Wiadomości z Polski: próba text miningu

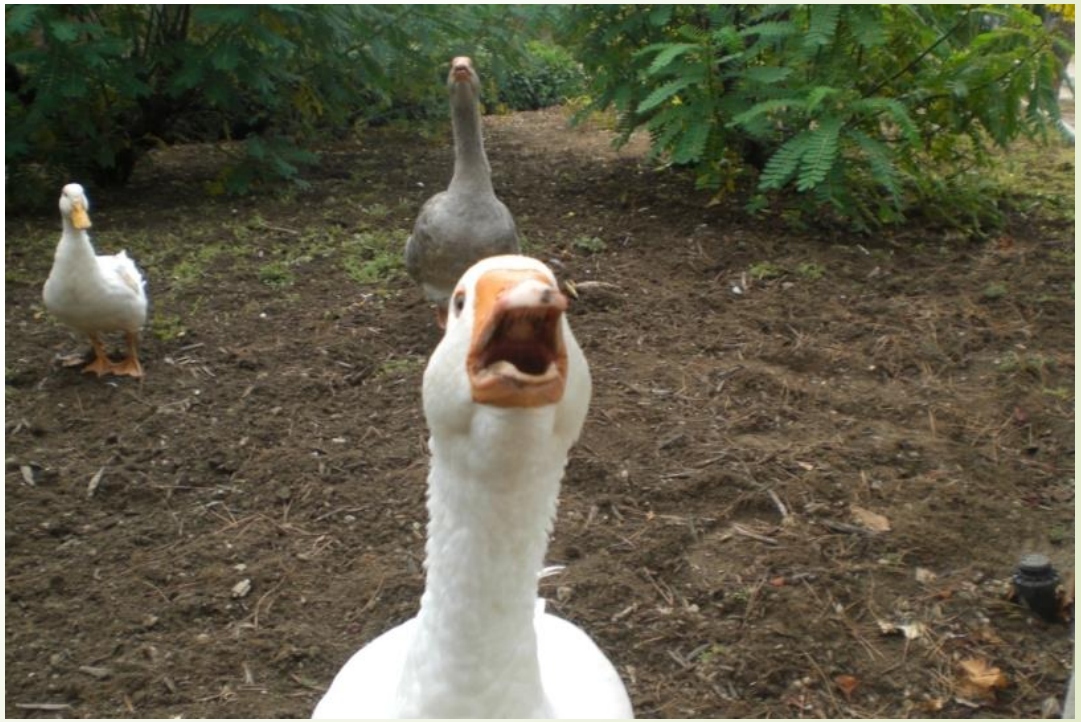
Piotr Orłowski

Agenda:

- Cel projektu
- Źródła i pozyskiwanie danych
- Zastosowane technologie
- Wstępna analiza danych
- Eksperyment: NER
- Eksperyment: klasyfikacja
- Podsumowanie i wnioski

Cel projektu

- Wypróbujemy NLP z językiem polskim



źródło obrazka: https://blog.dobryslownik.pl/polacy-nie-gesi-o-co-naprawde-chodzi-w-tym-slynnym-zdaniu/

Źródła i pozyskiwanie danych

- Metoda: web scrapping
- Źródła:
  - Bankier.pl
  - NaszDziennik.pl
  - onet.pl
  - PcWorld.pl
  - Plotek.pl
  - PolskieRadio24.pl
  - tvn24.pl
- Język danych: Polski

Zastosowane technologie

- Jupyter Lab
- Pandas
- Spacy
- Scikit Learn

Spacy – polski duży pipeline

pl_core_news_lg

Polish pipeline optimized for CPU. Components: tok2vec, morphologizer, tagger, parser, senter, ner, attribute_ruler, lemmatizer.

LANGUAGE	PL Polish
TYPE	CORE Vocabulary, syntax, entities, vectors
GENRE	NEWS written text (news, media)
SIZE	LG 583 MB
COMPONENTS	`tok2vec`, `morphologizer`, `parser`, `tagger`, `senter`, `attribute_ruler`, `lemmatizer`, `ner`
PIPELINE	`tok2vec`, `morphologizer`, `parser`, `tagger`, `attribute_ruler`, `lemmatizer`, `ner`
VECTORS	500k keys, 500k unique vectors (300 dimensions)

https://spacy.io/models/pl (od 04.11.2021)

Wstępna analiza danych

```
full_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2400 entries, 0 to 2399
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   title           2400 non-null   object
 1   title_int       2400 non-null   object
 2   lead_text       2400 non-null   object
 3   lead_text_int   2091 non-null   object
 4   link            2400 non-null   object
 5   text            2339 non-null   object
 6   when_published  2400 non-null   datetime64[ns]
 7   author          2008 non-null   object
 8   source          1432 non-null   object
 9   portal          2400 non-null   object
dtypes: datetime64[ns](1), object(9)
memory usage: 187.6+ KB
```

```
full_df.when_published.value_counts()
```

```
2022-01-19 19:58:00    4
2022-01-18 14:28:00    4
2022-01-19 12:03:00    4
2022-01-19 10:00:00    4
2022-01-18 11:35:00    4
                      ..
2020-09-25 11:33:00    1
2020-09-28 12:19:00    1
2020-10-02 11:20:00    1
2020-10-04 20:42:00    1
2022-01-14 20:04:00    1
Name: when_published, Length: 2172, dtype: int64
```

```
full_df.portal.value_counts()
```

```
bankier           300
naszdziennik      300
niebezpiecznik    300
onet              300
pcworld           300
plotek            300
pr24              300
tvn24             300
```

```
full_df.title.value_counts()
```

```
Jak kształtuje się liczba zakażonych                                  37
Prognoza pogody                                                        9
We wtorek w „Naszym Dzienniku”                                         6
W środę w „Naszym Dzienniku”                                           6
W piątek w „Naszym Dzienniku”                                          6
                                                                      ..
Numery ksiąg wieczystych w Geoportalu: Kara 100 tys. zł dla GGK, ale to nie koniec    1
```

```
counter=Counter(corpus)
most=counter.most_common()
pprint(most[0:10])
```

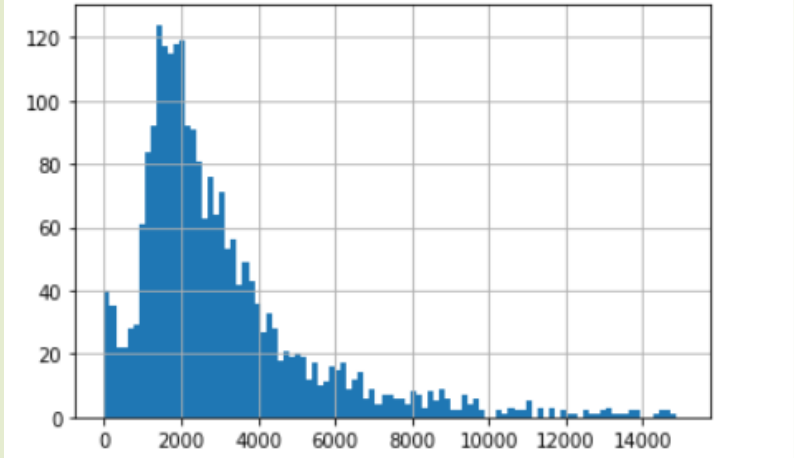
```
[('-', 52443),
 ('w', 41469),
 ('na', 26058),
 ('z', 25467),
 ('przyrost', 24071),
 ('i', 22434),
 ('do', 18291),
 ('się', 16632),
 ('że', 14087),
 ('nie', 12469)]
```

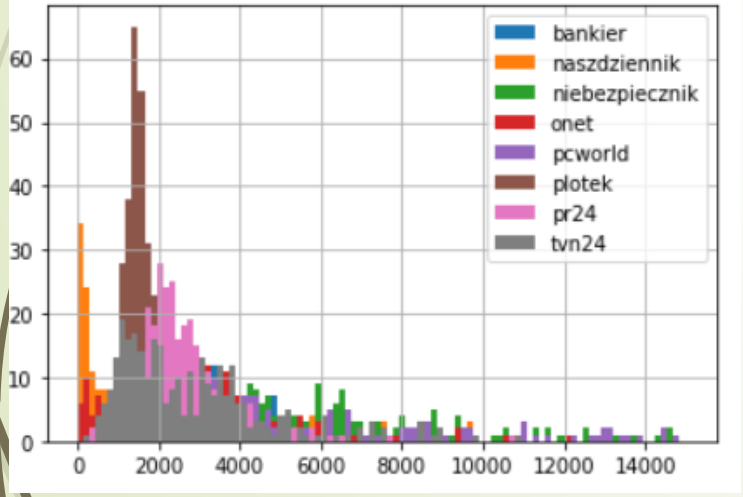
```
len(counter.keys())
```

```
149992
```

Wybrane wykresy

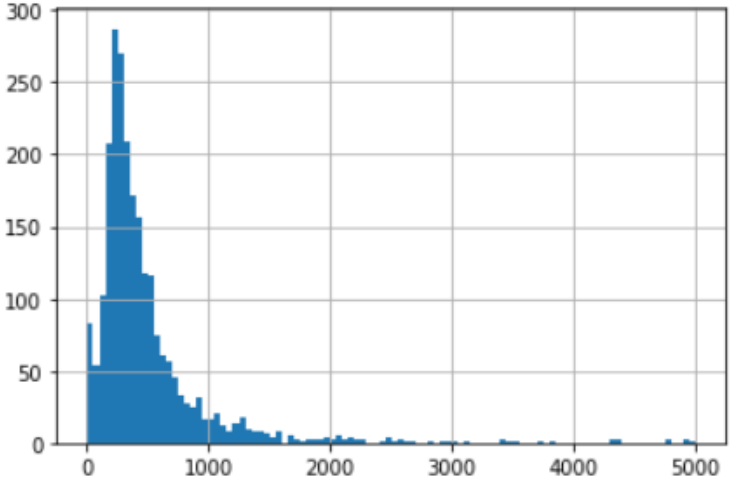
Długość artykułów na wszystkich portalach - histogram



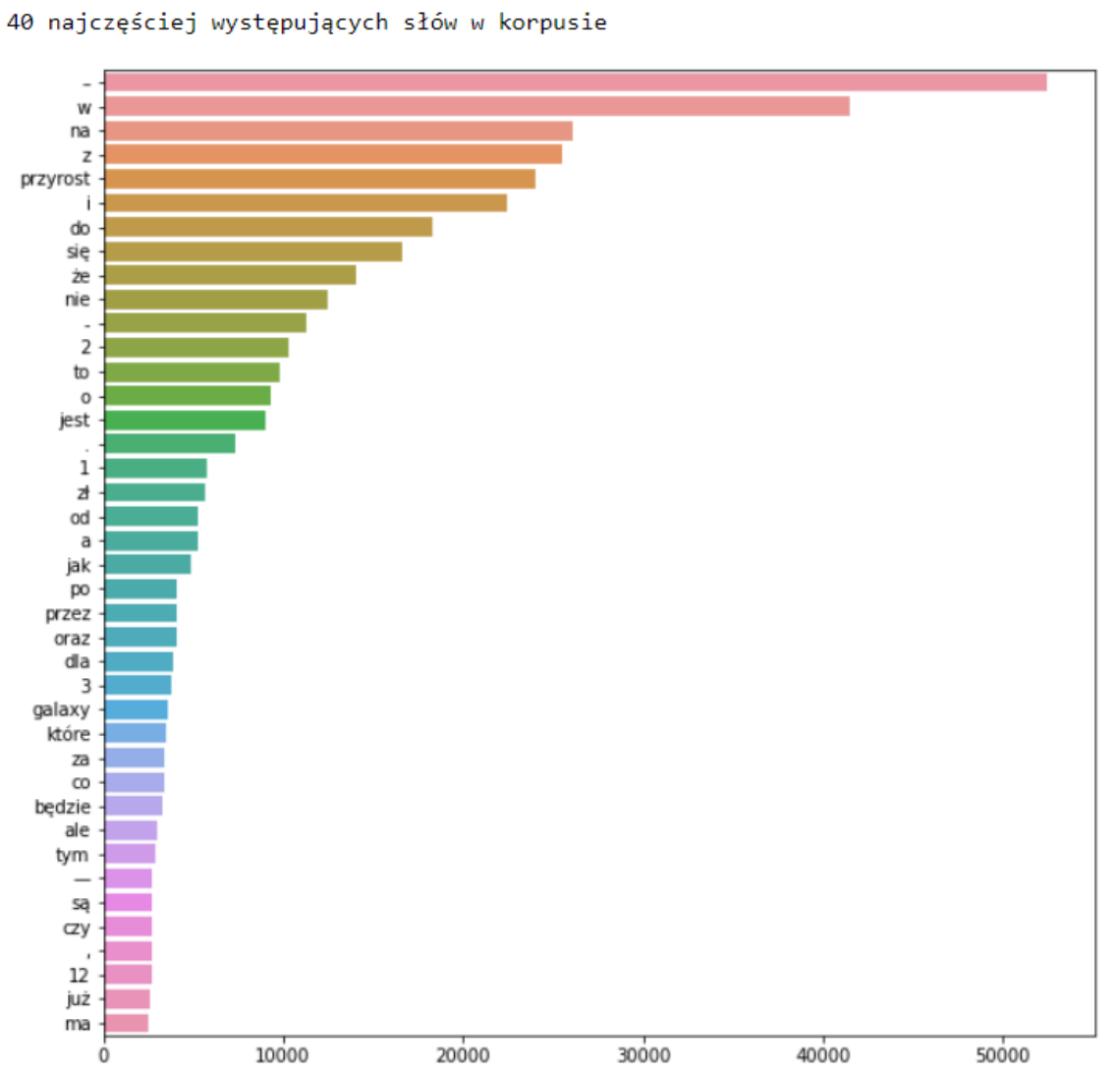


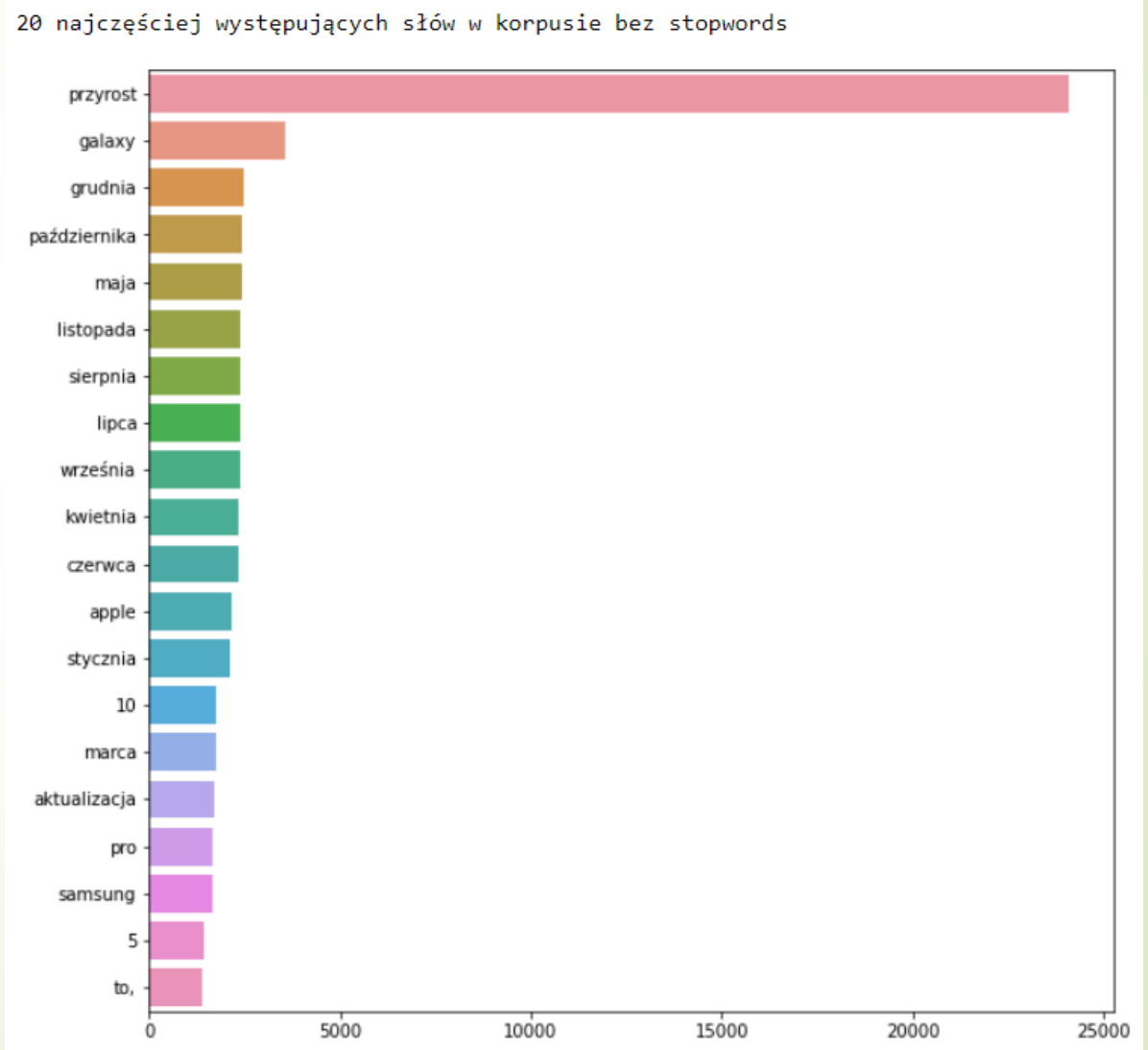
histogram popularności słów

```
count     2400.000000
mean       646.964583
std       1394.993872
min          9.000000
25%        226.750000
50%        348.000000
75%        567.000000
max      26725.000000
Name: lead_and_text, dtype: float64
```



Korpus


40 najczęściej występujących słów w korpusie
-
w
na
z
przyrost
i
do
się
że
nie
.
2
to
o
jest
.
1
zł
od
a
jak
po
przez
oraz
dla
3
galaxy
które
za
co
będzie
ale
tym
—
są
czy
,
12
już
ma
0
10000
20000
30000
40000
50000


20 najczęściej występujących słów w korpusie bez stopwords
przyrost
galaxy
grudnia
października
maja
listopada
sierpnia
lipca
września
kwietnia
czerwca
apple
stycznia
10
marca
aktualizacja
pro
samsung
5
to,
0
5000
10000
15000
20000
25000

Problemy…

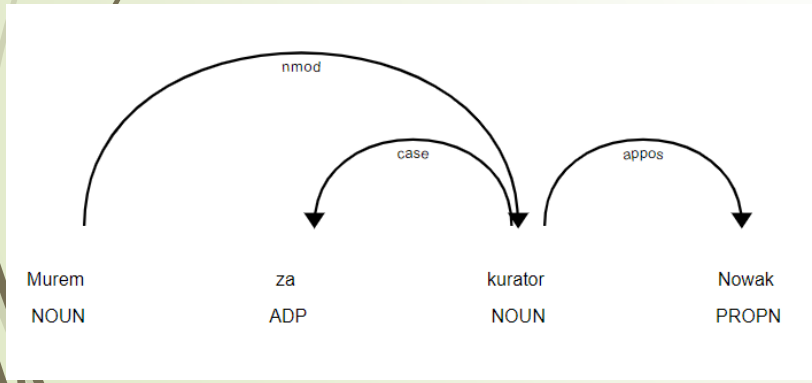
- Spacy – jest model języka polskiego ale… wiele rzeczy nie działa lub nie powstało (klasyfikacja, analiza sentymentu)
- Niewiele publikacji o NLP w języku Polskim, niewiele zespołów badawczych
- Niewiele dostępnych zbiorów uczących do użycia

Ale jednak dużo rzeczy działa ☺

Eksperyment: NER

Najpopularniejsze byty nazwane występujące w tytułach artykułów portal: full, korpus: 1272

```
[(('2022', 'date'), 121),
 (('polska', 'placeName'), 119),
 (('01.', 'date'), 114),
 (('polski', 'placeName'), 79),
 (('19.', 'date'), 44),
 (('pis', 'orgName'), 42),
 (('18.', 'date'), 34),
 (('„naszym dziennik”', 'geogName'), 33),
 (('rada medyczny', 'orgName'), 32),
 (('tusk', 'persName'), 22)]
```



Najpopularniejsze NER na bankierze (bez lematyzacji):

```
[(('Polsce', 'placeName'), 10),
 (('Polski', 'placeName'), 9),
 (('PGNiG', 'orgName'), 7),
 (('Polskiego', 'placeName'), 6),
 (('NIK', 'orgName'), 6),
 (('PGE', 'orgName'), 5),
 (('Ukrainy', 'placeName'), 5),
 (('Tusk', 'persName'), 5),
 (('Chin', 'placeName'), 4),
 (('Polskim', 'placeName'), 4),
 (('ZUS', 'orgName'), 4),
 (('NBP', 'orgName'), 4),
 (('Polska', 'placeName'), 4),
 (('KE', 'orgName'), 4),
 (('UE', 'orgName'), 4),
 (('Polacy', 'placeName'), 4),
 (('Polaków', 'placeName'), 4),
 (('USA', 'placeName'), 3),
 (('PiS', 'orgName'), 3),
```

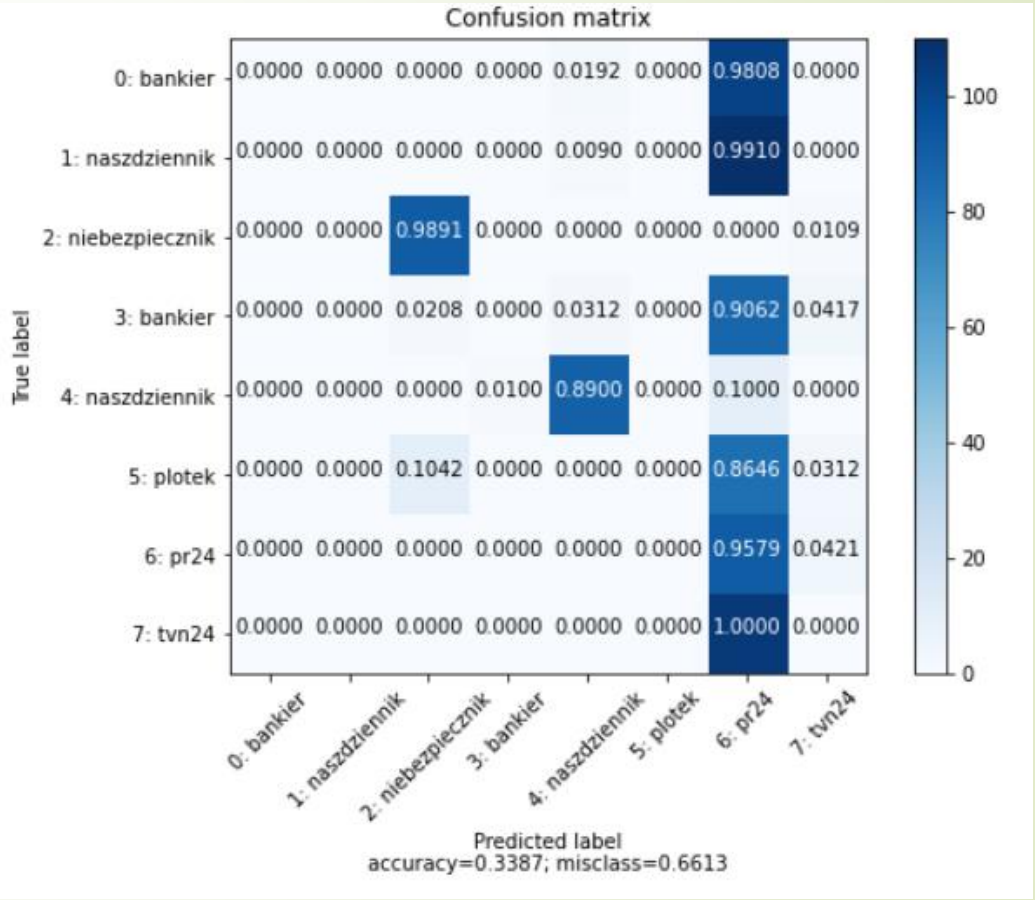
index: 373
title: Murem za kurator Nowak

Akcja Katolicka Archidiecezji Częstochowskiej orgName staje w obronie małopolskiej placeName kurator oświaty. Przystąpili Państwo do akcji „Murem za kurator Nowak”. persName – Nie ma zgody środowisk katolickich na odwołanie pani

Eksperyment: klasyfikacja

- Adaptive Boost + Count Vectorizer (Scikit-Learn)
- Dość słaba klasyfikacja

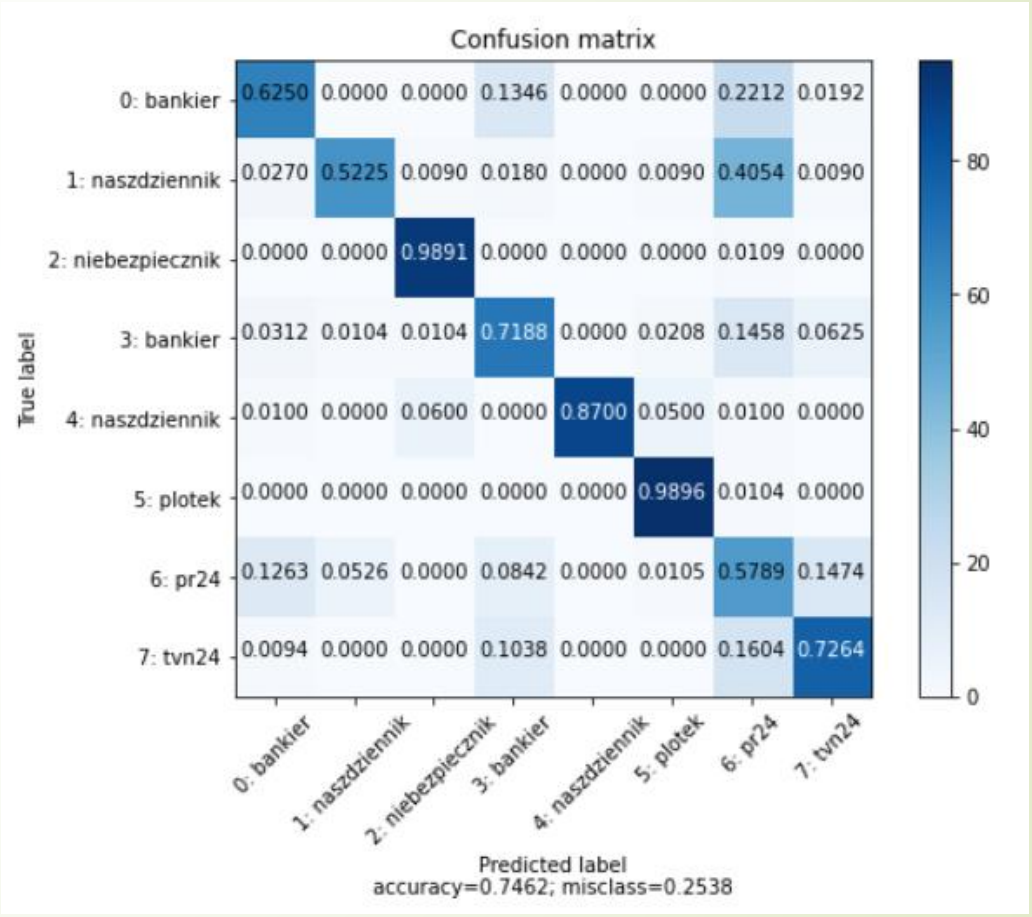
	precision	recall	f1-score	support
0	0.000	0.000	0.000	104
1	0.000	0.000	0.000	111
2	0.883	0.989	0.933	92
3	0.000	0.000	0.000	96
4	0.937	0.890	0.913	100
5	0.000	0.000	0.000	96
6	0.154	0.958	0.266	95
7	0.000	0.000	0.000	106
accuracy			0.339	800
macro avg	0.247	0.355	0.264	800
weighted avg	0.237	0.339	0.253	800



Eksperyment: klasyfikacja

- SVM + TfidfTransformer (Scikit-Learn)
- Przyzwoita klasyfikacja

	precision	recall	f1-score	support
0	0.765	0.625	0.688	104
1	0.906	0.523	0.663	111
2	0.919	0.989	0.953	92
3	0.663	0.719	0.690	96
4	1.000	0.870	0.930	100
5	0.913	0.990	0.950	96
6	0.350	0.579	0.437	95
7	0.770	0.726	0.748	106
accuracy			0.746	800
macro avg	0.786	0.753	0.757	800
weighted avg	0.789	0.746	0.755	800



Podsumowanie i wnioski

- Spacy dostarcza duży model języka polskiego i on działa (np. lematyzacja, tokenizacja, NER)… ale niekoniecznie można go wykorzystać w innych komponentach
- Klasyfikacja ze Scikit-learn działa całkiem nieźle z ustawieniami domyślnymi elementów pipeline
- NLP z językiem polskim to duża rzecz

Dziękuję za uwagę!

CMSPTCP@GMAIL.COM