Dubinska analiza i otkrivanje znanja iz podataka

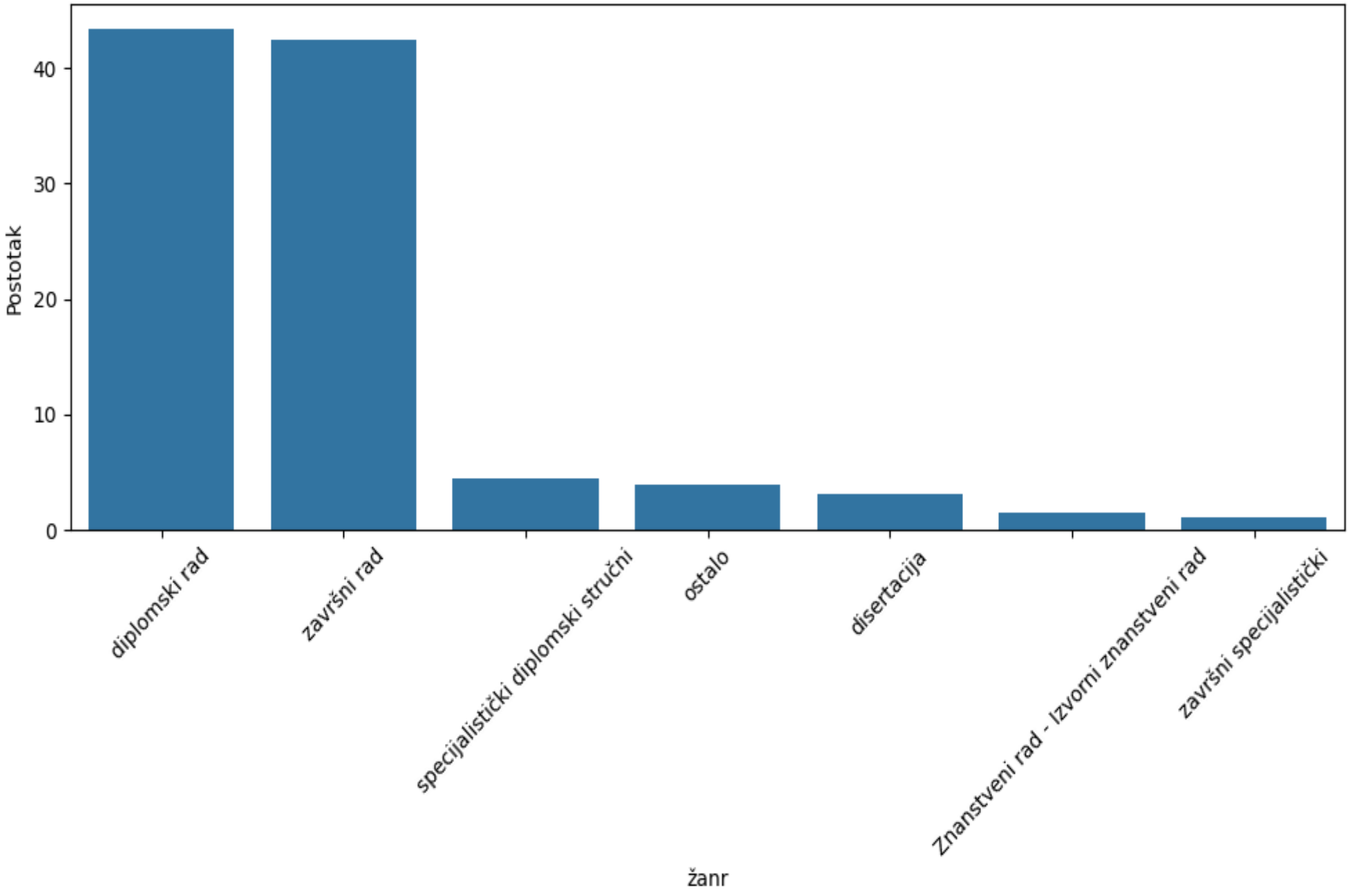
Korištenje Python paketa za statističku analizu i vizualizaciju podataka i analiza algoritma strojnog učenja i dubinske analize podataka

dr.sc. Damir Korenčić
Institut Ruđer Bošković
Laboratorij za strojno učenje i reprezentacije znanja

10. lipnja 2026.

Dabar repozitorij

- Centralni repozitorij akademskih radova
  - Diplomski radovi
  - Doktorati
  - ...
- Dabar skup podataka
  - Sakupljanje podataka (crawling)
  - Uređivanje
  - **Čišćenje**
- 265.790 radova (kraj 2025.)
- `https://dabar.srce.hr`


Postotak
0
10
20
30
40
diplomski rad
završni rad
specijalistički diplomski stručni
ostalo
disertacija
Znanstveni rad - Izvorni znanstveni rad
završni specijalistički
žanr

Matplotlib i Seaborn

- Matplotlib
  - Potrebna pozornost na detalje (mnogo koda)
  - Ne radi s Pandas DataFrame klasom
  - Zadane vrijednosti (defaults)
- Seaborn
  - API “iznad” Matplotlib-a
  - Jednostavne funkcije za standardne grafike
  - Integriran s DataFrame klasom
  - Razumne zadane vrijednosti (defaults)

Pandas DataFrame i Series

```python
def load_dframe_from_pkl(file_path: Path) -> DataFrame:
    with open(file_path, 'rb') as f:
        data = pickle.load(f)
    assert isinstance(data, DataFrame)
    return data
...
df = load_dframe_from_pkl('dabar.pkl') # DataFrame - tablica
df['genre_hrv'] # Series - indeksirani niz, kolumna u tablici
df[df['genre_hrv']=='diplomski rad'] # DataFrame - pod-tablica
```

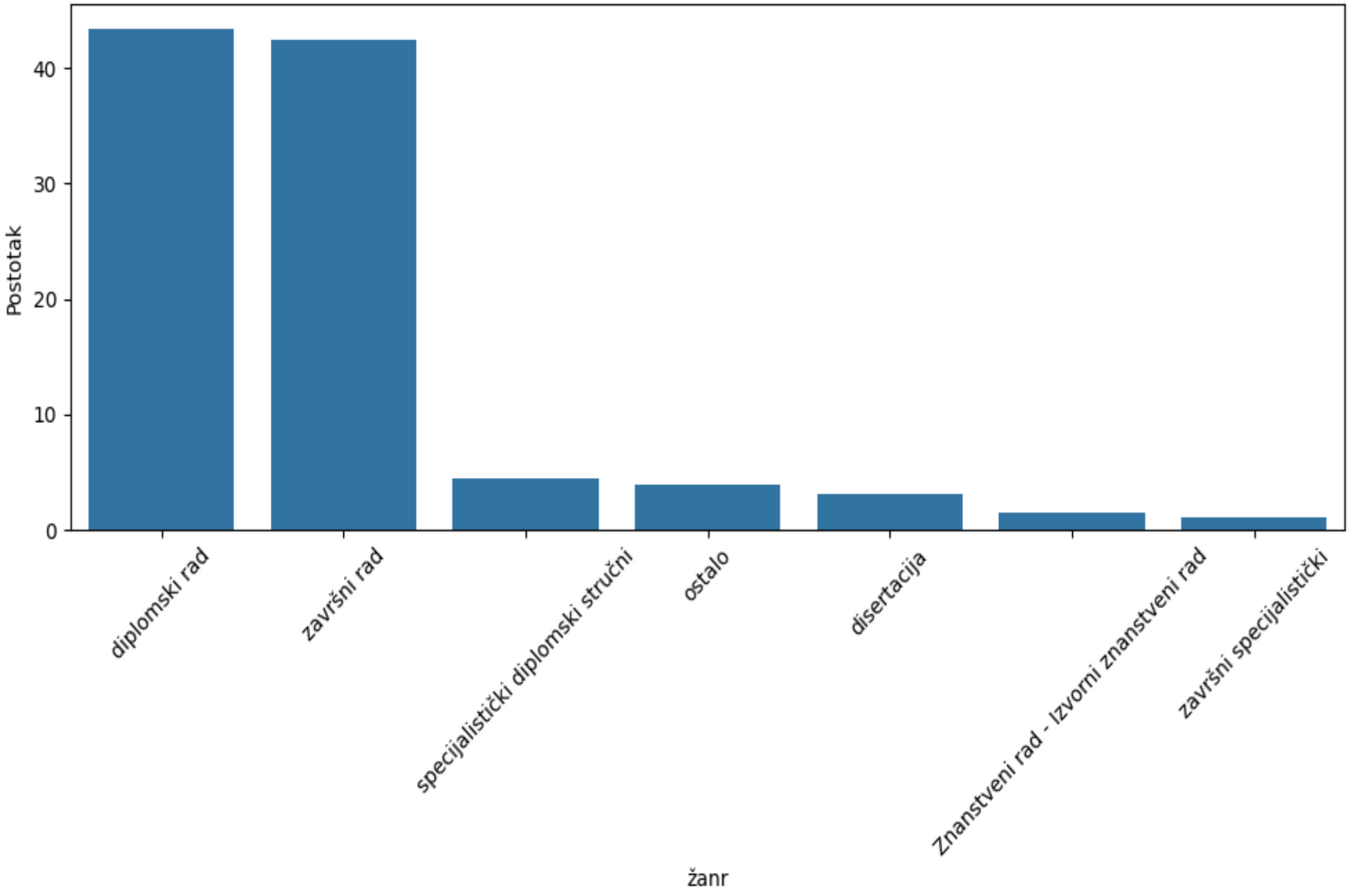
```
df['genre_hrv']
[(0, 'diplomski rad'), (1, 'završni rad'), …]

df['genre_hrv']=='diplomski rad'
[(0, True), (1, False), …]
```

Pandas DataFrame i Series

```
1 df['genre_hrv'].value_counts() # Series!
```

```
 diplomski rad, 115296
završni rad, 112667
specijalistički diplomski stručni, 11882
```


Postotak
0
10
20
30
40
diplomski rad
završni rad
specijalistički diplomski stručni
ostalo
disertacija
Znanstveni rad - Izvorni znanstveni rad
završni specijalistički
žanr

Matplotlib i Seaborn

```
df = load_dframe_from_pkl(DABAR_RAW_DATASET)
genres = df['genre_hrv']
```

Matplotlib:

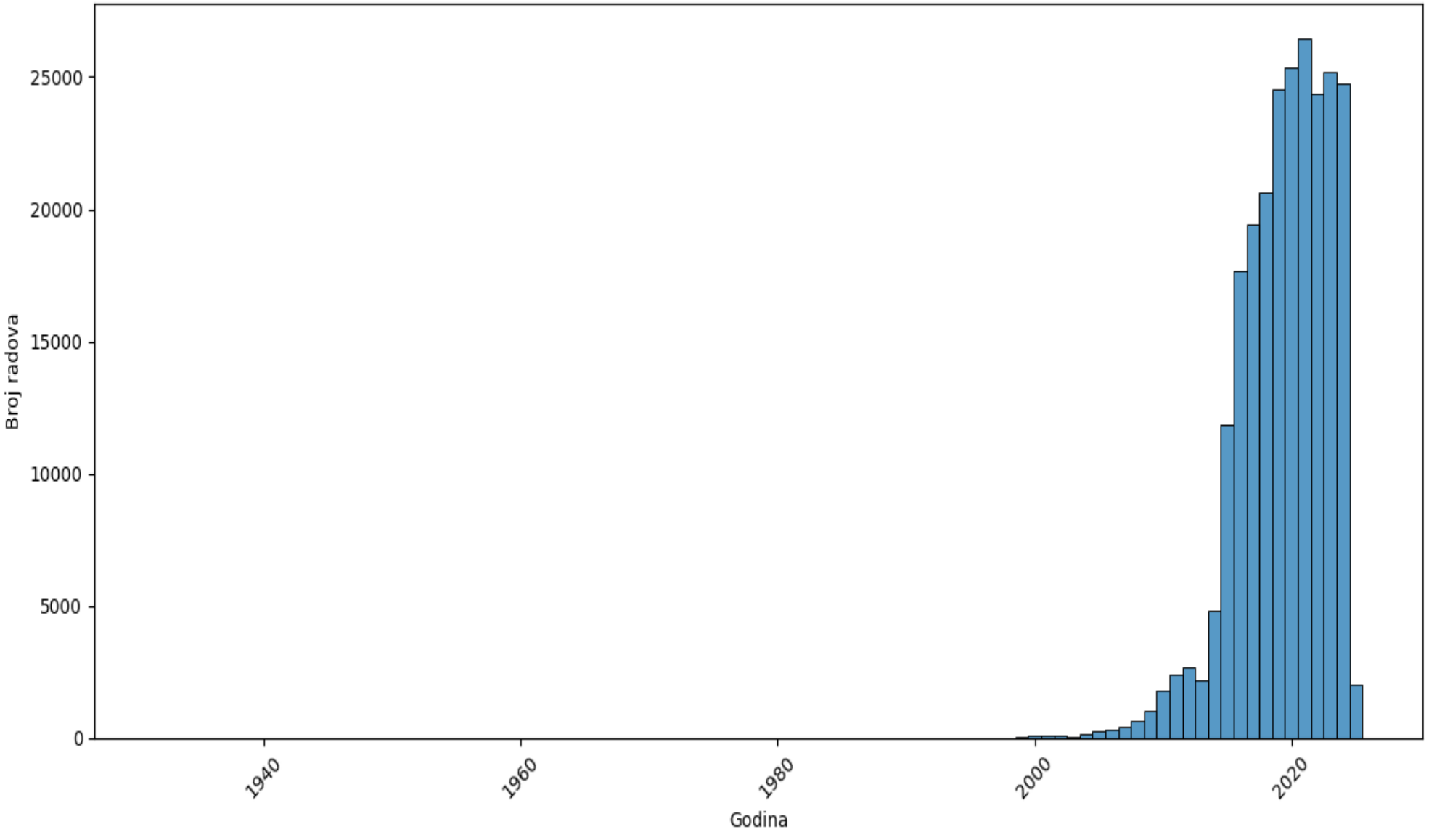
```
counts = genres.value_counts() # Series
heights = counts / counts.sum() * 100 # Series
x = range(len(counts))
plt.bar(x, heights.values)
```

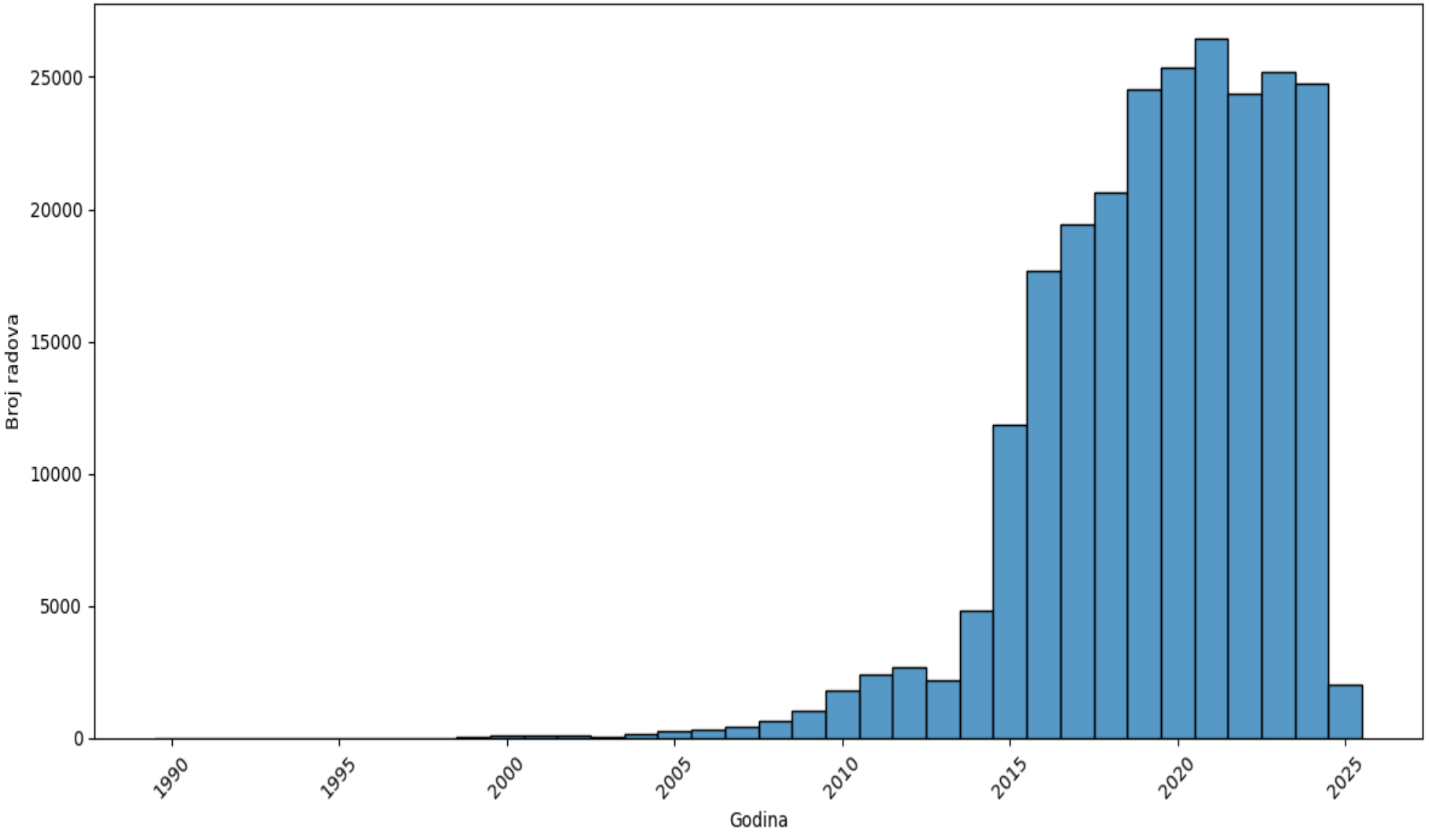
Seaborn:

```
sns.countplot(x=genres, order=genres.value_counts().index,
↪  stat='percent')
```

Seaborn+Matplotlib

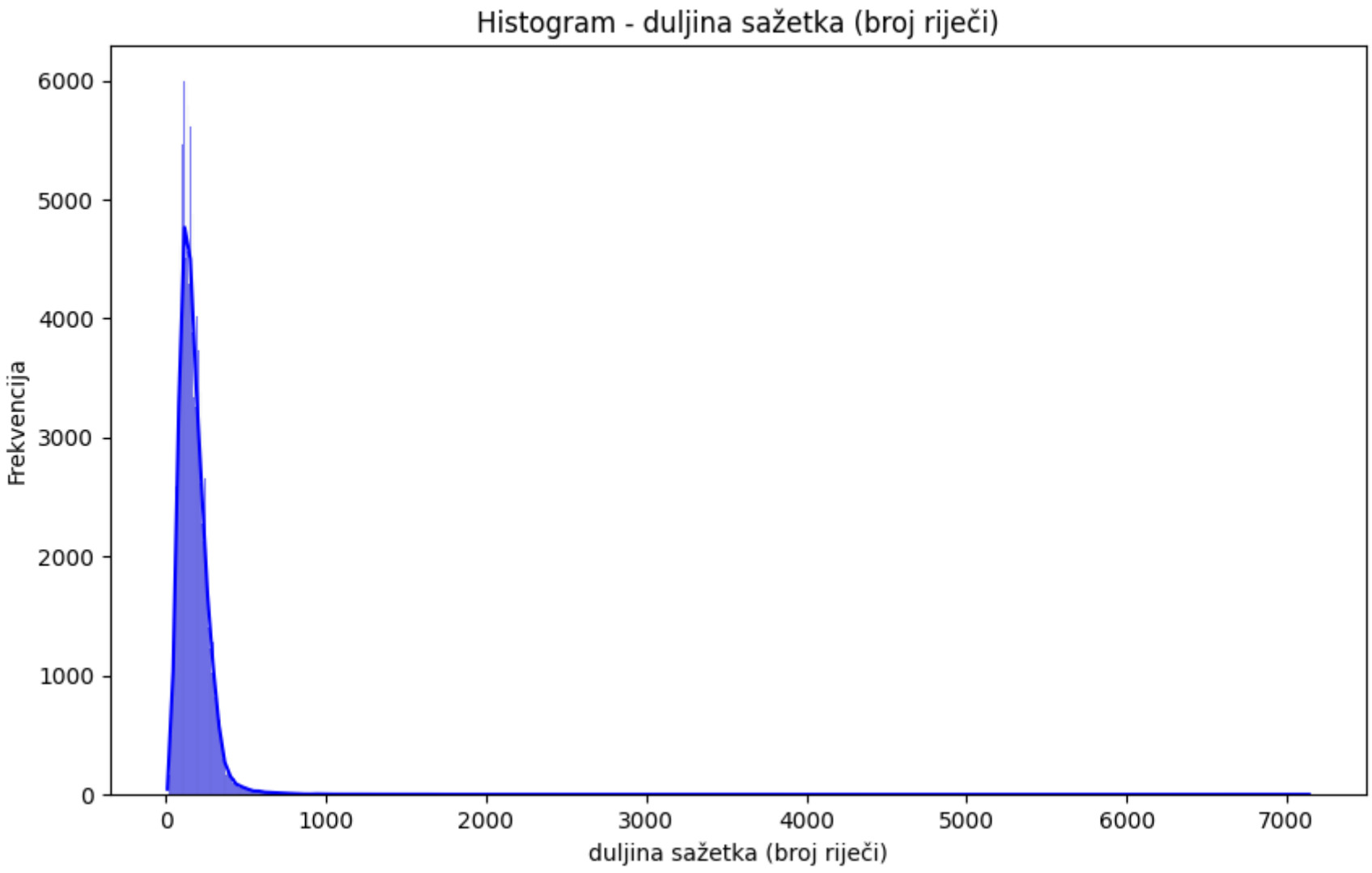
```
1 from matplotlib import pyplot as plt
2 plt.figure() # init. display
3 ...
4 sns.countplot(x=values, order=values.value_counts().index,
  ↪  stat='percent')
5 ...
6 plt.ylabel('postotak') # customize the plot
7 plt.xticks(rotation=45)
```

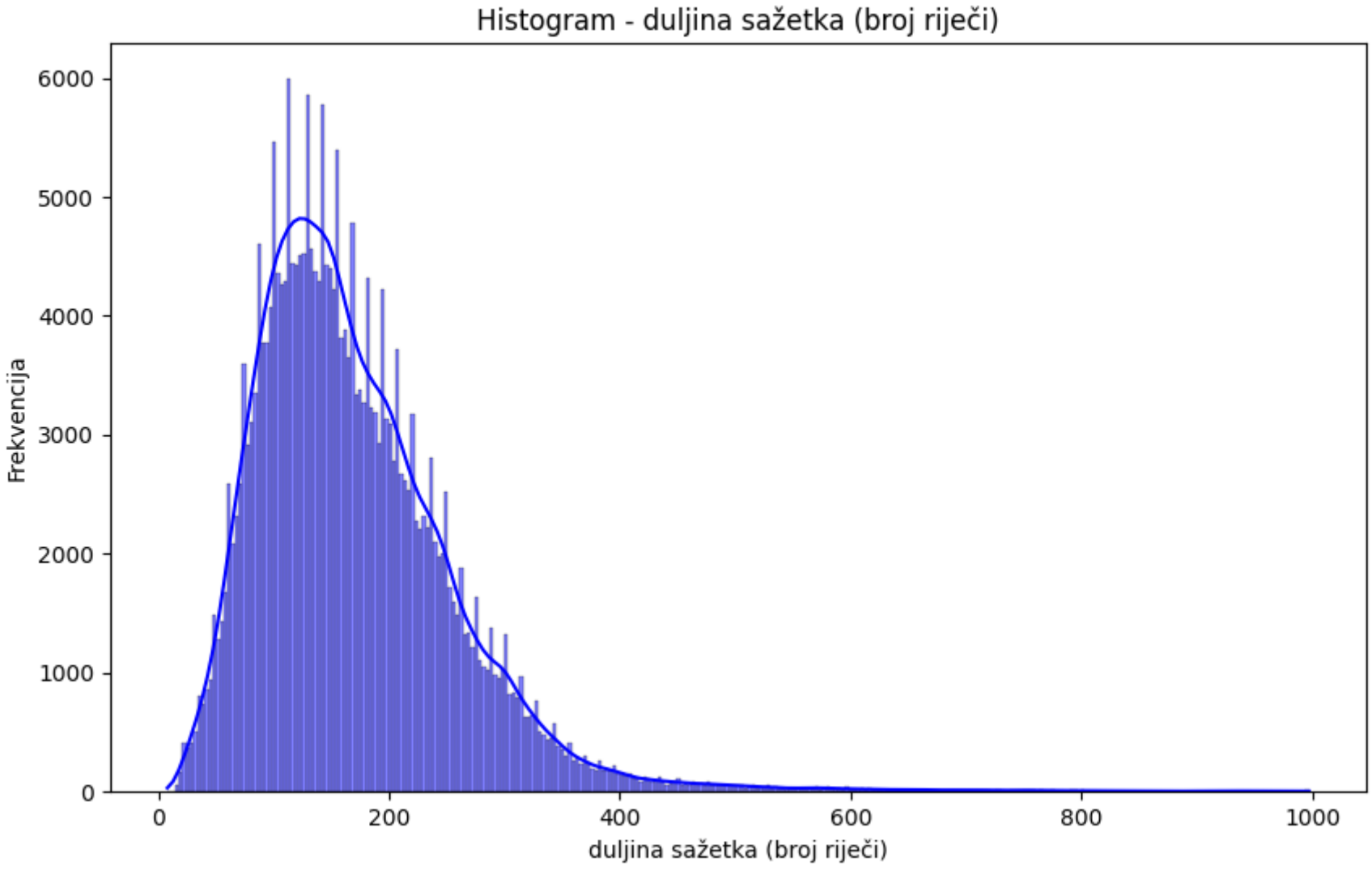

Broj radova
25000
20000
15000
10000
5000
0
1940
1960
1980
2000
2020
Godina


Broj radova
0
5000
10000
15000
20000
25000
1990
1995
2000
2005
2010
2015
2020
2025
Godina

Pandas i Seaborn

```
dates = pd.to_datetime(df['date_defended'], errors='coerce').dropna()
dates = dates[dates >= pd.to_datetime(cutoff_start)]
years = dates.dt.year
sns.histplot(x=years, discrete=True)
```


Histogram - duljina sažetka (broj riječi)
Frekvencija
duljina sažetka (broj riječi)
0
1000
2000
3000
4000
5000
6000
7000


Histogram - duljina sažetka (broj riječi)
Frekvencija
duljina sažetka (broj riječi)
0
1000
2000
3000
4000
5000
6000
0
200
400
600
800
1000

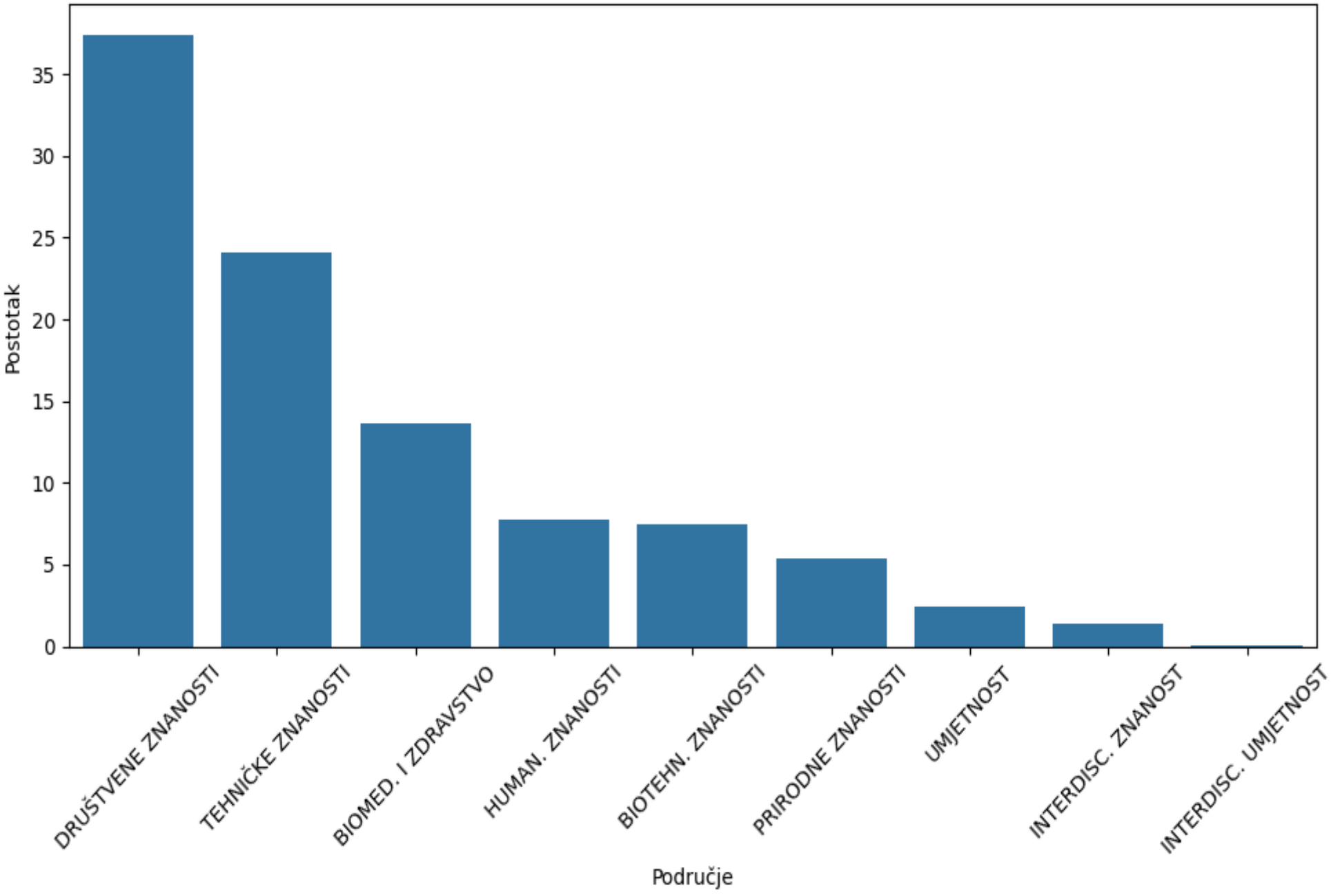
Pandas i Seaborn

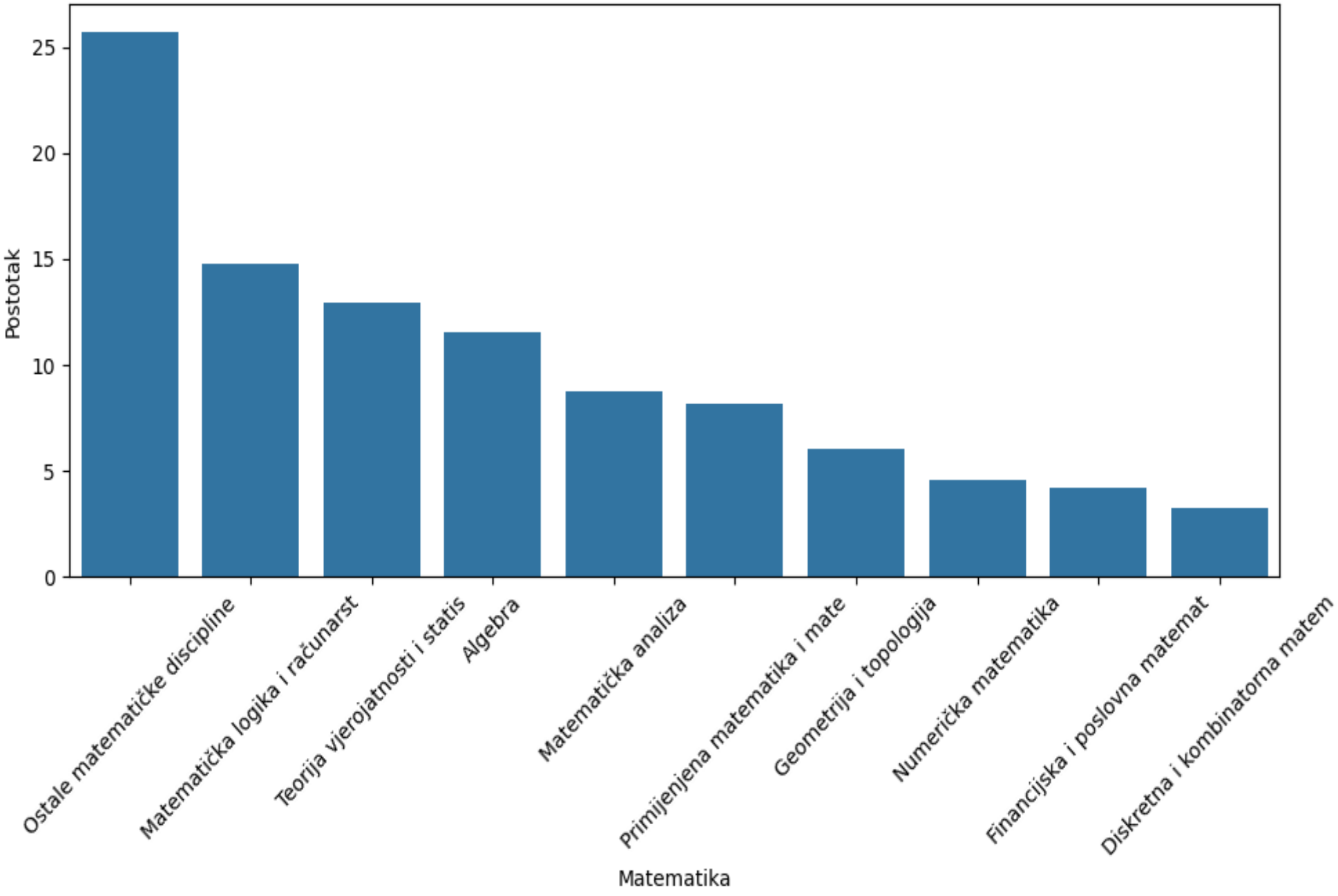
```
summary = numbers.describe() # Series
print(summary) # Series!
...
sns.histplot(numbers, bins="auto", color='blue', kde=True)
```

```
count 262287.000000
mean 171.845753
std 104.487170
min 8.000000
25% 110.000000
50% 154.000000
75% 214.000000
max 7141.000000
```

Znanstveno-umjetnička taksonomija

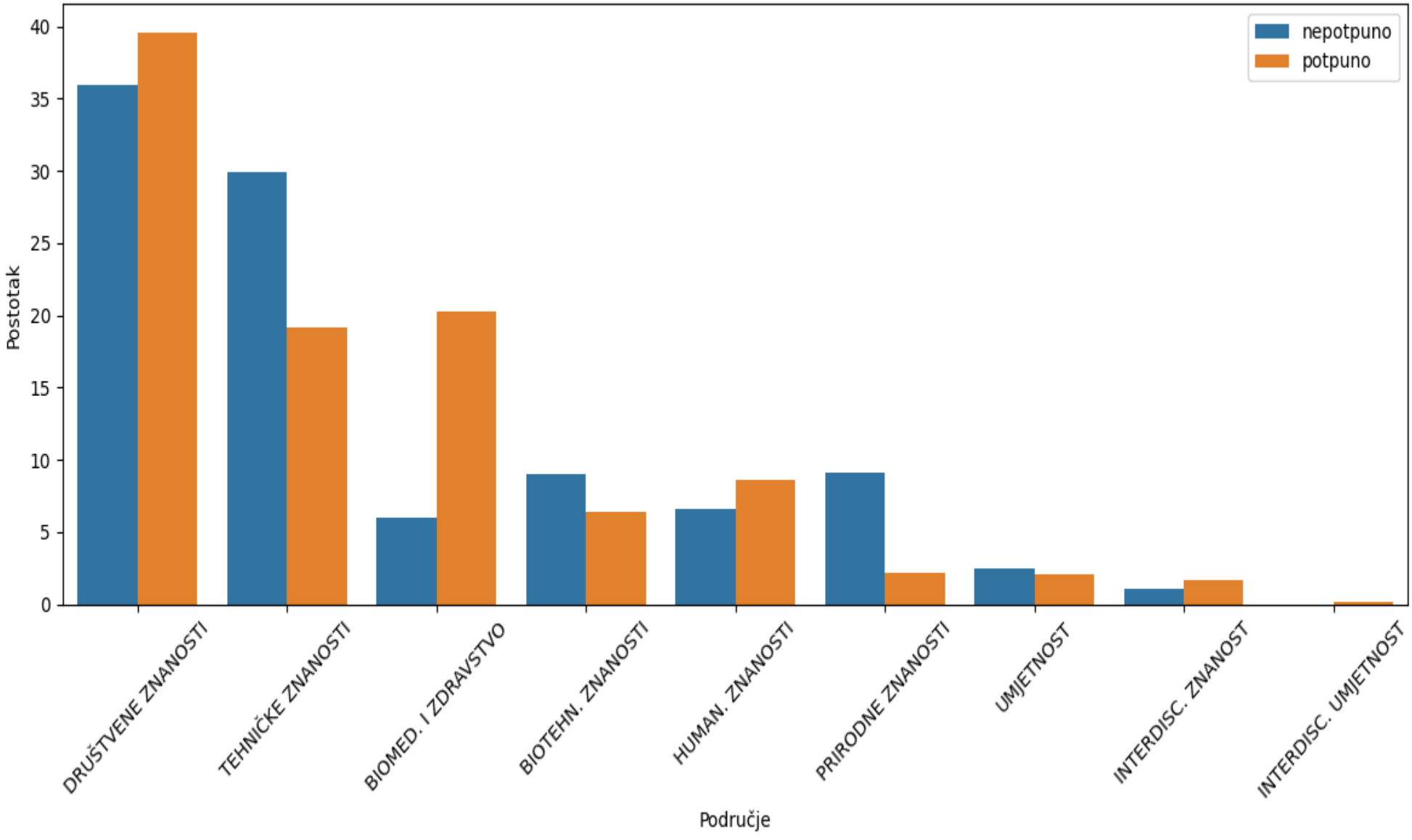
- Područje $\rightarrow$ Polje $\rightarrow$ Grana
- Prirodne znanosti $\rightarrow$ Matematika $\rightarrow$ Algebra
- Nacionalno vijeće za znanost: “Pravilnik o znanstvenim i umjetničkim područjima, poljima, i granama” (2009)


Postotak
0
5
10
15
20
25
30
35
DRUŠTVENE ZNANOSTI
TEHNIČKE ZNANOSTI
BIOMED. I ZDRAVSTVO
HUMAN. ZNANOSTI
BIOTEHN. ZNANOSTI
PRIRODNE ZNANOSTI
UMJETNOST
INTERDISC. ZNANOST
INTERDISC. UMJETNOST
Područje


Postotak
0
5
10
15
20
25
Ostale matematičke discipline
Matematička logika i računarst
Teorija vjerojatnosti i statis
Algebra
Matematička analiza
Primijenjena matematika i mate
Geometrija i topologija
Numerička matematika
Financijska i poslovna matemat
Diskretna i kombinatorna matem
Matematika

Dabar – taksonomija

- Potpuni unos:
  Prirodne znanosti $\rightarrow$ Matematika $\rightarrow$ Algebra
- Nepotpuni unos:
  Prirodne znanosti $\rightarrow$ Matematika


nepotpuno
potpuno
Postotak
0
5
10
15
20
25
30
35
40
DRUŠTVENE ZNANOSTI
TEHNIČKE ZNANOSTI
BIOMED. I ZDRAVSTVO
BIOTEHN. ZNANOSTI
HUMAN. ZNANOSTI
PRIRODNE ZNANOSTI
UMJETNOST
INTERDISC. ZNANOST
INTERDISC. UMJETNOST
Područje

Pregled tema u skupu podataka

- ključne riječi: 'dramsko pismo', 'filmsko pismo'
- normalizacija: 'dramskoPismo', 'filmskoPismo'

```
1  from wordcloud import WordCloud
2  ...
3  for subjects in df[keyword_column]:
4     subjects_list.extend([normalize(subject) for subject in subjects
      ↪  if subject])
5  ...
6  wordcloud = WordCloud(width=3200, height=1600,
   ↪  background_color='white', normalize_plurals=False)
7  wordcloud.generate(' '.join(subjects_list)) # jedan dokument
8  plt.figure(figsize=(20, 10))
9  plt.imshow(wordcloud, interpolation='bilinear')
```

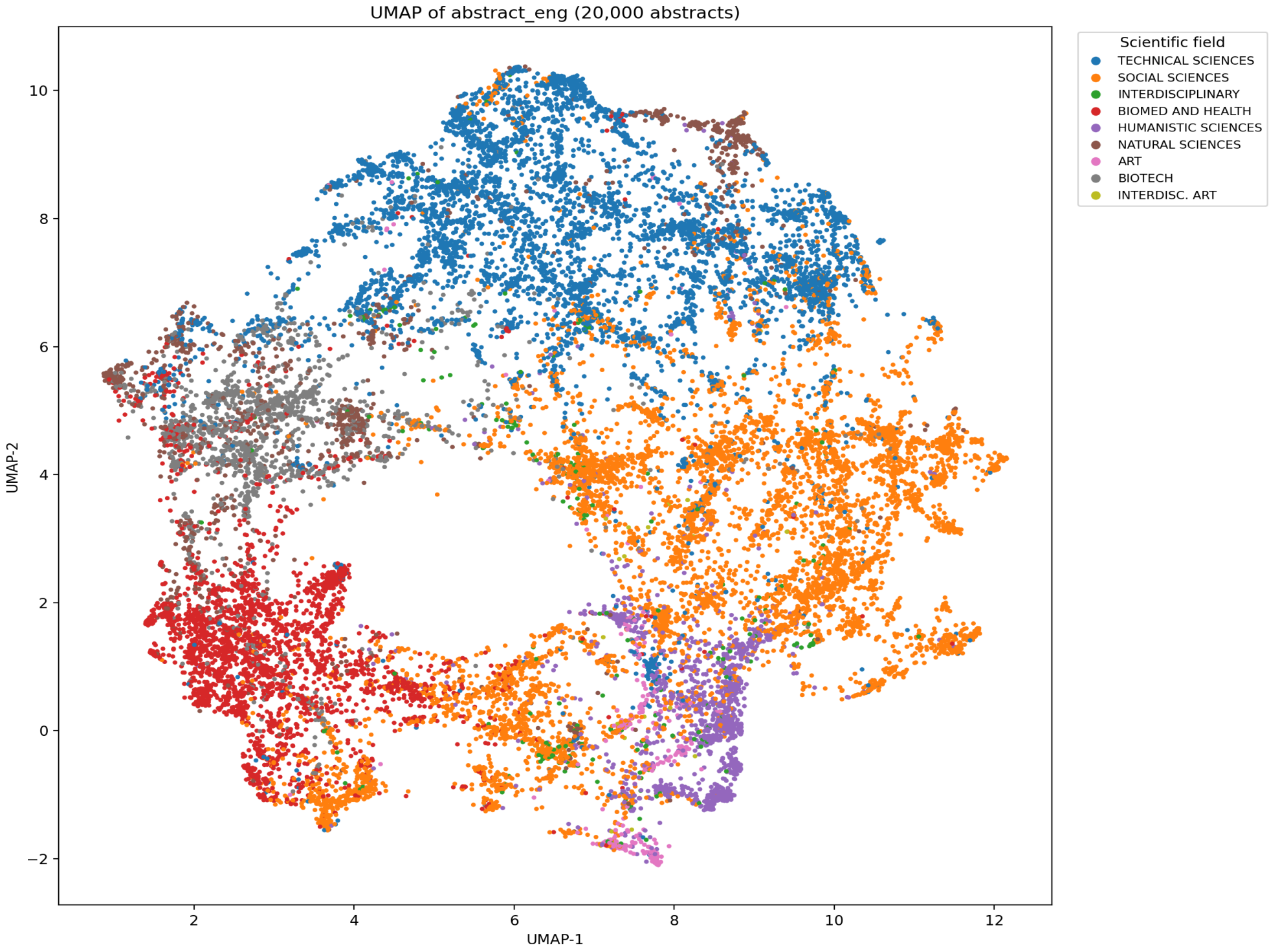

hrvatska
ključneRiječi
turizam
analiza
komunikacija
republikaHrvatska
razvoj
djeca
sigurnost
kvalitetaŽivota
organizacija
kvaliteta
edukacija
motivacija
prevencija
obrazovanje
istraživanje
održiviRazvoj
digitalizacija
covid
mediji
društveneMreže
tehnologija
europskaUnija
marketing
medicinskaSestra
automatizacija
model
umjetnaInteligencija
poduzeće
potrošači
projekt
globalizacija
poduzetništvo
promet
energija
inovacije
integracija
identitet
zaštita
arhitektura
oglašavanje
simulacija
tržište
računovodstvo
stavovi
strategija
troškovi
matlab
zaštitaOkoliša
ponašanjePotrošača
zdravstvenaNjega
trgovina
videoigre
ruralniTurizam
klimatskePromjene
mentalnoZdravlje
digitalniMarketing
ultrazvuk
dizajn
električnaEnergija
kompetencije
python
proizvodnja
dubrovnik
prehrana
hrvatskiJezik
kultura
fizioterapija
bazaPodataka
internet
ljudskiPotencijali
aplikacija
metodaKonačnihElemenata
kulturnaBaština
internetStvari
optimizacija
upravljanje
energetskaUčinkovitost
proces
liječenje
adolescenti
usporedba
poljoprivreda
konkurentnost
turističkaDestinacija
osiguranje
učinkovitost
hotelijerstvo
trendovi
webAplikacija
konstrukcija
održavanje
strojnoUčenje
percepcija
temperatura
boja
rizik
modeliranje
nogomet
roditelji
terapija
zaposlenici
metode
sustav
korozija
voda
umjetnost
arduino
gospodarstvo
programiranje
promocija
obitelj
logistika
mladi
bol
inkluzija
sport
klasifikacija
diplomskiIspit
koncert
stres
društvenoOdgovornoPoslovanje
financiranje
fotografija
glazba
etika
utjecaj
dijete
zdravlje
android
gradZagreb
održivost
rehabilitacija
brend
hotel
učenici
transport
kreativnost
mjerenje
node js
algoritam
tradicija
sestrinstvo
politika
anketa
sestrinskaSkrb
zavarivanje
eu
d
prodaja
zagreb
igra
odnosiSJavnošću
emocije
dijagnostika
menadžment
društvo
pandemija
povijest
zadovoljstvo
pretilost
poslovanje
kurikulum
žene
Stoljeće
ekologija
kulturniTurizam
okoliš
ambalaža
c
malaISrednjaPoduzeća
studenti
html css
feminizam
proračun
kriza
pandemijaCovid
planiranje
tjelesnaAktivnost
obnovljiviIzvoriEnergije
financijskiIzvještaji
senzori
diskriminacija
mobilnaAplikacija
ključneRiječiUniouUredniku
medicinaInformacijskiSustav
konkurentskaPrednost
web aplikacija
znanje
ljudskaPrava
suradnja
istra
film
implementacija
medicinskaSestra tehničar
digitalnaTransformacija
trudnoća
zakon


financijskiIzvještaji
ruralniTurizam
marketing
studenti
društvenoOdgovornoPoslovanje
percepcija
emocije
diskriminacija
hotelijerstvo
algoritam
menadžment
europskaUnija
obitelj
medicina medicina
troškovi
konstrukcija
fotografija
voda
igra
eu
trudnoća
transport
temperatura
pretilost
programiranje
proizvodnja
društveneMreže
tehnologija
promocija
zdravlje
etika
poljoprivreda
arhitektura
računovodstvo
rehabilitacija
model
gradZagreb
tradicija
komunikacija
kriza
terapija
proračun
stavovi
sustav
inovacije
logistika
sport
sestrinstvo
oglašavanje
liječenje
nogomet
umjetnost
brend
hotel
ekologija
dijagnostika
odnosiSJavnošću
arduino
internet
simulacija
android
identitet
električnaEnergija
održivost
učenici
dijete
politika
python
poslovanje
covid
strategija
razvoj
djeca
motivacija
zaštita
strojnoUčenje
znanje
adolescenti
feminizam
fizioterapija
održiviRazvoj
klimatskePromjene
web aplikacija
energija
rizik
webAplikacija
pandemijaCovid
utjecaj
republikaHrvatska
učenje
senzori
pandemija
anketa
dubrovnik
aplikacija
turističkaDestinacija
bazaPodataka
videoigre
implementacija
korozija
kreativnost
diplomskiIspit
koncert
konkurentnost
povijest
kulturnaBaština
mediji
kvalitetaŽivota
sigurnost
okoliš
promet
zakon
prevencija
istraživanje
planiranje
poduzetništvo
umjetnaInteligencija
kulturniTurizam
trendovi
kompetencije
upravljanje
digitalniMarketing
prehrana
poduzeće
održavanje
žene
Stoljeće
organizacija
trgovina
mladi
roditelji
kultura
projekt
zdravstvenaNjega
potrošači
mjerenje
društveniMediji
modeliranje
zadovoljstvo
glazba
konkurentskaPrednost
obrazovanje
dizajn
metode
stres
bol
društvo
hrvatskiJezik
tržište
edukacija
tjelesnaAktivnost
globalizacijaoptimizacija
automatizacija
medicinskaSestra
metodaKonačnihElemenata
kvaliteta
ponašanjePotrošača
digitalizacija
film
mobilnaAplikacija
ultrazvuk
osiguranje
informacijskiSustav
suradnja
prodaja
node js
html css
učinkovitost
gospodarstvo
zavarivanje
boja
proces
kurikulum
nasilje
zaposlenici
matlab
socijalniRad
ključneRiječiUnioUrednik
politika
integracija
internetStvari
usporedba
inkluzija
malaISrednjaPoduzeća
financiranje
klasifikacija
sestrinskaSkrb
zaštitaOkoliša
ljudskiPotencijali
digitalnaTransformacija
medicinskaSestra
tehničar
mentalnoZdravlje
obnovljiviIzvoriEnergije

UMAP 2D vizualizacija

- "Semantic Embedding"
  - Sažeci tekstova $\rightarrow$ visokodim. vektor
  - `sentence-transformers/all-mpnet-base-v2`
- UMAP algoritam
  - Konstrukcija grafa u visokodim. prostoru
    - broj susjeda
    - mjera udaljenosti
  - Aproksimacija grafa u 2D prostoru

```
reducer = umap.UMAP(n_components=2, n_neighbors=30,
↪  min_dist=min_dist, metric='cosine', random_state=random_state)
return reducer.fit_transform(embeddings)
```


UMAP of abstract_eng (20,000 abstracts)
Scientific field
TECHNICAL SCIENCES
SOCIAL SCIENCES
INTERDISCIPLINARY
BIOMED AND HEALTH
HUMANISTIC SCIENCES
NATURAL SCIENCES
ART
BIOTECH
INTERDISC. ART
UMAP-1
UMAP-2
10
8
6
4
2
0
−2
2
4
6
8
10
12

UMAP 2D vizualizacija

- **Alat za istraživanje**
- Alat za zaključivanje (?!)
  - Validacija rezultata
  - U kombinaciji s drugim argumentima
- Varijacija parametara (i random seed-a)
  - Uočavanje novih obrazaca
  - Robustnost zaključaka

Alati za analizu podataka

- Visual Studio Code
  - `sudo snap install --classic code`
- Jupyter ekstenzija

  `"jupyter.notebookFileRoot": "$workspaceFolder"`
- Data Wrangler ekstenzija
- Jupyter integriran sa GitHub Copilot-om

AI za analizu podataka

- Jednostavan kod (u pravilu)
- Lakša provjera rezultata
- ...
- Katalog alata i tehnika (čovjek)
  -> Implementacija (AI)

The End...

Hvala na pažnji!