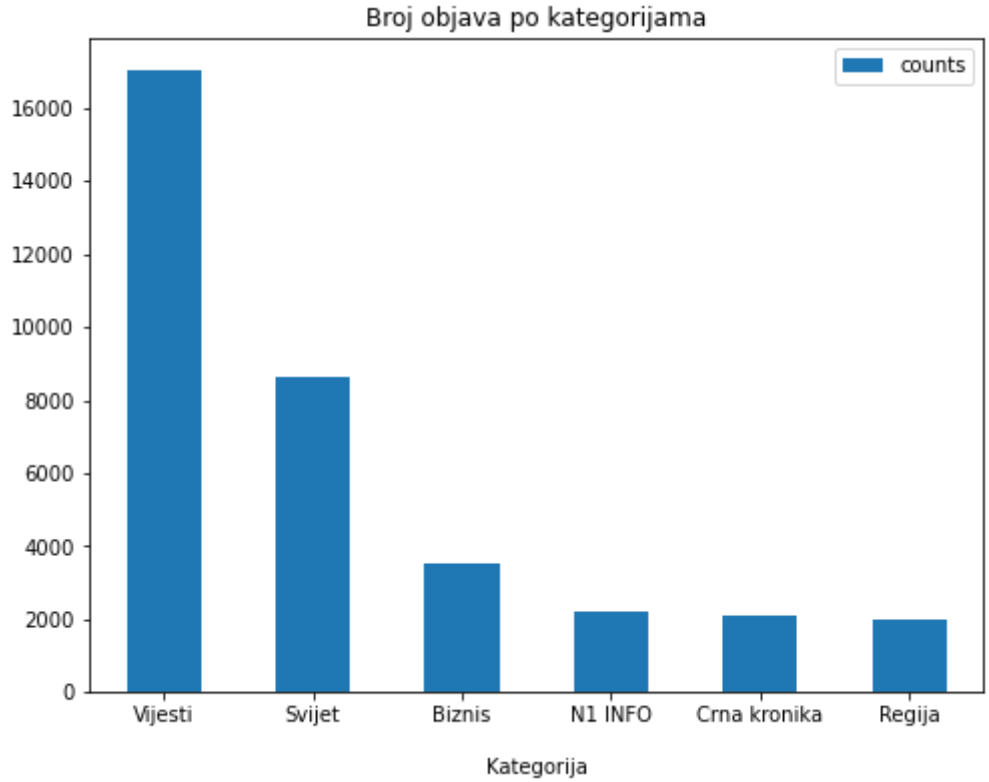
Naziv kolumne	Tip podatka	Opis podatka
id	int	Jedinstveni identifikator članka
datetime	datetime	Datum i vrijeme (YYYY-MM-DD HH:MM:SS) članka
date	date	Datum (YYYY-MM-DD) članka
month	date	Datum (YYYY-MM-01) članka
category_name	string	Naziv kategorije glavne članka
title	string	Naslov članka
author	string	Autor članka
text	string	Tekst članka
categories	string (list)	Kategorije članka (Python lista u obliku tekstualnog niza)
tags	string (list)	Tagovi članka (Python lista u obliku tekstualnog niza)
covid_related	boolean	Članak je vezan uz COVID tematiku
vaccine_related	boolean	Članak je vezan uz tematiku cijepljenja
anti_related	boolean	Članak je vezan uz tematiku antimaskera/antivaksera
soj_related	boolean	Članak je vezan uz tematiku novih sojeva

wordcount-n1info.csv

Naziv kolumne	Tip podatka	Opis podatka
month	date	Datum (YYYY-MM-01) članka
title	string	Konkatenacija naslova i teksta članka

In [ ]:

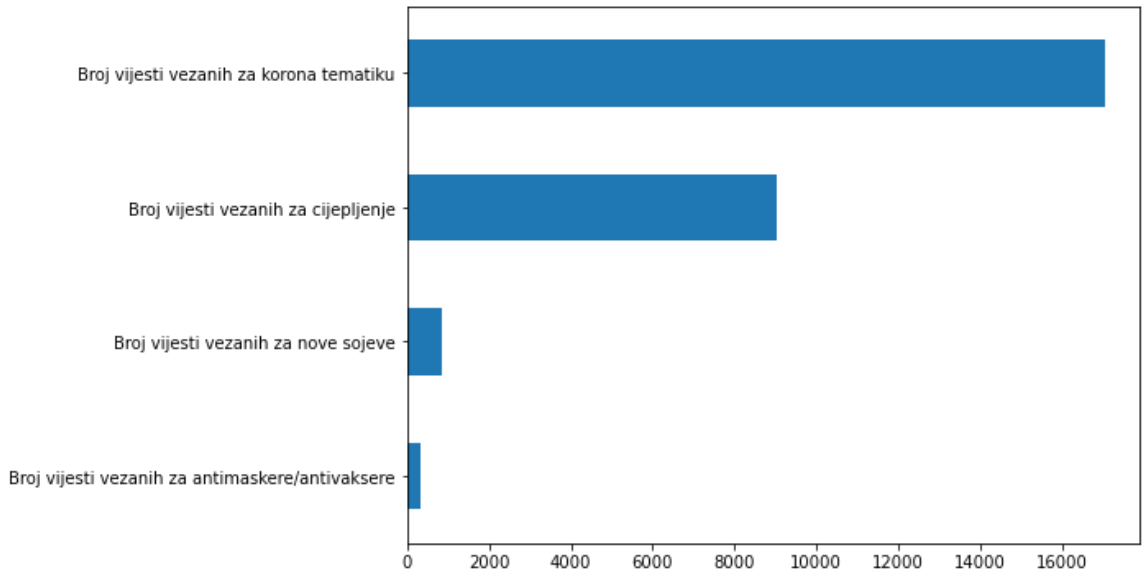
```
fig, ax = plt.subplots(1, figsize=(8, 6))

category_df = df.groupby('category_name').size() \
    .reset_index(name='counts') \
    .sort_values('counts', ascending=False) \
    .head(6)

most_common_categories = category_df['category_name'].unique()
category_df.plot(x='category_name', kind='bar', rot=0, ax=ax)

ax.set_title('Broj objava po kategorijama')
ax.set_xlabel('\nKategorija')

plt.show()
```



In [ ]:

```
num_anti_related    = len(df[df['anti_related'] == True])
num_soj_related     = len(df[df['soj_related'] == True])
num_vaccine_related = len(df[df['vaccine_related'] == True])
num_covid_related   = len(df[df['covid_related'] == True])

print('a) Ukupan broj objava na portalu za vremenski period 1.1.2021. - 31.12.2021.:
{:,.0f}'.format(len(df)))
print('b) Broj vijesti vezanih za korona tematiku: {:,.0f}'.format(num_covid_related))
print('c) Broj vijesti vezanih za cijepljenje: {:,.0f}'.format(num_vaccine_related))
print('d) Broj vijesti vezanih za antimaskere/antivaksere: {:,.0f}'.format(num_anti_rel
ated))
print('*) Broj vijesti vezanih za nove sojeve: {:,.0f}'.format(num_soj_related))

pd.DataFrame(
    {
        'count': [
            num_anti_related,
            num_soj_related,
            num_vaccine_related,
            num_covid_related,
        ]
    }, index=[
        'Broj vijesti vezanih za antimaskere/antivaksere',
        'Broj vijesti vezanih za nove sojeve',
        'Broj vijesti vezanih za cijepljenje',
        'Broj vijesti vezanih za korona tematiku'
    ]
).plot.barh(legend=None, figsize=(8, 6))

plt.show()
```

```
a) Ukupan broj objava na portalu za vremenski period 1.1.2021. - 31.12.202
1.: 50,616
b) Broj vijesti vezanih za korona tematiku: 17,036
c) Broj vijesti vezanih za cijepljenje: 9,048
d) Broj vijesti vezanih za antimaskere/antivaksere: 321
*) Broj vijesti vezanih za nove sojeve: 853
```



In [ ]:

```
day_df = df[['date', 'covid_related', 'anti_related', 'vaccine_related']].groupby('dat
e').sum()
day_df['total'] = df.groupby('date').size()
day_df = day_df[['total', 'covid_related', 'vaccine_related', 'anti_related']]
day_df.rename(rename_map, axis=1)
```

Out[ ]:

date	Ukupno objava	Objave vezane uz korona tematiku	Objave vezane uz tematiku cijepljenja	Objave vezane uz tematiku antimaskera/antivaksera
2021-01-01	31	11	3	0
2021-01-02	72	28	12	0
2021-01-03	74	27	9	0
2021-01-04	140	49	16	0
2021-01-05	134	51	30	0
...	...	...	...	...
2021-12-27	64	26	18	1
2021-12-28	101	30	19	1
2021-12-29	133	38	22	5
2021-12-30	149	52	26	5
2021-12-31	85	29	12	1

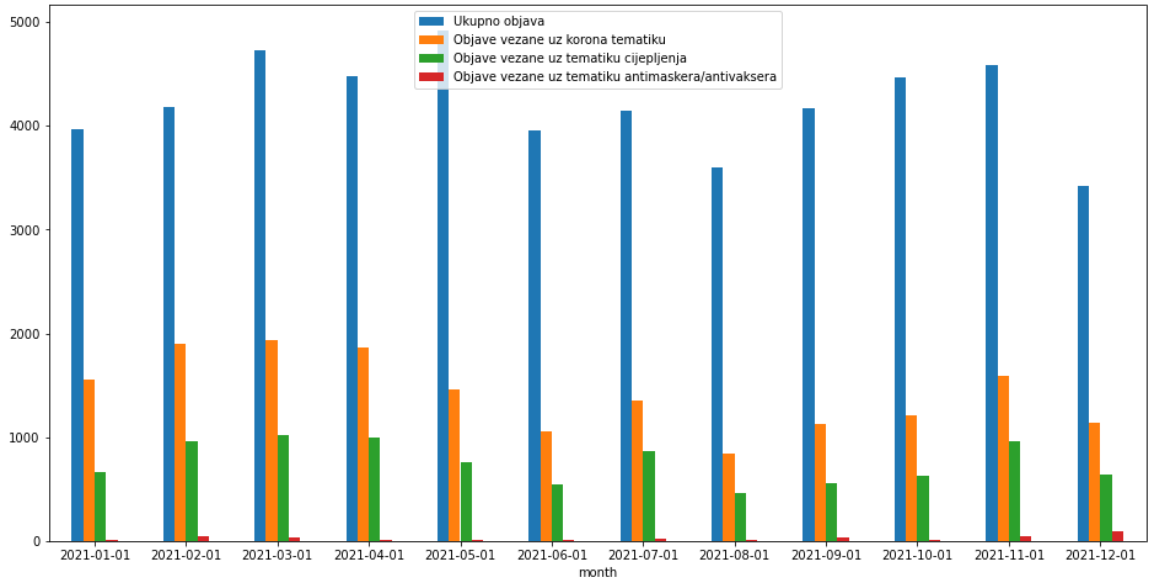
360 rows × 4 columns

In [ ]:

```
month_df = df[['month', 'covid_related', 'anti_related', 'vaccine_related']].groupby('m
onth').sum()
month_df['total'] = df.groupby('month').size()
month_df = month_df[['total', 'covid_related', 'vaccine_related', 'anti_related']].rena
me(rename_map, axis=1)
month_df.plot(kind='bar', rot=0, figsize=(16, 8))
```

Out[ ]:

```
<AxesSubplot:xlabel='month'>
```



In [ ]:

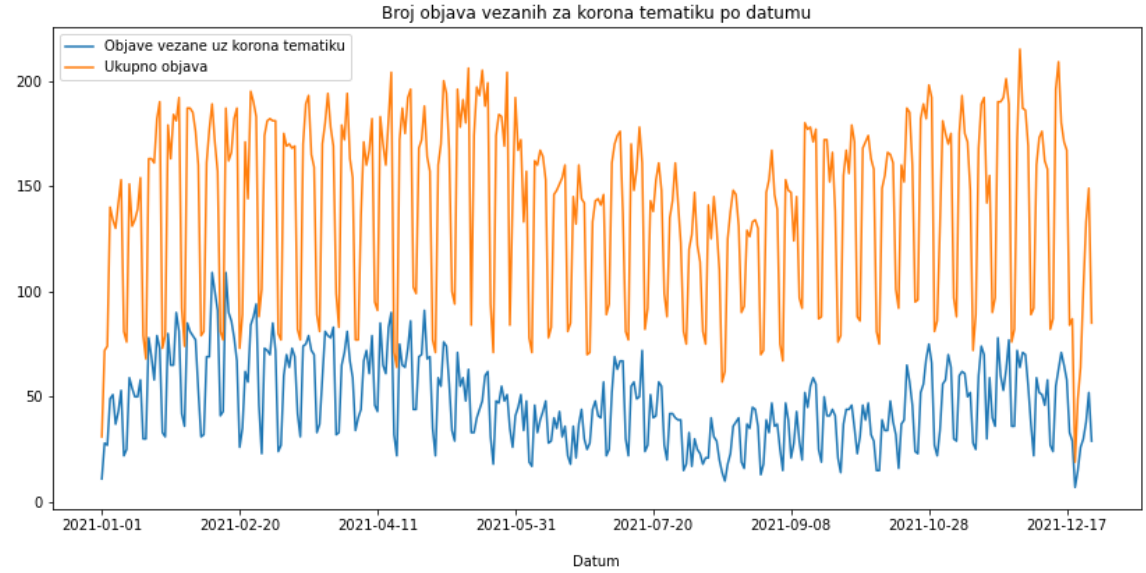
```
month_df.rename(rename_map, axis=1)
```

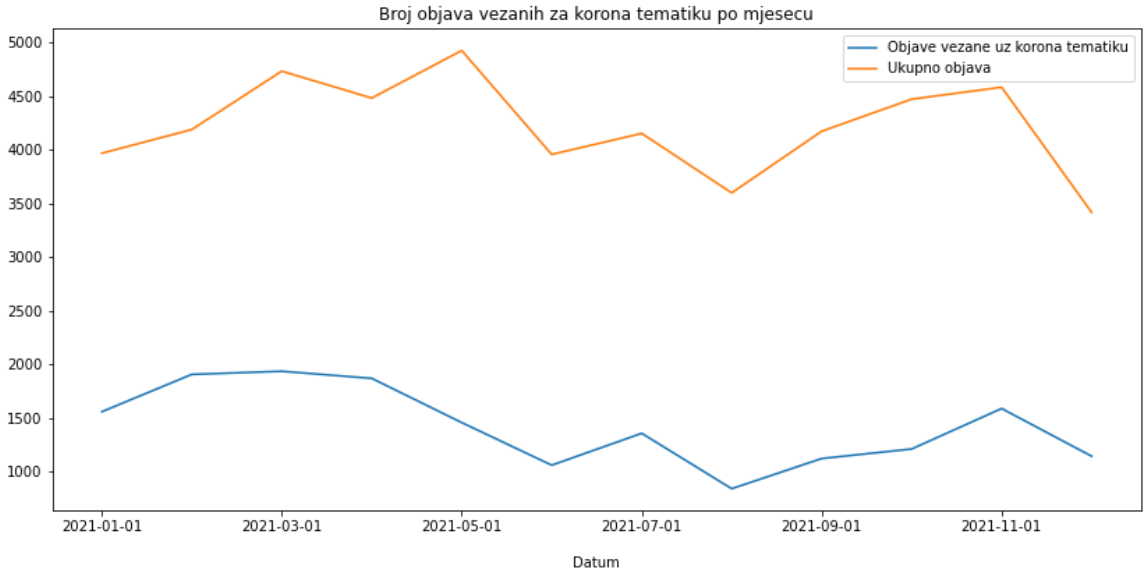
Out[ ]:

month	Ukupno objava	Objave vezane uz korona tematiku	Objave vezane uz tematiku cijepljenja	Objave vezane uz tematiku antimaskera/antivaksera
2021-01-01	3965	1557	665	6
2021-02-01	4186	1905	959	43
2021-03-01	4730	1934	1014	28
2021-04-01	4479	1868	997	13
2021-05-01	4922	1456	763	9
2021-06-01	3954	1059	548	6
2021-07-01	4150	1356	860	21
2021-08-01	3596	840	462	15
2021-09-01	4170	1121	558	30
2021-10-01	4469	1210	623	14
2021-11-01	4580	1587	965	46
2021-12-01	3415	1143	634	90

In [ ]:

```
fig, ax = plt.subplots(nrows=2, figsize=(14, 14))

df[df['covid_related'] == True].groupby('date').size() \
    .reset_index(name='counts') \
    .sort_values('date') \
    .set_index('date') \
    .plot(ax=ax[0], legend=None)

df.groupby('date').size() \
    .reset_index(name='counts') \
    .sort_values('date') \
    .set_index('date') \
    .plot(ax=ax[0], legend=None)


df[df['covid_related'] == True].groupby('month').size() \
    .reset_index(name='counts') \
    .sort_values('month') \
    .set_index('month') \
    .plot(ax=ax[1], legend=None)

df.groupby('month').size() \
    .reset_index(name='counts') \
    .sort_values('month') \
    .set_index('month') \
    .plot(ax=ax[1], legend=None)

ax[0].set_title('Broj objava vezanih za korona tematiku po datumu')
ax[0].set_xlabel('\nDatum')
ax[0].legend([rename_map[name] for name in ['covid_related', 'total']])
ax[1].set_title('Broj objava vezanih za korona tematiku po mjesecu')
ax[1].set_xlabel('\nDatum')
ax[1].legend([rename_map[name] for name in ['covid_related', 'total']])

plt.show()
```





Slijedeći grafikon (Word Cloud) prikazuje ukupnu frekvenciju pojavljivanja tagova u člancima za period od 1.1.2021 - 31.12.2021.

In [ ]:

```
cattags_df = df[['date', 'month', 'category_name', 'tags']].dropna() \
    .explode('tags') \
    .reset_index(drop=True) \
    .set_index('category_name')

tag_frequency = cattags_df.groupby('tags').size() \
    .reset_index(name='counts') \
    .sort_values('counts', ascending=False) \
    .set_index('tags')

fig, ax = plt.subplots(figsize=(14, 7))
wordcloud = WordCloud(width=800, height=500, background_color='white').generate_from_fr
equencies(tag_frequency.to_dict()['counts'])
ax.imshow(wordcloud)
ax.axis('off')
ax.set_title('Ukupna frekvencija pojavljivanja tagova u člancima')
plt.show()
```

Slijedeći grafikon (Word Cloud) prikazuje frekvenciju pojavljivanja tagova u člancima po pojedinoj kategoriji za period od 1.1.2021 - 31.12.2021.

In [ ]:

```
counts_df = cattags_df.groupby(['category_name', 'tags']).size() \
    .reset_index(name='counts') \
    .sort_values('counts', ascending=False) \
    .set_index('category_name')

print("Frekvencija pojavljivanja tagova za pojedinu kategoriju")
ncols = 2
nrows = len(most_common_categories) / ncols
fig, axs = plt.subplots(nrows=int(nrows), ncols=ncols, figsize=(nrows*5, ncols*7))
for i, category_name in enumerate(most_common_categories):
    frequencies = {
        item['tags']: item['counts']
        for item in counts_df.loc[category_name].to_dict(orient='records')
    }
    wordcloud = WordCloud(background_color='white').generate_from_frequencies(frequenci
es)

    axs[i // 2, i % 2].set_title("Kategorija: " + category_name)
    axs[i // 2, i % 2].imshow(wordcloud, interpolation='bilinear')
    axs[i // 2, i % 2].axis('off')
```

```
Frekvencija pojavljivanja tagova za pojedinu kategoriju
```

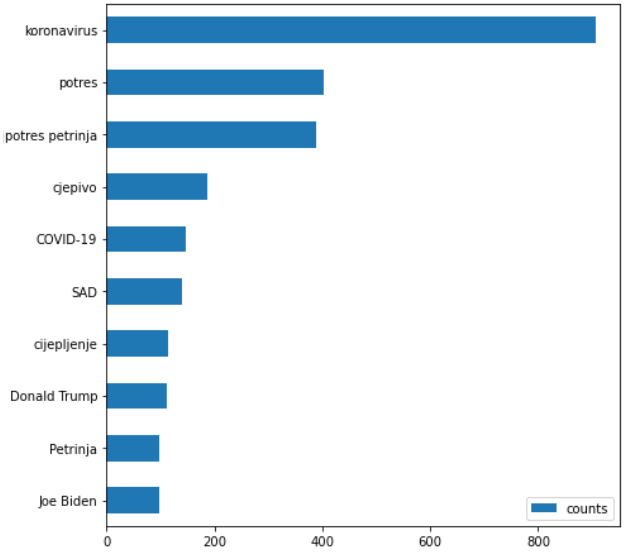
Slijedeći grafikon (Word Cloud) prikazuje frekvenciju pojavljivanja top 25 tagova u člancima po mjesecu za period od 1.1.2021 - 31.12.2021. S desne je strane "bar" graf koji prikazuje frekvenciju pojavljivanja top 10 tagova u člancima po mjesecu.

In [ ]:

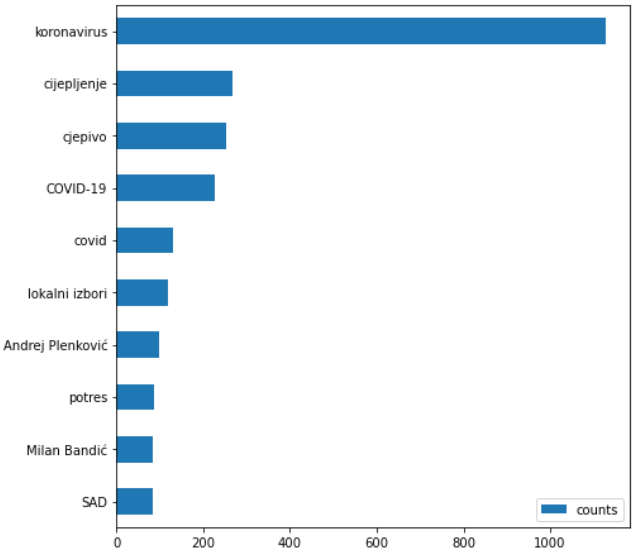
```
monthtags_df = cattags_df.groupby(['month', 'tags']).size() \
    .reset_index(name='counts') \
    .sort_values('counts', ascending=False) \
    .set_index('month')

print("Frekvencija pojavljivanja tagova po mjesecu")
ncols = 2
nrows = len(monthtags_df.index.unique())
fig, axs = plt.subplots(nrows=int(nrows), ncols=ncols, figsize=(ncols*8, nrows*9))
for i, month in enumerate(sorted(monthtags_df.index.unique())):
    month_items = monthtags_df.loc[month]
    frequencies = {
        item['tags']: item['counts']
        for item in month_items.to_dict(orient='records')[:25]
    }
    wordcloud = WordCloud(background_color='white', height=400, width=600).generate_fro
m_frequencies(frequencies)

    axs[i, 0].set_title("Datum: " + str(month)[:7])
    axs[i, 0].imshow(wordcloud, interpolation='bilinear')
    axs[i, 0].axis('off')

    month_items.head(10).sort_values('counts').plot.barh(x='tags', ax=axs[i, 1])
    axs[i, 1].set_ylabel(None)
```

Frekvencija pojavljivanja tagova po mjesecu

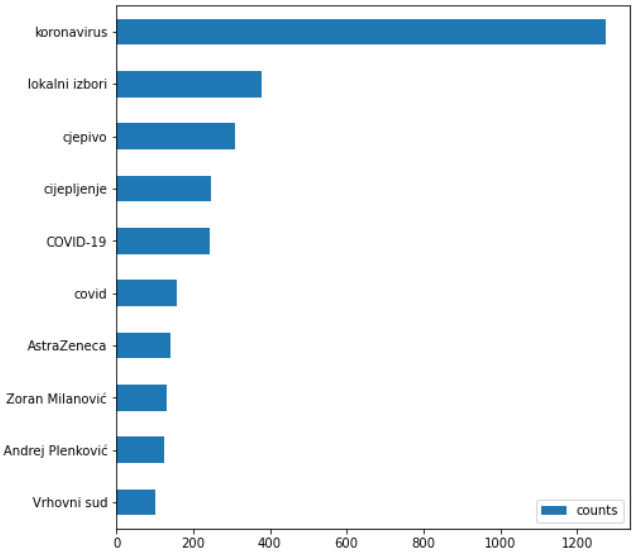

Datum: 2021-01
potres u Petrinji
koronavirus
Glina
potres petrinja
Sisak
Slovenija
Njemačka
Ivo Žinić
policija
Joe Biden
potres
covid
cijepljenje
HDZ
Žarko Tušek
SAD
koronavirus cjepivo
cjepivo
Donald Trump
AstraZeneca
sabor
COVID-19
Rusija
Petrinja
Andrej Plenković


koronavirus
potres
potres petrinja
cjepivo
COVID-19
SAD
cijepljenje
Donald Trump
Petrinja
Joe Biden
counts
0
200
400
600
800

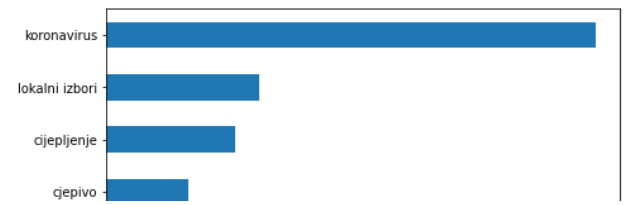

Datum: 2021-02
SDP
HGK
policija
HDZ
Slovenija
Donald Trump
AstraZeneca
SAD
koronavirus
Andrej Plenković
epidemiološke mjere
cjepivo
covid
Joe Biden
Rusija
potres petrinja
potres
Zoran Milanović
cijepljenje
Njemačka
COVID-19
lokalni izbori
Milan Bandić
Zagreb
horoskop

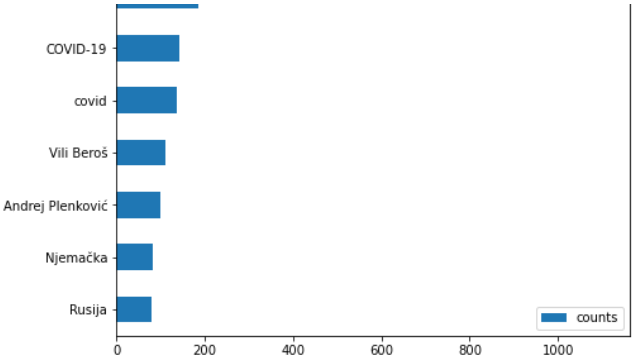

koronavirus
cijepljenje
cjepivo
COVID-19
covid
lokalni izbori
Andrej Plenković
potres
Milan Bandić
SAD
counts
0
200
400
600
800
1000


Datum: 2021-03
Vrhovni sud
Slovenija
turizam
lokalni izbori
Njemačka
potres
koronavirus
Milan Bandić
SAD
cijepljenje
cjepivo
BiH
Tomislav Tomašević
Srbija
COVID-19
Zdravko Mamić
distribucija N1 u Hrvatskoj
Zlata Đurđević
AstraZeneca
Zagreb
Zdravlje
Andrej Plenković
covid
Zoran Milanović
HDZ

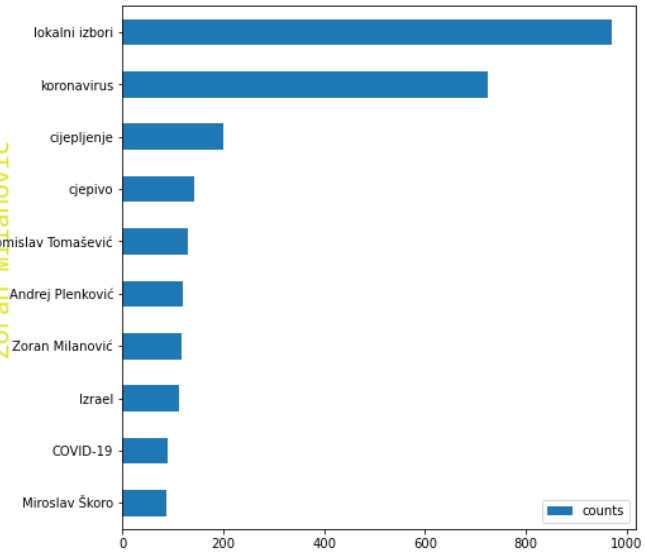

koronavirus
lokalni izbori
cjepivo
cijepljenje
COVID-19
covid
AstraZeneca
Zoran Milanović
Andrej Plenković
Vrhovni sud
counts
0
200
400
600
800
1000
1200


Datum: 2021-04
COVID-19
AstraZeneca
lokalni izbori
Njemačka
policija

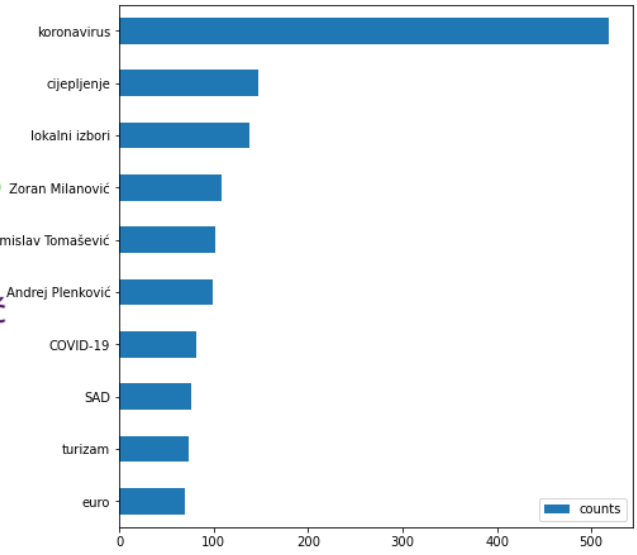

koronavirus
lokalni izbori
cijepljenje
cjepivo



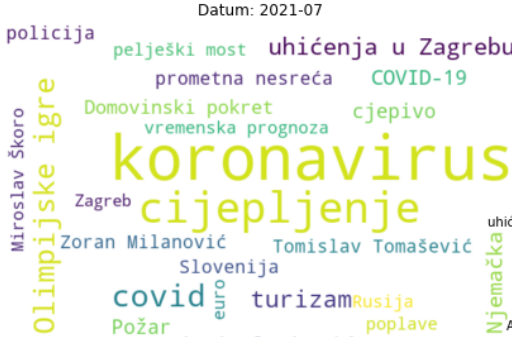
Datum: 2021-05

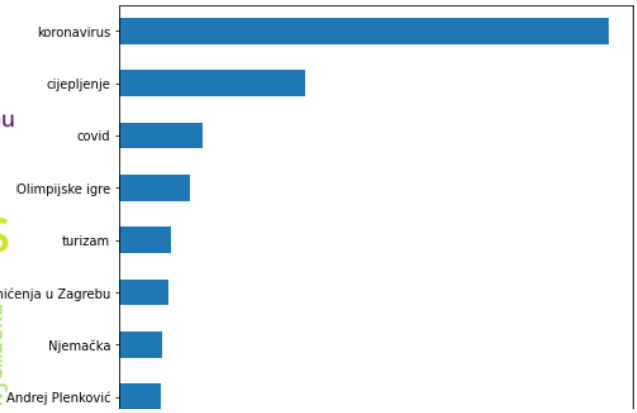


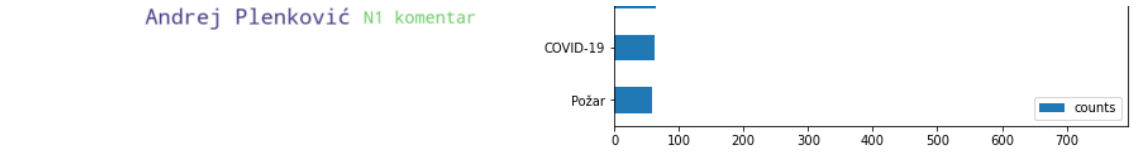
Datum: 2021-06

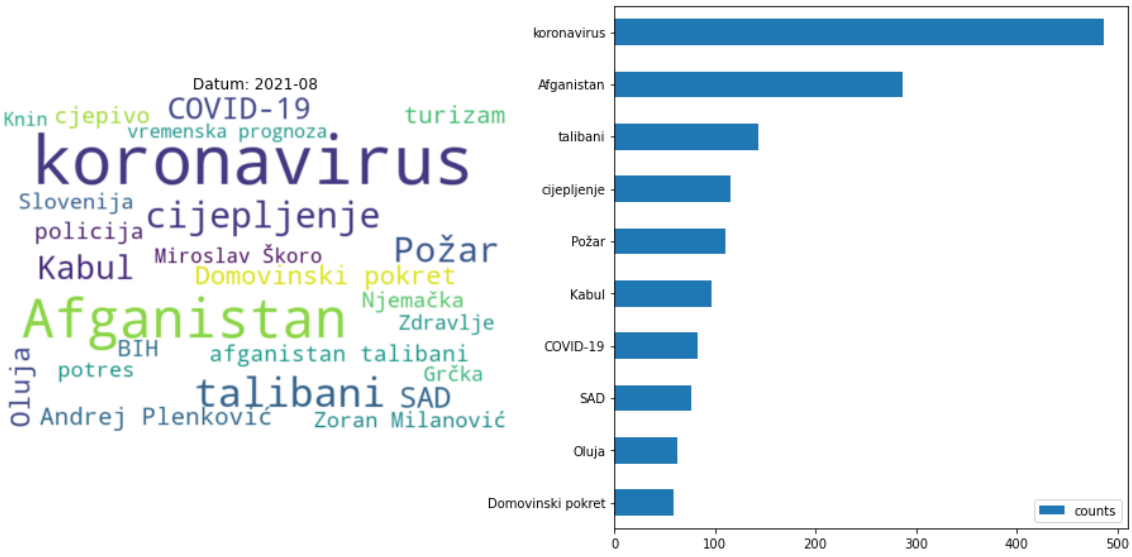


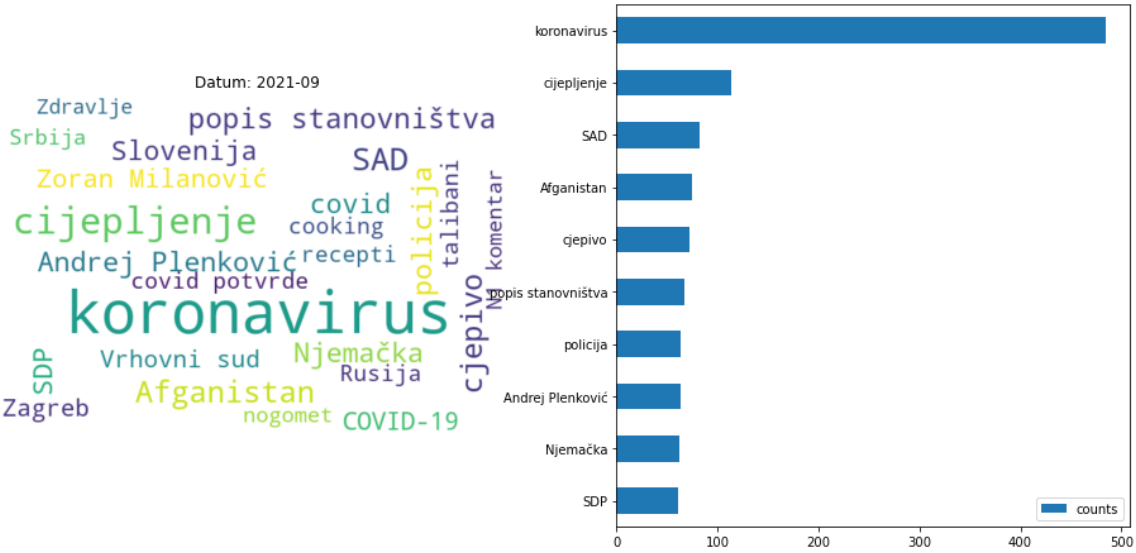
Datum: 2021-07

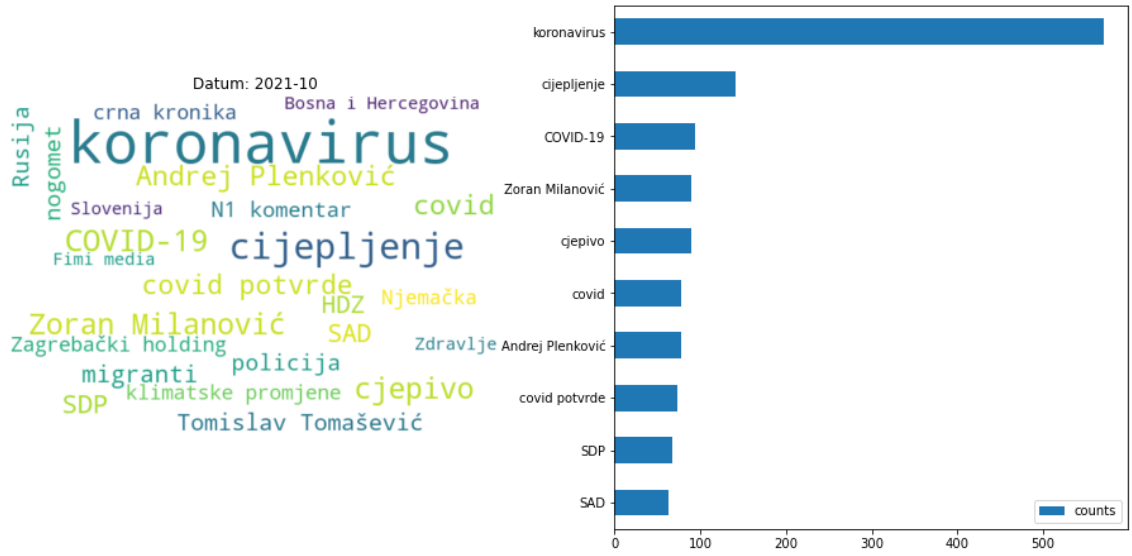


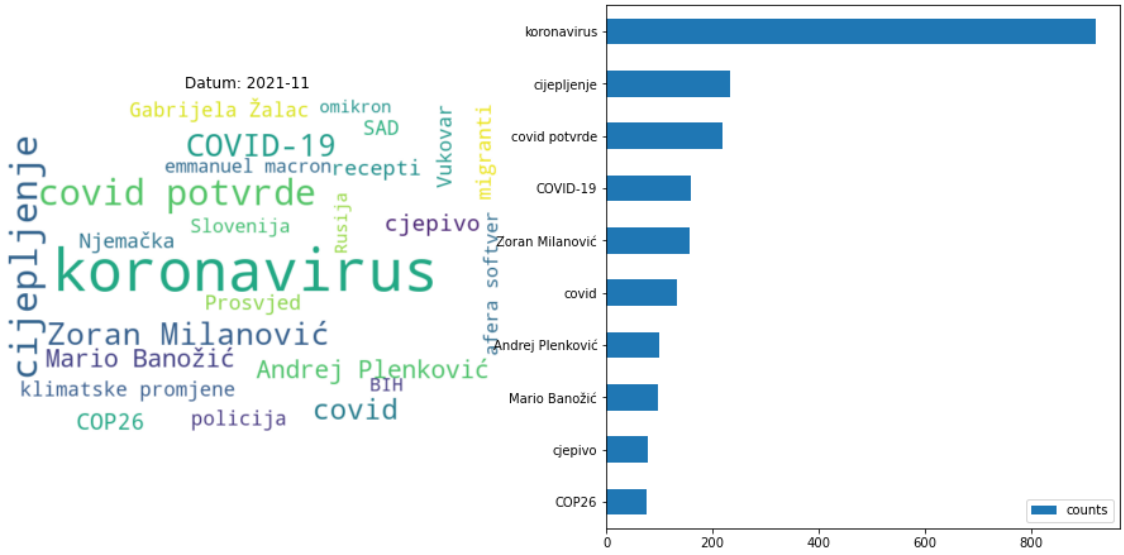


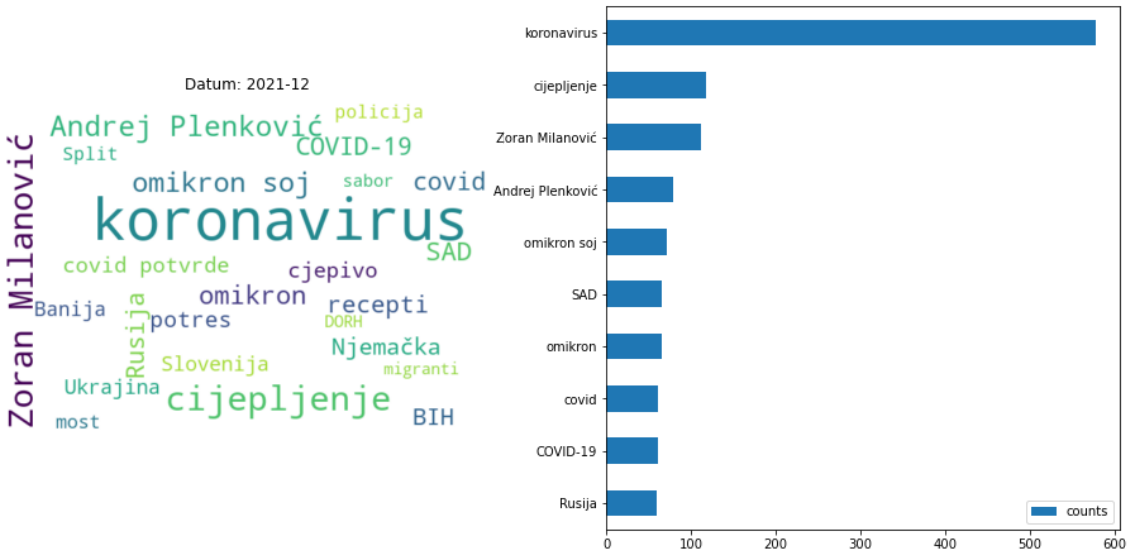

Andrej Plenković
N1 komentar
COVID-19
Požar
counts


Datum: 2021-08
koronavirus
cijepljenje
COVID-19
Afganistan
Kabul
Požar
talibani
SAD
Domovinski pokret
Njemačka
Zdravlje
Slovenija
policija
Miroslav Škoro
turizam
Knin
cjepivo
vremenska prognoza
BIH
potres
Oluja
Andrej Plenković
Zoran Milanović
Grčka
afganistan talibani
koronavirus
Afganistan
talibani
cijepljenje
Požar
Kabul
COVID-19
SAD
Oluja
Domovinski pokret
counts


Datum: 2021-09
koronavirus
cijepljenje
SAD
Afganistan
cjepivo
popis stanovništva
policija
Andrej Plenković
Njemačka
SDP
Zdravlje
Srbija
Slovenija
Zoran Milanović
cijepljenje
Andrej Plenković
covid potvrde
Vrhovni sud
Zagreb
Afganistan
nogomet
COVID-19
Njemačka
Rusija
covid
cooking
recepti
policija
talibani
komentar
cjepivo
counts


Datum: 2021-10
koronavirus
cijepljenje
COVID-19
Zoran Milanović
cjepivo
covid
Andrej Plenković
covid potvrde
SDP
SAD
Rusija
nogomet
crna kronika
Bosna i Hercegovina
Andrej Plenković
Slovenija
N1 komentar
COVID-19
Fimi media
cijepljenje
covid potvrde
HDZ
Njemačka
Zoran Milanović
Zagrebački holding
SAD
Zdravlje
migranti
policija
SDP
klimatske promjene
cjepivo
Tomislav Tomašević
counts





Primijećuje se da je u svakom mjesecu tag "koronavirus" najčešći, osim u svibnju (2021-05) kada je tag "lokalni izbori" najzastupljenija vijest.

U nastavku slijedi tablični prikaz unigrama za top 25 tagova u člancima po mjesecu za period od 1.1.2021 - 31.12.2021.

In [ ]:

```
for i, month in enumerate(sorted(monthtags_df.index.unique())):
    print("Ispis unigrama za {}".format(str(month)[:7]))
    print(monthtags_df.loc[month][:25].reset_index(drop=True))
    print()
```

Ispis unigrama za 2021-01

	tags	counts
0	koronavirus	906
1	potres	403
2	potres petrinja	388
3	cjepivo	187
4	COVID-19	147
5	SAD	141
6	cijepljenje	115
7	Donald Trump	113
8	Joe Biden	97
9	Petrinja	97
10	Sisak	75
11	Glina	67
12	Andrej Plenković	64
13	Njemačka	59
14	potres u Petrinji	58
15	Ivo Žinić	52
16	Slovenija	52
17	Žarko Tušek	47
18	HDZ	45
19	koronavirus cjepivo	43
20	policija	43
21	AstraZeneca	38
22	Rusija	37
23	covid	37
24	sabor	36

Ispis unigrama za 2021-02

	tags	counts
0	koronavirus	1127
1	cijepljenje	267
2	cjepivo	254
3	COVID-19	227
4	covid	131
5	lokalni izbori	118
6	Andrej Plenković	100
7	potres	87
8	SAD	84
9	Milan Bandić	84
10	Njemačka	80
11	AstraZeneca	62
12	Zoran Milanović	62
13	HDZ	55
14	potres petrinja	54
15	Slovenija	50
16	Rusija	49
17	SDP	45
18	policija	42
19	Zagreb	40
20	Joe Biden	39
21	HGK	37
22	Donald Trump	36
23	horoskop	36
24	epidemiološke mjere	36

Ispis unigrama za 2021-03

	tags	counts
0	koronavirus	1274
1	lokalni izbori	378
2	cjepivo	310

3	cijepljenje	248
4	COVID-19	242
5	covid	158
6	AstraZeneca	142
7	Zoran Milanović	130
8	Andrej Plenković	124
9	Vrhovni sud	101
10	Milan Bandić	101
11	potres	74
12	HDZ	71
13	SAD	69
14	distribucija N1 u Hrvatskoj	64
15	Zagreb	59
16	Srbija	56
17	Njemačka	55
18	Zdravlje	52
19	Slovenija	50
20	Zdravko Mamić	48
21	Zlata Đurđević	46
22	BIH	45
23	Tomislav Tomašević	44
24	turizam	43

Ispis unigrama za 2021-04

	tags	counts
0	koronavirus	1107
1	lokalni izbori	345
2	cijepljenje	291
3	cjepivo	186
4	COVID-19	142
5	covid	137
6	Vili Beroš	111
7	Andrej Plenković	99
8	Njemačka	83
9	Rusija	80
10	AstraZeneca	74
11	Zdravlje	63
12	SAD	62
13	Zagreb	60
14	Zoran Milanović	60
15	Slovenija	56
16	turizam	53
17	potres	53
18	policija	53
19	ozlijeđeno dijete	43
20	Jasenovac	41
21	Joe Biden	41
22	veledrogerije	40
23	sabor	39
24	Superliga	39

Ispis unigrama za 2021-05

	tags	counts
0	lokalni izbori	969
1	koronavirus	723
2	cijepljenje	201
3	cjepivo	142
4	Tomislav Tomašević	129
5	Andrej Plenković	119
6	Zoran Milanović	118
7	Izrael	112

8	COVID-19	91
9	Miroslav Škoro	87
10	Zagreb	82
11	Palestina	75
12	covid	70
13	SDP	63
14	HDZ	63
15	Njemačka	59
16	Zdravlje	55
17	SAD	55
18	turizam	55
19	policija	54
20	nogomet	50
21	Split	46
22	Slovenija	46
23	Gaza	44
24	Dijana Zadravec	44

Ispis unigrama za 2021-06

	tags	counts
0	koronavirus	518
1	cijepljenje	147
2	lokalni izbori	138
3	Zoran Milanović	108
4	Tomislav Tomašević	102
5	Andrej Plenković	99
6	COVID-19	82
7	SAD	76
8	turizam	73
9	euro	70
10	euro 2020	65
11	cjepivo	61
12	Zagreb	59
13	Njemačka	57
14	nogomet	54
15	Rusija	52
16	SDP	51
17	europsko nogometno prvenstvo 2020.	49
18	Joe Biden	48
19	Split	46
20	Hrvatska nogometna reprezentacija	45
21	Zdravlje	44
22	N1 komentar	41
23	covid	40
24	potres	39

Ispis unigrama za 2021-07

	tags	counts
0	koronavirus	757
1	cijepljenje	287
2	covid	128
3	Olimpijske igre	110
4	turizam	80
5	uhićenja u Zagrebu	77
6	Njemačka	67
7	Andrej Plenković	65
8	COVID-19	62
9	Požar	59
10	cjepivo	58
11	Domovinski pokret	54
12	prometna nesreća	51

13	policija	49
14	Slovenija	48
15	Tomislav Tomašević	47
16	Zoran Milanović	47
17	poplave	44
18	pelješki most	42
19	euro	41
20	Rusija	40
21	vremenska prognoza	39
22	N1 komentar	39
23	Miroslav Škoro	38
24	Zagreb	36

Ispis unigrama za 2021-08

	tags	counts
0	koronavirus	486
1	Afganistan	286
2	talibani	143
3	cijepljenje	115
4	Požar	111
5	Kabul	96
6	COVID-19	83
7	SAD	77
8	Oluja	63
9	Domovinski pokret	59
10	turizam	52
11	Andrej Plenković	52
12	afganistan talibani	46
13	cjepivo	46
14	policija	46
15	BIH	45
16	Njemačka	41
17	potres	41
18	Zoran Milanović	41
19	Slovenija	40
20	Zdravlje	38
21	Grčka	38
22	Miroslav Škoro	36
23	vremenska prognoza	36
24	Knin	35

Ispis unigrama za 2021-09

	tags	counts
0	koronavirus	484
1	cijepljenje	114
2	SAD	82
3	Afganistan	75
4	cjepivo	72
5	popis stanovništva	67
6	Andrej Plenković	64
7	policija	64
8	Njemačka	63
9	SDP	61
10	Slovenija	61
11	covid	61
12	Zoran Milanović	57
13	COVID-19	55
14	Vrhovni sud	53
15	Zagreb	52
16	talibani	50
17	covid potvrde	48

18	Rusija	48
19	recepti	47
20	cooking	46
21	N1 komentar	43
22	Srbija	42
23	nogomet	42
24	Zdravlje	39

Ispis unigrama za 2021-10

	tags	counts
0	koronavirus	570
1	cijepljenje	142
2	COVID-19	94
3	cjepivo	89
4	Zoran Milanović	89
5	Andrej Plenković	78
6	covid	78
7	covid potvrde	74
8	SDP	68
9	SAD	63
10	HDZ	57
11	migranti	55
12	Tomislav Tomašević	53
13	nogomet	53
14	policija	49
15	Rusija	49
16	N1 komentar	47
17	klimatske promjene	43
18	crna kronika	42
19	Zagrebački holding	42
20	Njemačka	39
21	Bosna i Hercegovina	38
22	Zdravlje	37
23	Fimi media	37
24	Slovenija	36

Ispis unigrama za 2021-11

	tags	counts
0	koronavirus	922
1	cijepljenje	233
2	covid potvrde	218
3	COVID-19	159
4	Zoran Milanović	157
5	covid	132
6	Andrej Plenković	99
7	Mario Banožić	97
8	cjepivo	78
9	COP26	76
10	recepti	69
11	Prosvjed	63
12	migranti	62
13	Njemačka	60
14	Vukovar	56
15	Gabrijela Žalac	55
16	klimatske promjene	55
17	SAD	53
18	policija	52
19	Slovenija	48
20	emmanuel macron	48
21	afera softver	46
22	BIH	42

```
23              Rusija      40
24             omikron      38

Ispis unigrama za 2021-12
                tags  counts
0         koronavirus     577
1         cijepljenje     117
2    Zoran Milanović      111
3   Andrej Plenković       79
4         omikron soj       72
5             omikron       65
6                 SAD       65
7           COVID-19        61
8              covid        61
9             Rusija        60
10           recepti        57
11          Njemačka        51
12               BIH        50
13            potres        48
14           cjepivo        46
15     covid potvrde        45
16          Slovenija       42
17          Ukrajina        42
18            Banija        40
19              most        38
20          policija        36
21             Split        35
22             sabor        33
23          migranti        31
24              DORH        30
```

Slijedeći blok definira dijakritičke znakove, znakove koji će se izbrisati iz ukupnog zbira tekstova i naslova članaka (`replace_chars`) te zaustavne riječi koje su odabrane ručnim pregledom unigrama iz ukupnog zbira tekstova i naslova članaka.

In [ ]:

```
diacritics = {
    'č': 'c',
    'ć': 'c',
    'š': 's',
    'ž': 'z'
}
replace_chars = [
    '”','\t','\n','!','#','$','%','&','/','(',')','[',']','{','}','=',
    '?','*','\\','|','€','÷','×',',','.','_','@',':','"',';',':','\xa0'
]
stop_words = [
    'protiv', 'hrvatskoj', 'ljudi', 'dana', 'novi', 'novih', 'posto', 'komentar', 'nov
e', 'evo', 'traži', 'biste', 'broj',
    'je','u','i','da','se','na','za','su','od','s','će','a','koji','o','to','ne','što',
'kako','bi','putem','te','iz','do',
    'nije','koje','biti','rekao','kao','ali','ili','koja','zbog','sve','smo','može','-'
,'po','jer','sa','još','ako','bilo',
    'li','oko','ima','prema','bio','sam','dok','kada','mogu','pa','prije','već','nego',
'nisu','kod','uz','treba','ih','mi',
    'osoba','sada','tako','bez','tri','no','ga','neće','kaže','nema','rekla','toga','dv
a','ćemo','bila','danas','jedan',
    'gdje','kojima','kazao','kad','između','tome','godina','oni','ni','imaju','ove','ti
jekom','među','taj','tu','on',
    'nekoliko','koju','radi','dvije','pitanje','svoje','vrijeme','svi','dio','im','dalj
e','bih','nešto','osobe','prvi',
    'je','jedna','bili','također','njih','mu','odnosno','vrlo','ovo','uvijek','način',
'četiri','n1','tom','dodao','mora',
    'kojoj','druge','imamo','onda','24','koliko','ponedjeljak','ja','nas','kojem','godi
ne','neki','poput','pet','smatra',
    'odnosu','pod','navodi','dan','ono','neke','tko','preko','kroz','čak','kuna','drug
i','piše','one','sata','utorak',
    'srijedu','ta','istaknuo','suda','svoj','ovaj','mislim','ona','malo','niti','tim',
'petak','jednom','ovom','četvrtak',
    'sad','kojih','bude','svim','pri','vam','bit','tek','šest','čega','mogao','možete',
'svojim','sati','svoju','znači',
    'mogli','tjedna','riječ','obzirom','osim','zato','zašto','ste','iako','tog','žele',
'čemu','time','stvari','naše',
    'njima','toj','tada','komentirao','kojeg','kazala','doći','tiče','mogla','tih','na
d','isto','imao','drugom','jednu',
    'svom','bismo','se','ipak','bile','ova','n','samo','nam','imati','moći', 'nakon','v
iše','dosad','milijuna'
]
stop_words_no_diacritics = []
for stop_word in stop_words:
    if any(key in stop_word for key in diacritics.keys()):
        for key, value in diacritics.items():
            stop_word = stop_word.replace(key, value)
        stop_words_no_diacritics.append(stop_word)

def get_word_count(text):
    word_count = {}
    for char in replace_chars:
        text = text.replace(char, ' ')
    all_text = text.lower().split(" ")
    for word in all_text:
        if word and not word.isnumeric() and word not in stop_words and word not in sto
p_words_no_diacritics:
            try:
                word_count[word] += 1
```

```
            except KeyError:
                word_count[word] = 1
    sorted_word_count = {k: v for k, v in sorted(word_count.items(), key=lambda x: x[1
], reverse=True)[:25]}
    return sorted_word_count
```

Slijedeći blok konkatenira sve naslove i tekstove na razini mjeseca u ukupan zbir naslova i tekstova članaka. Zbog dugog izvođenja agregacije, sadržaj je spremljen u wordcount-n1info.csv.

In [ ]:

```
if not os.path.isfile(MONTH_WORDCOUNT_PATH):
    month_titletext_df = df[['month', 'title', 'text']].groupby('month').sum()
    month_titletext_df.to_csv(MONTH_WORDCOUNT_PATH, encoding='utf-8')
else:
    month_titletext_df = pd.read_csv(MONTH_WORDCOUNT_PATH, encoding='utf-8', index_col=
'month')
```

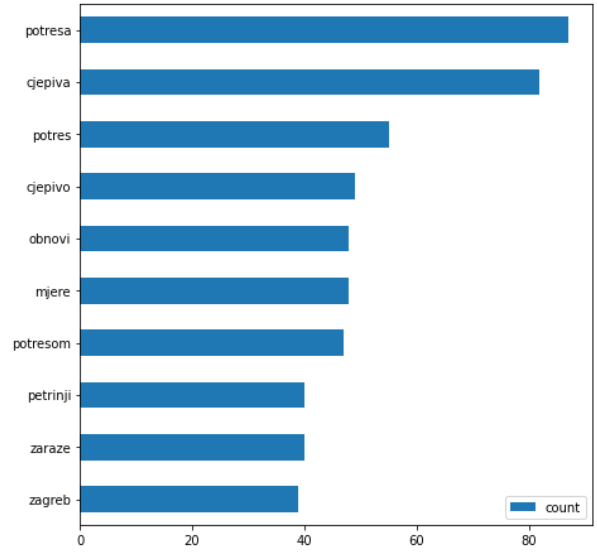
U nastavku slijedi Word Cloud prikaz top 25 najčešćih pojavljivanja riječi ukupnom zbiru naslova i tekstova članaka po mjesecu za period od 1.1.2021 - 31.12.2021. S desne je strane "bar" graf koji prikazuje frekvenciju pojavljivanja top 10 riječi u člancima po mjesecu.

In [ ]:

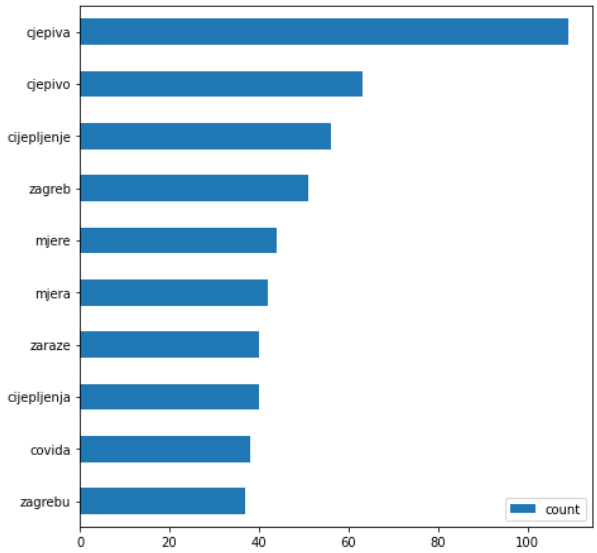
```
print("Najčešće riječi po mjesecu")
ncols = 2
nrows = len(month_titletext_df.index.unique())
fig, axs = plt.subplots(nrows=int(nrows), ncols=ncols, figsize=(ncols*8, nrows*9))
for i, month in enumerate(sorted(month_titletext_df.index.unique())):
    month_items = month_titletext_df.loc[month]
    frequencies = get_word_count(month_items['title'])
    wordcloud = WordCloud(background_color='white', height=400, width=600).generate_fro
m_frequencies(frequencies)

    axs[i, 0].set_title("Datum: " + str(month)[:7])
    axs[i, 0].imshow(wordcloud, interpolation='bilinear')
    axs[i, 0].axis('off')

    pd.DataFrame(frequencies.values(), frequencies.keys(), columns=['count']) \
        .head(10) \
        .sort_values('count') \
        .plot.barh(ax=axs[i, 1])
    axs[i, 1].set_ylabel(None)
```

Najčešće riječi po mjesecu

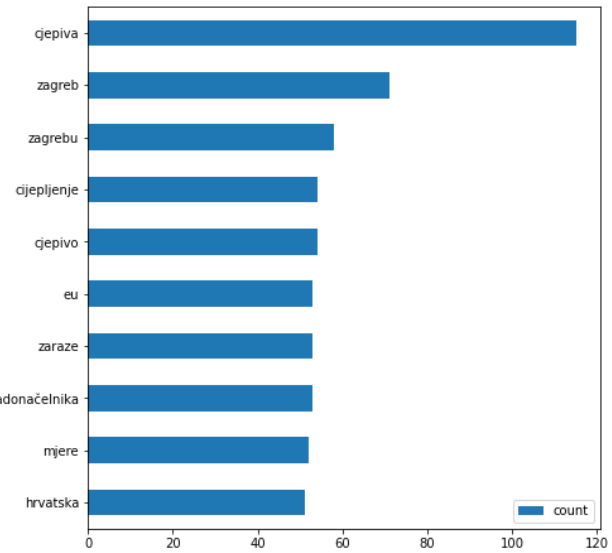
Datum: 0



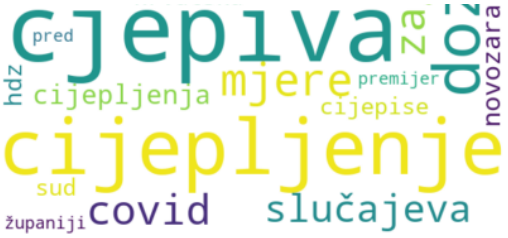
Datum: 1

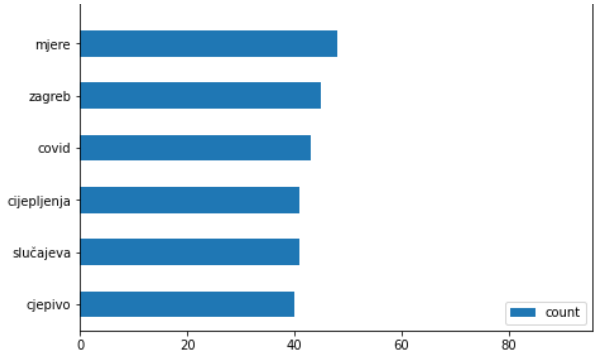


Datum: 2

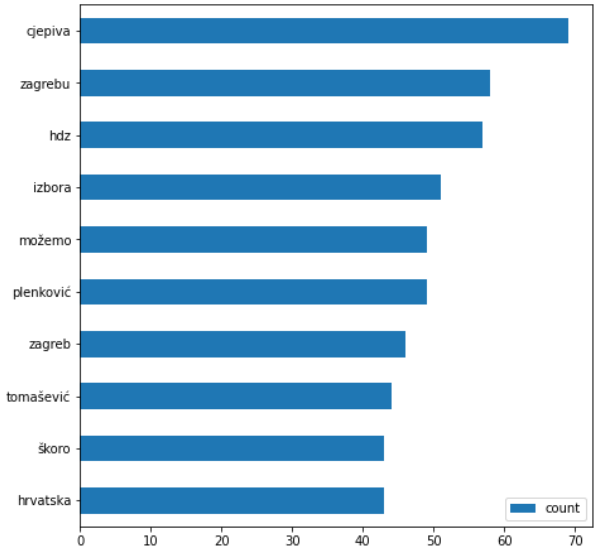


Datum: 3

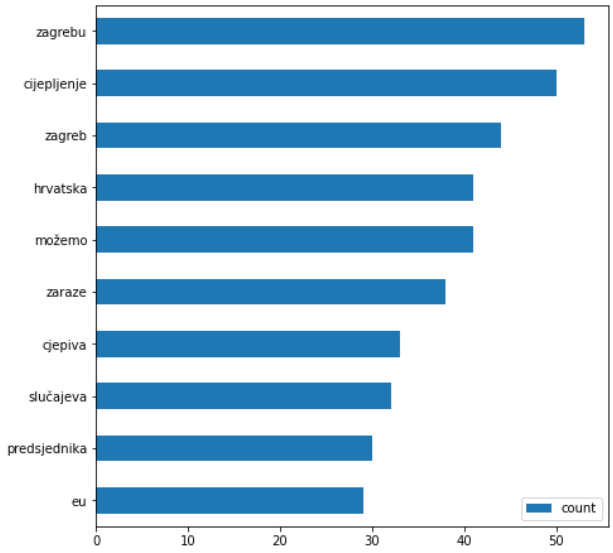




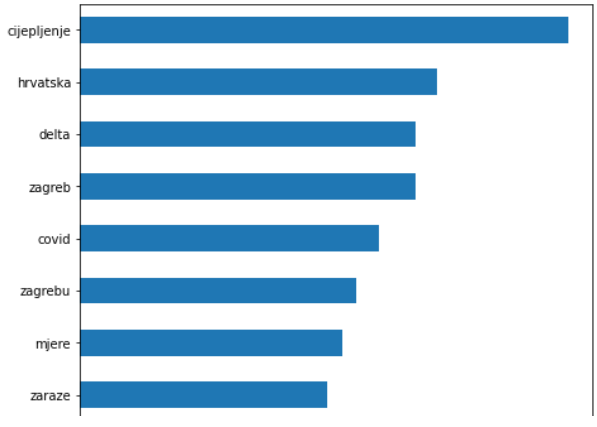
Datum: 4

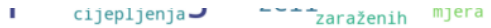


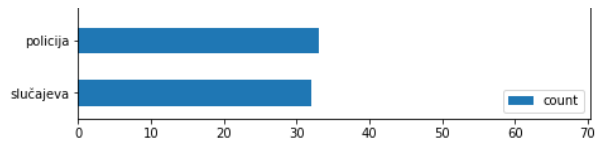
Datum: 5



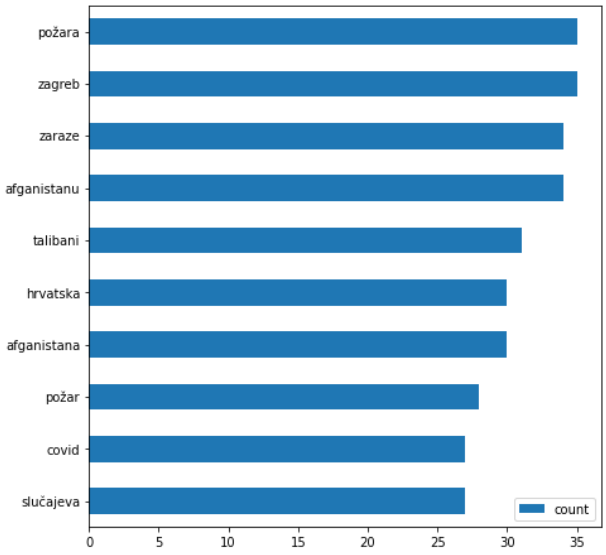
Datum: 6



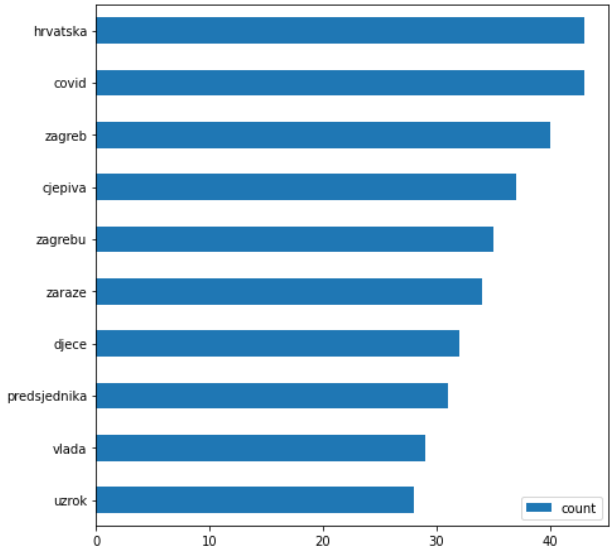




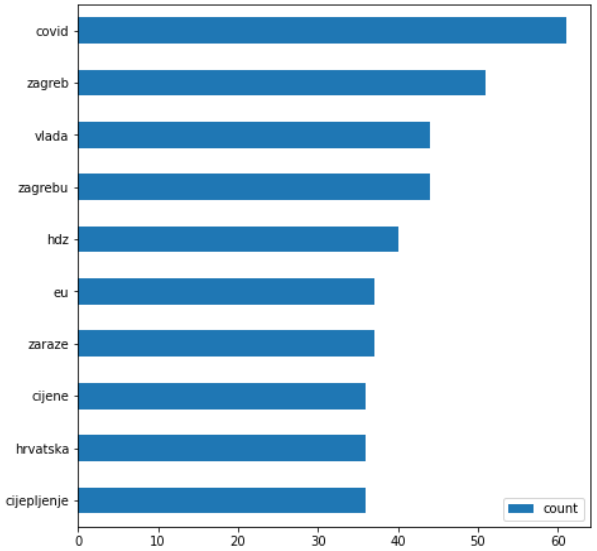
Datum: 7

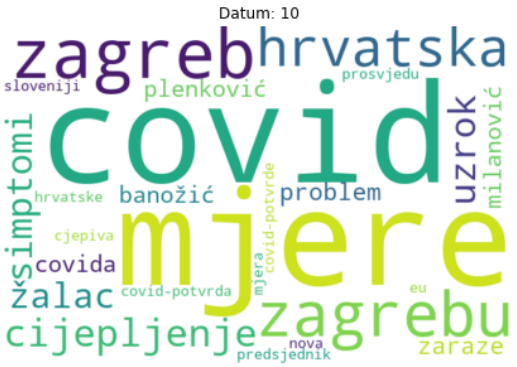


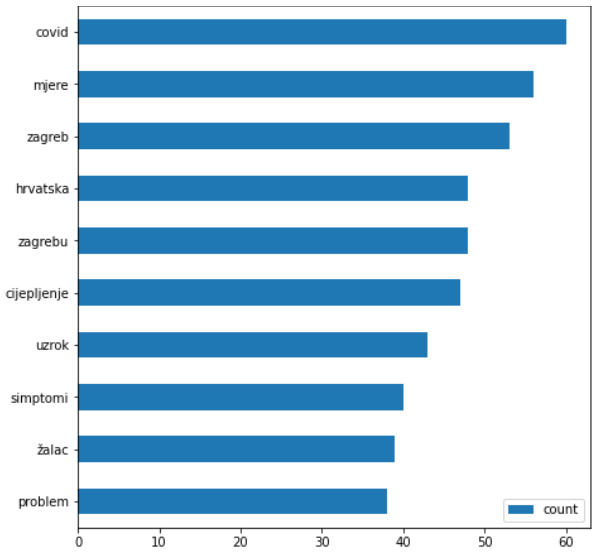
Datum: 8

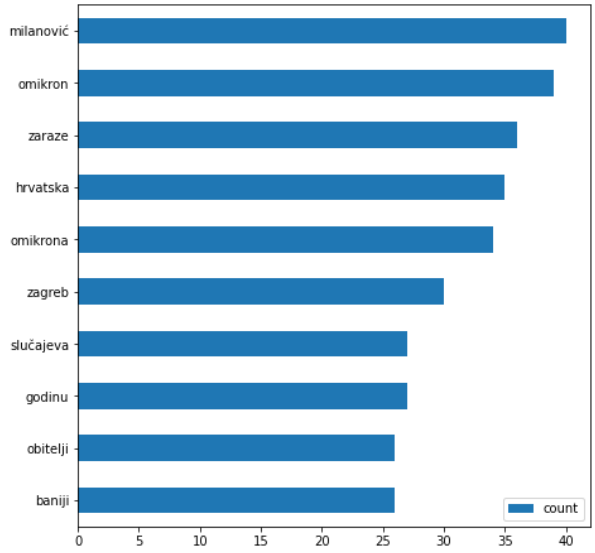


Datum: 9









Slijedeći blok izračunava i prikazuje Jaccardov indeks sličnosti za svaka dva mjeseca u razdoblju od 1.1.2021 do 31.12.2021. Svaki mjesec na x-osi predstavlja Jaccardov indeks sličnosti između ispisanog i prethodnog mjeseca.

In [ ]:

```
def jaccard_similarity(list1, list2):
    s1 = set(list1)
    s2 = set(list2)
    return float(len(s1.intersection(s2)) / len(s1.union(s2)))

index = month_titletext_df.index[1:]
data = {'jaccard_similarity': []}
for i, month in enumerate(month_titletext_df.index):
    if i == 0:
        continue
    prev_month = month_titletext_df.index[i-1]
    curr_month = month_titletext_df.index[i]
    prev_month_text = month_titletext_df.loc[prev_month]['title']
    curr_month_text = month_titletext_df.loc[curr_month]['title']

    prev_word_count = get_word_count(prev_month_text)
    curr_word_count = get_word_count(curr_month_text)
    data['jaccard_similarity'].append(jaccard_similarity(prev_word_count.keys(), curr_w
ord_count.keys()))

pd.DataFrame(data=data, index=index).plot(kind='bar', rot=30, figsize=(14, 8))
```

Out[ ]:

<AxesSubplot:xlabel='month'>