Text analysis

ELIZAVETA LASHKEVICH
Business Analytics & Big Data Systems
211

Data

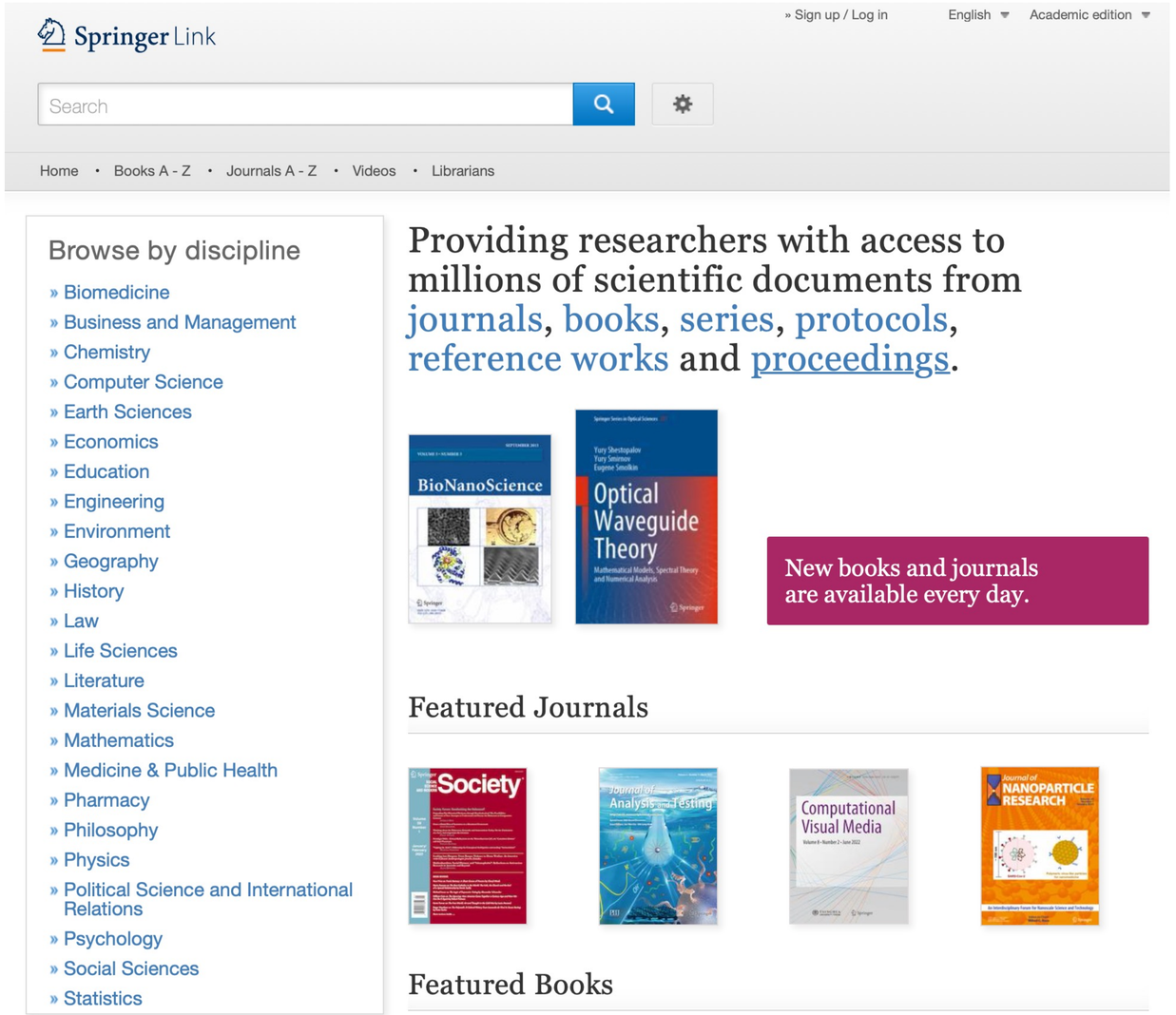
Data source

Springer

Articles, researches by applied mathematics from SpringerLink

Data example

```
['Marina De Géa Neves, Ronei Jesus Poppi, Márcia Cristina Breitkreitz,',
 'Authentication of plant-based protein powders and classification of adulterants as whey, soy protein, and wheat using FT-NIR
in tandem with OC-PLS and PLS-DA models,',
 'Food Control,',
 'Volume 132,',
 '2022,',
 '108489,',
 'ISSN 0956-7135,',
 'https://doi.org/10.1016/j.foodcont.2021.108489.',
 '(https://www.sciencedirect.com/science/article/pii/S0956713521006277)',
 'Abstract: This study aimed at developing a non-invasive and rapid method to determine the authenticity of plant-based protein
powders (free of soy, lactose, and gluten), and classify possible adulterations in the powders using near-infrared spectroscopy
(NIR) and chemometric tools. Three potential powder adulterants were investigated: soy protein, whey (lactose source), and whea
t (gluten source). The goal was to achieve untargeted and targeted detection to solve problems related to the authentication of
the protein powders and the classification of the adulterants. For this purpose, the OC-PLS (one-class partial least squares) m
odel was used for authentication and the PLS2-DA (partial least squares discriminant analysis) model was used to classify the a
dulterants. VIP (variable importance in projection) scores were used to confirm the main relevant variables and spectral ranges
were responsible for each class in PLS2-DA. Laboratory samples were prepared by adding 10, 15, 20, 25, 30, 35 and 40% (w/w) of
each adulterant into pure plant-based protein powder samples. In total, 47 pure plant-based protein powder samples and 144 adul
terated samples were analyzed. The analysis results indicate a promising way of combining one-class (OC-PLS) with multiclass (P
LS-DA) methods, in tandem with NIR to investigate plant-based protein powders. Due to the speed, high sensitivity, and specific
ity of the methodology, and no requirement of sample preparation, the proposed methodology could be successfully used in a rang
e of 10–40% of adulteration, to verify the authenticity of the plant-based protein powders and to classify adulterants into soy
```



Data processing

Data loading

```
In [6]: text_1="text_1.txt"
        text_2="text_2.txt"
        text_3="text_3.txt"
        text_4="text_4.txt"
        text_5="text_5.txt"
        text_6="text_6.txt"
        text_7="text_7.txt"
        text_8="text_8.txt"
        text_9="text_9.txt"
        text_10="text_10.txt"
```

```
In [7]: os.chdir(path_input)
        text_rdd_1=sc.textFile(text_1)
        text_rdd_2=sc.textFile(text_2)
        text_rdd_3=sc.textFile(text_3)
        text_rdd_4=sc.textFile(text_4)
        text_rdd_5=sc.textFile(text_5)
        text_rdd_6=sc.textFile(text_6)
        text_rdd_7=sc.textFile(text_7)
        text_rdd_8=sc.textFile(text_8)
        text_rdd_9=sc.textFile(text_9)
        text_rdd_10=sc.textFile(text_10)
```

```
In [8]: text_rdd = text_rdd_2.union(text_rdd_1)
        text_rdd = text_rdd.union(text_rdd_3)
        text_rdd = text_rdd.union(text_rdd_4)
        text_rdd = text_rdd.union(text_rdd_5)
        text_rdd = text_rdd.union(text_rdd_6)
        text_rdd = text_rdd.union(text_rdd_7)
        text_rdd = text_rdd.union(text_rdd_8)
        text_rdd = text_rdd.union(text_rdd_9)
        text_rdd = text_rdd.union(text_rdd_10)
```

Data cleaning

```
def lower_clean_str(x):
  punc='!"#$%&\'()*+,./:;<=>?@[\\]^_`{|}~—'
  lowercased_str = x.lower()
  for ch in punc:
    lowercased_str = lowercased_str.replace(ch, '')
    lowercased_str = lowercased_str.replace('  ', ' ')
  return lowercased_str
```

```
text_rdd = text_rdd.map(lower_clean_str)
```

```
text_rdd.take(2)
```

```
['marina de géa neves ronei jesus poppi márcia cristina breitkreitz',
 'authentication of plantbased protein powders and classification of adulterants as whey soy protein and wheat using ftnir in tan
dem with ocpls and plsda models']
```

Stopwords exclusion

```
from nltk.corpus import stopwords
stopwords =stopwords.words('english')
stopwords
```

```
  'hers',
  'herself',
  'it',
  "it's",
  'its',
```

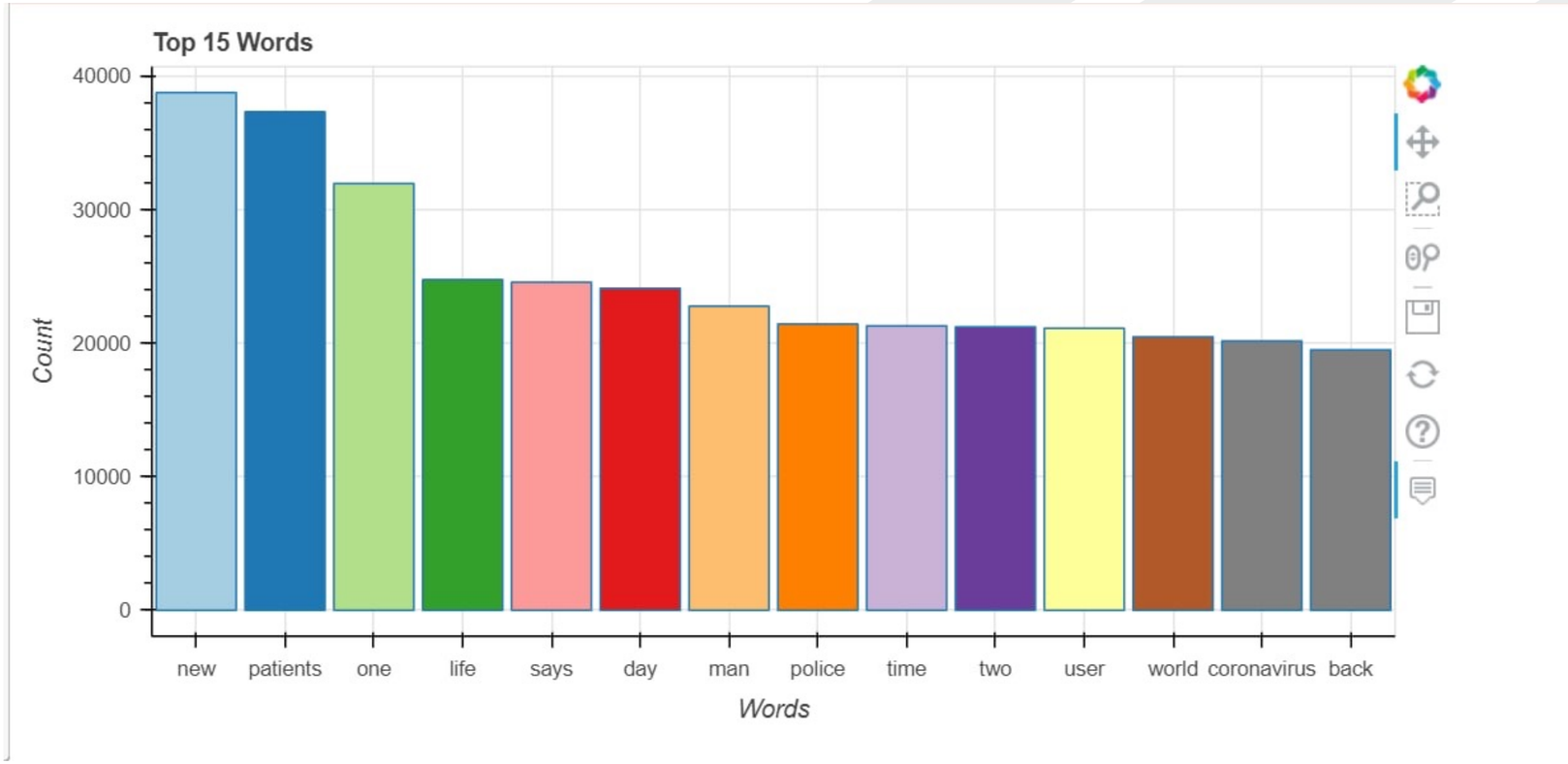
Top 15 most common words

Count	Item
38746	new
37312	patients
31935	one
24747	life
24553	says
24093	day
22762	man
21428	police
21284	time
21212	two
21106	user
20461	world
20154	coronavirus
19488	back

Data analysis

Top 15 most common words

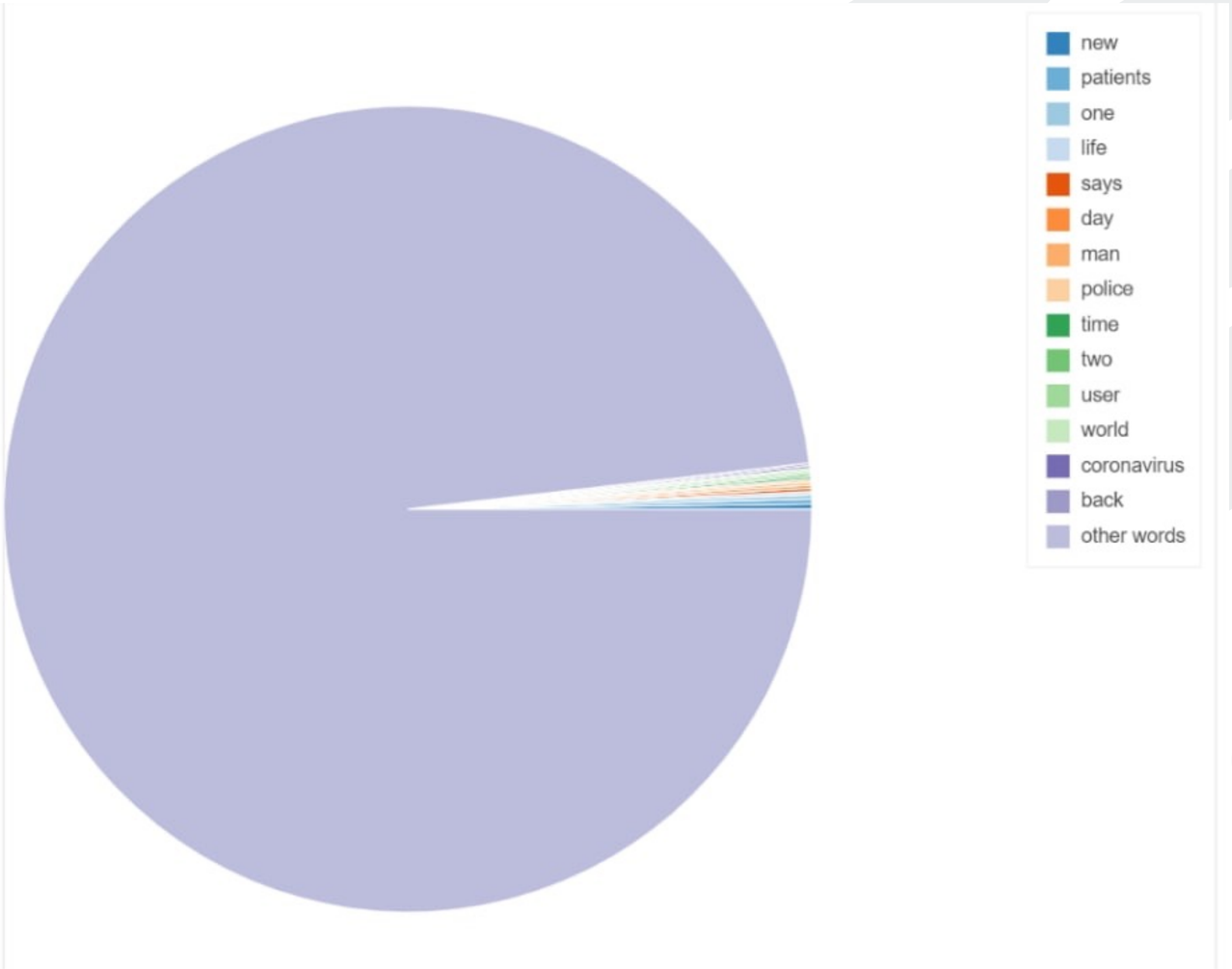
Count	Item
38746	new
37312	patients
31935	one
24747	life
24553	says
24093	day
22762	man
21428	police
21284	time
21212	two
21106	user
20461	world
20154	coronavirus
19488	back



Data analysis

Top 15 most common words with rate

Count	Item	Rate %
38746	new	0.20300985451309728
37312	patients	0.19549640457318654
31935	one	0.16732358705094103
24747	life	0.12966202626427548
24553	says	0.1286455623254033
24093	day	0.1262353900992116
22762	man	0.11926160915777421
21428	police	0.11227210970181821
21284	time	0.11151762100492342
21212	two	0.111140376656476
21106	user	0.110584989143484
20461	world	0.10720550852197602
20154	coronavirus	0.10559698053623504
19488	back	0.10210747031309657
18736492	other words	98.1699405101381



Thank you

evlashkevich@edu.hse.ru