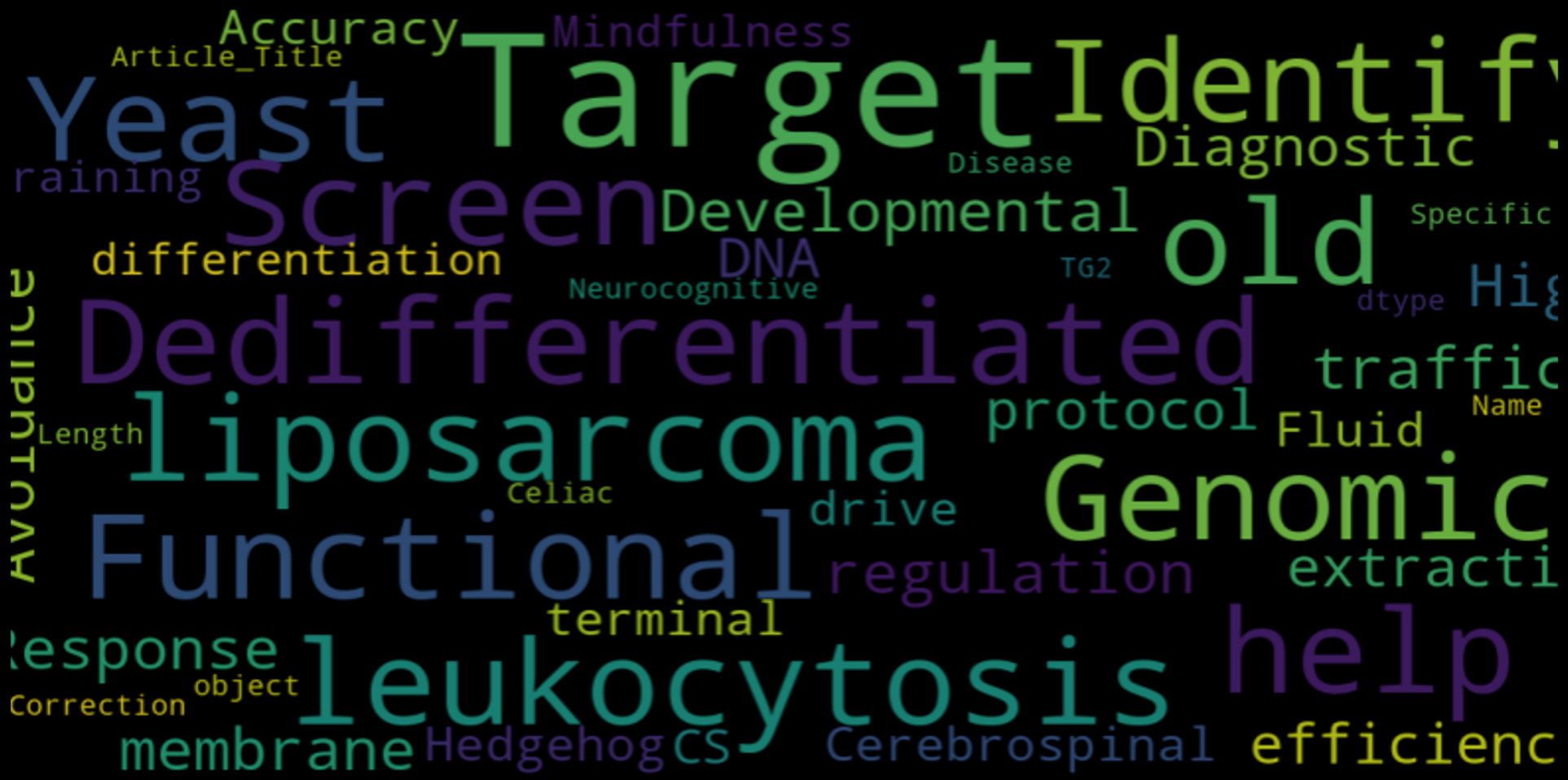
Natural Language Processing (CP8207)

(Assignment 1)

By Urmi Patel (501064008)



Steps:

1. Information retrieval

2. Pick 5 useful attributes from .nxml file
    I. Article-id (pub id type = “pmc”)
    II. Journal-Id
    III. Publisher-Name
    IV. Article-Title
    V. Abstract (Background + Methods + Results + Conclusion)

3. Convert xml data into csv file (by traversing each sub-folder)

```
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/26/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/21/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/07/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/38/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/00/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/36/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/09/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/31/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/30/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/37/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/08/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/01/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/06/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/39/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/24/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/23/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/15/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/12/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/13/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/14/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/22/
/Users/urmi/Documents/NLP/Assignment_1/pmc-text-00/25/
       Article_Id               Journal_Id                    Publisher_Name  \
0         2213094                J Exp Med  The Rockefeller University Press
1         2203997       World J Surg Oncol                     BioMed Central
2         2211553                PLoS Pathog         Public Library of Science
3         2199963                J Cell Biol  The Rockefeller University Press
4         2203094            J Gen Physiol  The Rockefeller University Press
...           ...                      ...                               ...
199995    2233773     Int J Plant Genomics    Hindawi Publishing Corporation
199996    2226121            J Gen Physiol  The Rockefeller University Press
199997    2223773  J Biophys Biochem Cytol  The Rockefeller University Press
199998    2229684  J Biophys Biochem Cytol  The Rockefeller University Press
199999    2228996            J Gen Physiol  The Rockefeller University Press

                                      Article_Title  \
0                                       Too old to help
1       Dedifferentiated liposarcoma with leukocytosis...
2       A Functional Genomic Yeast Screen to Identify ...
3       Developmental regulation of membrane traffic o...
4                              The Avoidance Response in
...                                                   ...
199995  Progress in Understanding and Sequencing the G...
199996  Destruction of Sodium Conductance Inactivation...
199997  The Use of Carbon Films to Support Tissue Sect...
199998  FURTHER OBSERVATIONS ON THE FINE STRUCTURE OF ...
199999                          Actions of ryanodine

                                             Abstract
0
1       Granulocyte-colony-stimulating factor (G-CSF) ...
2
3
4
...                                                 ...
199995
199996
199997
199998
199999

[200000 rows x 5 columns]
urmi@Urmis-MacBook-Pro Downloads % clear
```

XML to CSV

- A total of 200000 rows and 5 columns extracted from one folder
- After merging, a total of 733328 rows and 5 columns were extracted

```
In [3]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 733328 entries, 0 to 733327
Data columns (total 6 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Unnamed: 0      733328 non-null  int64
 1   Article_Id      733328 non-null  int64
 2   Journal_Id      733328 non-null  object
 3   Publisher_Name  693991 non-null  object
 4   Article_Title   722952 non-null  object
 5   Abstract        219729 non-null  object
dtypes: int64(2), object(4)
memory usage: 33.6+ MB
```

Preprocessing / Cleaning Steps

- Tokenization
- Stripped the tokenized abstract
- Removed punctuation
- Removed stopwords
- Made all text in lower case
- Applied stemmer (nltk.stem.PorterStemmer)
- Applied lemmatizer (nltk.stem.WordNetLemmatizer)

Results…A Small Sample of the Entire Dataset

Original Abstract:

- Odorant binding proteins (OBPs) are believed to shuttle odorants from the environment to the underlying odorant receptors, for which they could potentially serve as odorant presenters.

Tokenized Abstract:

- ['Odorant', 'binding', 'proteins', '(', 'OBPs', ')', 'are', 'believed', 'to', 'shuttle', 'odorants', 'from', 'the', 'environment', 'to', 'the', 'underlying', 'odorant', 'receptors', ',', 'for', 'which', 'they', 'could', 'potentially', 'serve', 'as', 'odorant', 'presenters', '.']

Removing Punctuations:

- ['Odorant', 'binding', 'proteins', 'OBPs', 'are', 'believed', 'to', 'shuttle', 'odorants', 'from', 'the', 'environment', 'to', 'the', 'underlying', 'odorant', 'receptors', 'for', 'which', 'they', 'could', 'potentially', 'serve', 'as', 'odorant', 'presenters']

Results…

- ['Odorant', 'binding', 'proteins', 'OBPs', 'are', 'believed', 'to', 'shuttle', 'odorants', 'from', 'the', 'environment', 'to', 'the', 'underlying', 'odorant', 'receptors', 'for', 'which', 'they', 'could', 'potentially', 'serve', 'as', 'odorant', 'presenters’]

Removing STOPWORDS

- ['odorant', 'binding', 'proteins', 'obps', 'believed', 'shuttle', 'odorants', 'environment', 'underlying', 'odorant', 'receptors', 'could', 'potentially', 'serve', 'odorant', 'presenters’]

Stemming

- ['odor', 'bind', 'protein', 'obp', 'believ', 'shuttl', 'odor', 'environ', 'underli', 'odor', 'receptor’, 'could', 'potenti', 'serv', 'odor', 'present’]

Lemmatizing

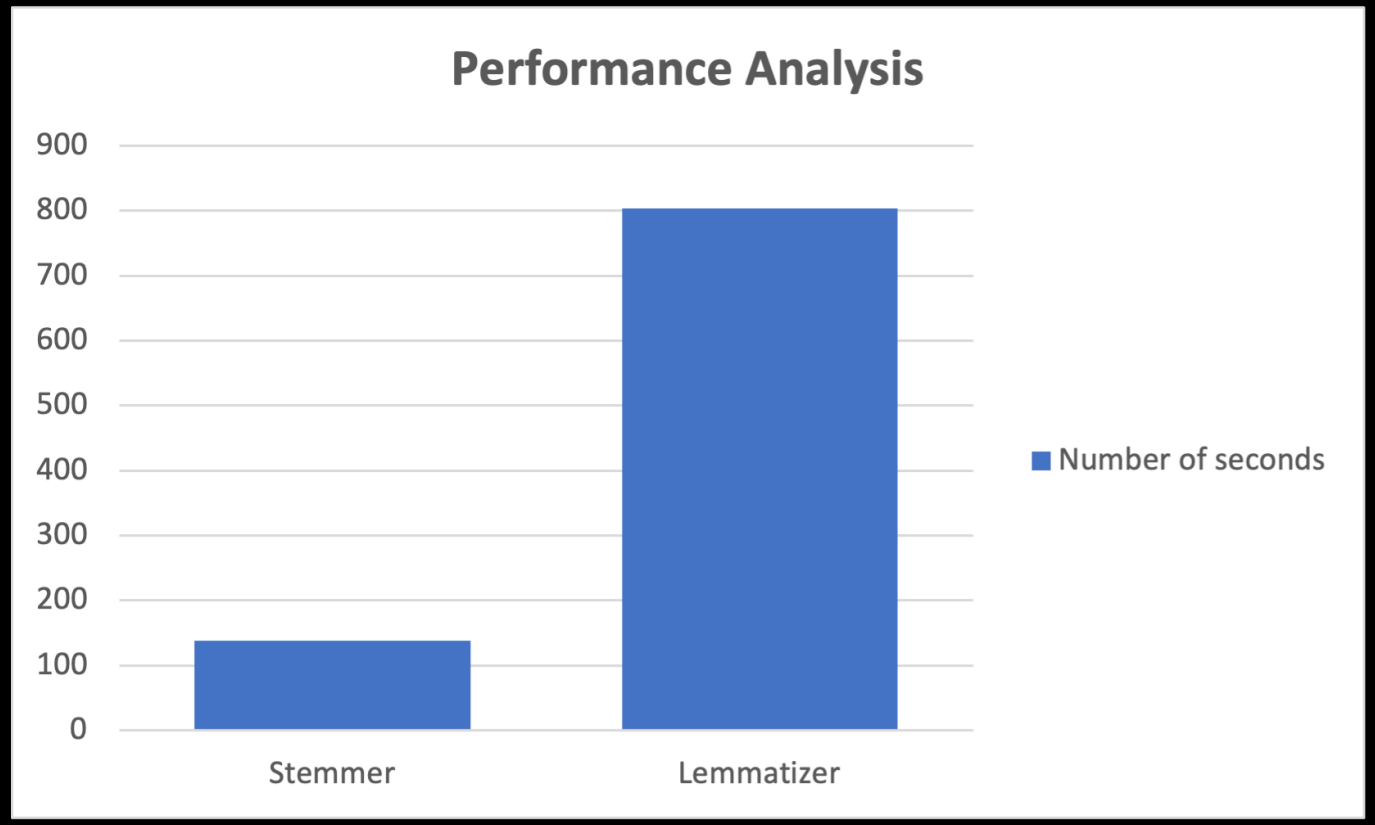
- ['odorant', 'binding', 'protein', 'obps', 'believed', 'shuttle', 'odorants', 'environment', 'underlying', 'odorant', 'receptor', 'could', 'potentially', 'serve', 'odorant', 'presenter’]

Stemming VS Lemmatization

- Converting a word into its base form
- Stem might not be an actual word
- Follow Rule based Algorithm
- Less performance time

- Converting word into some meaningful base form
- Lemma is an actual language word
- Check for meaningful root word in dictionary
- Higher performance time

Performance Analysis



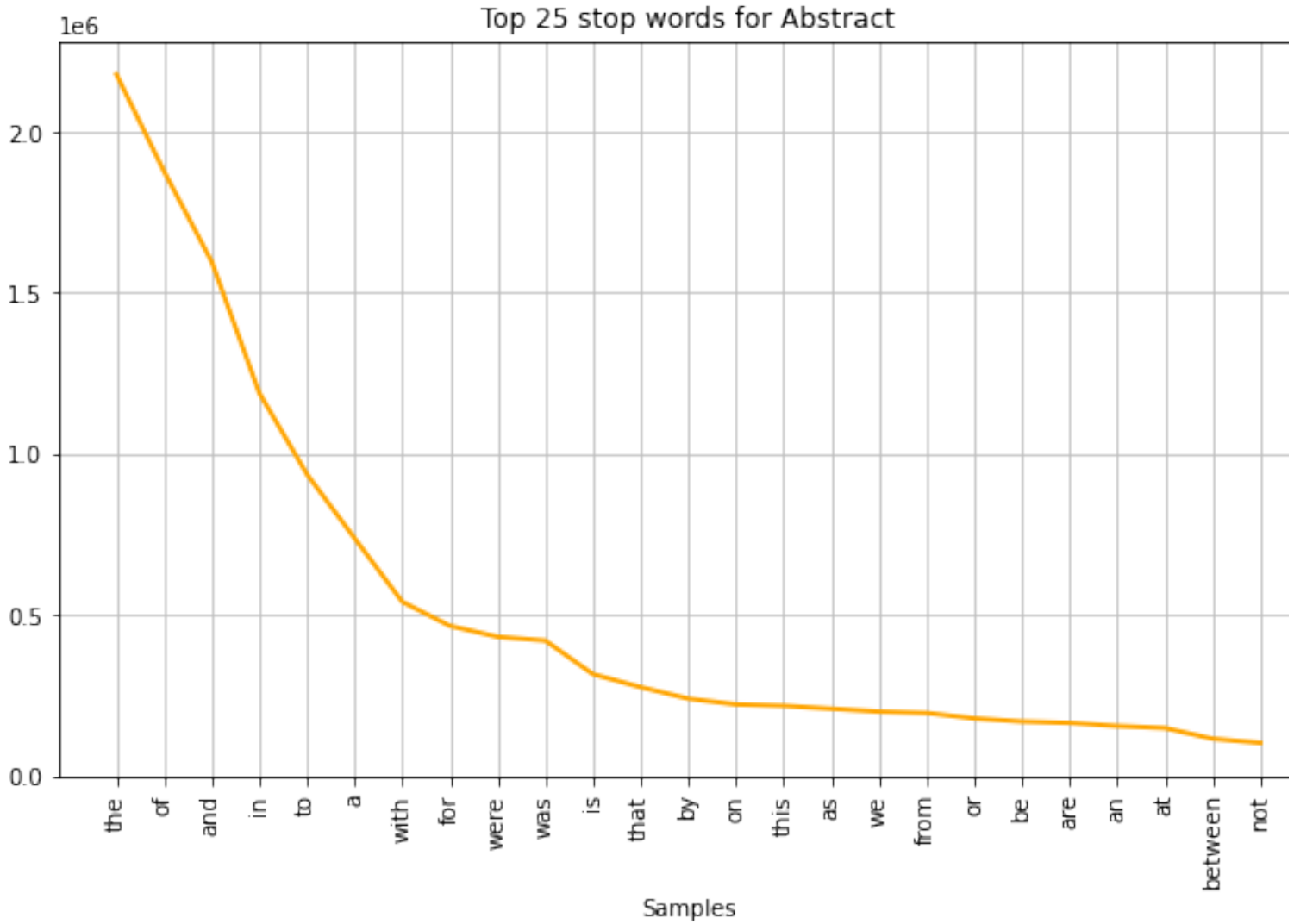
- Process of stemming takes around 2-2.5 minutes
- Lemmatizing takes around 12-13 minutes.

Is it worth use of lemmatization ?

- It gives language an accurate result. If we are working with a language-based application where language is an important part, lemmatization should be used.
- If the focus is on performance speed, then stemming should be used.
- For example: if someone wants to find the most frequent words..?

    - They can use stemming as it is much faster and gives a similar type of output as lemmatization.

Top 25 Used STOPWORDS



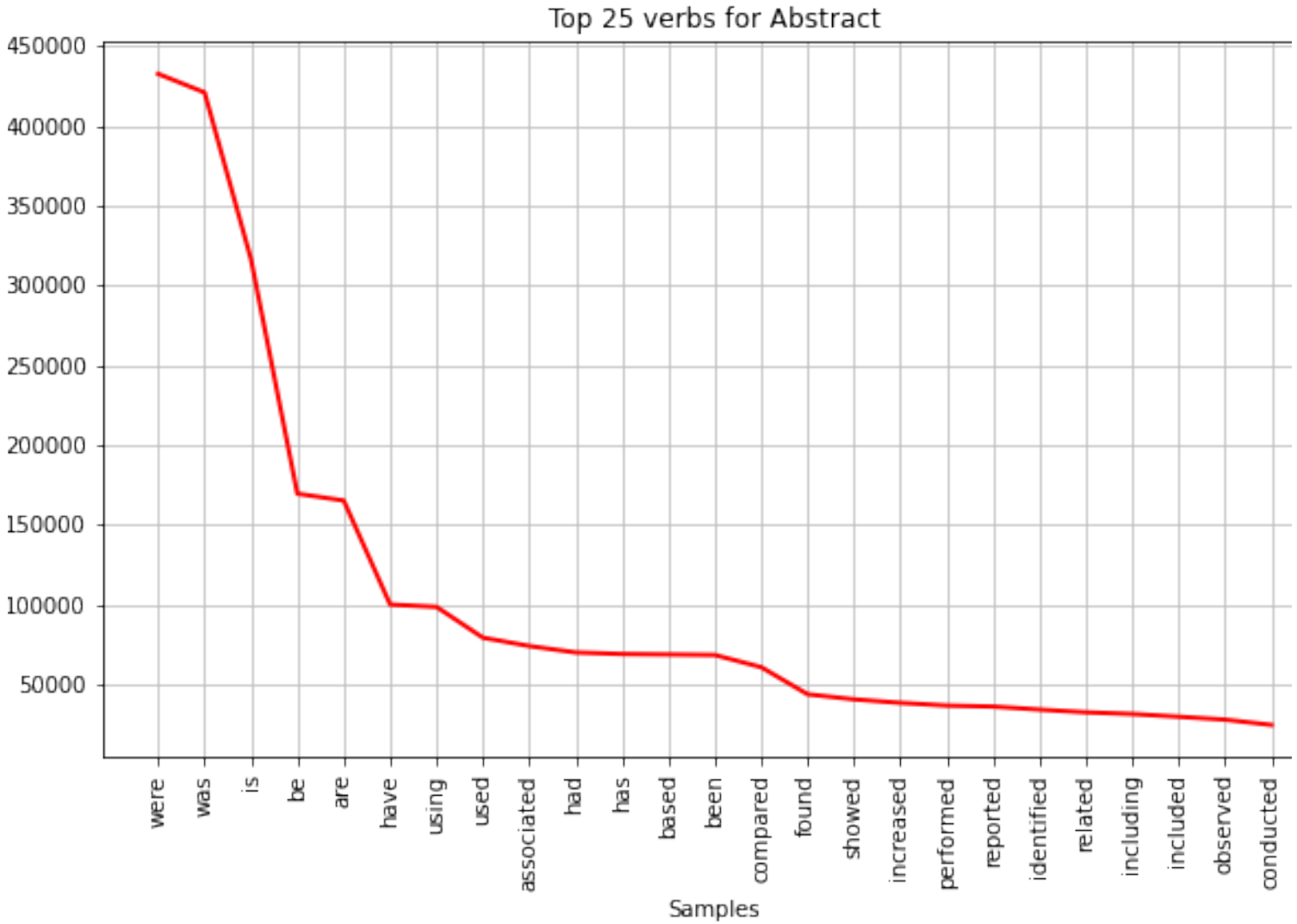
STOPWORD:

Generally, do not provide any meaningful information

“The” - most used stopword inside the abstract part of the article

“not” - least used stopword

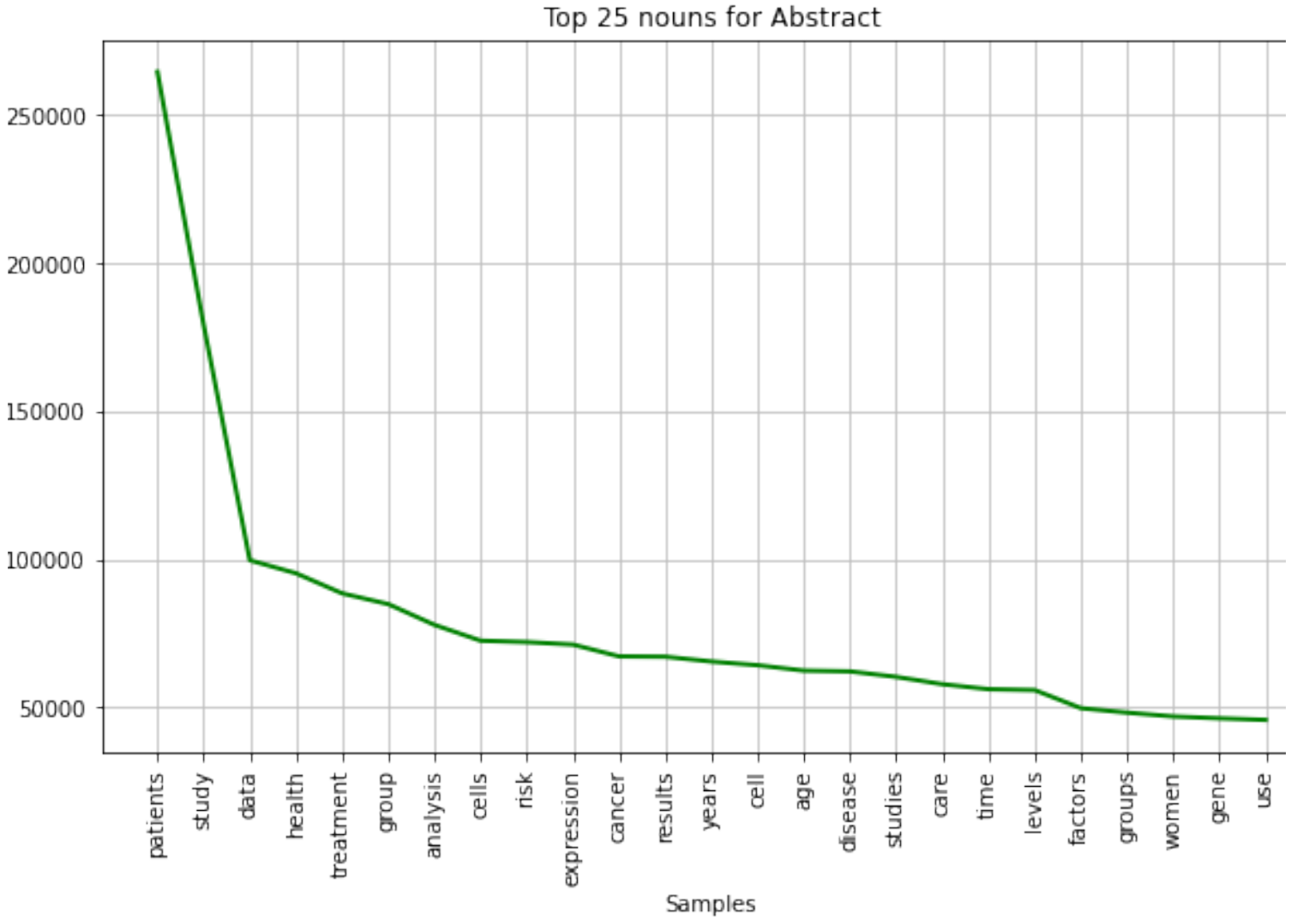
Top 25 Used Verbs



Verb class includes the words referring to actions and processes.

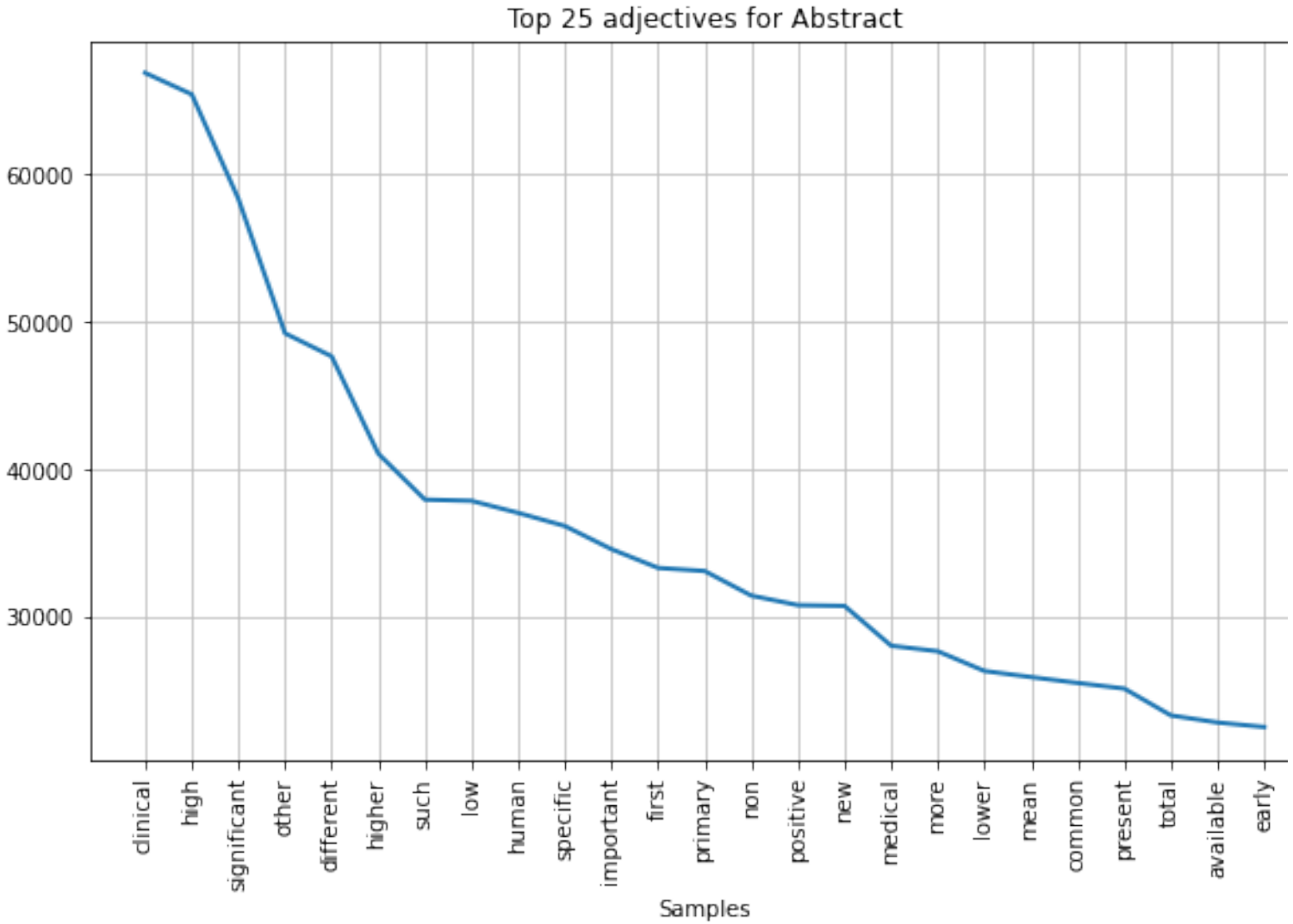
“were” is the most and “conducted” is the least used verbs inside the abstract data

Top 25 Used Nouns



A word that identifies a person, place or thing, or names

Top 25 Used Adjectives



Describes properties or qualities.

“Clinical” is the most used adjective in abstract data.

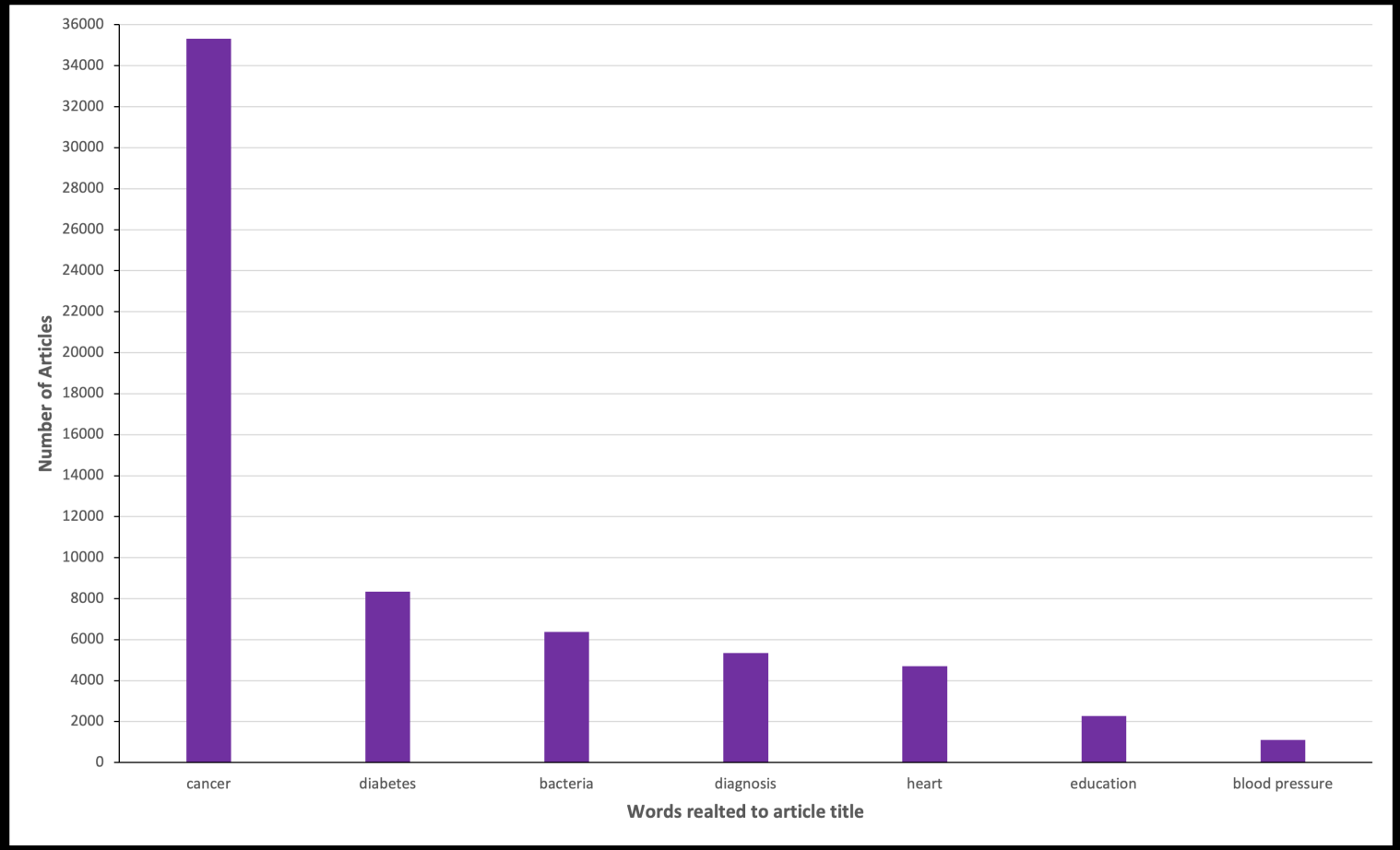
“early” is the least used adjective for abstract data

What percentage of articles published was by a particular publisher in 2014?

Out[14]:

	Publisher_Name
BioMed Central	24.262562
Public Library of Science	15.482622
The Rockefeller University Press	7.811485
Hindawi Publishing Corporation	5.446180
International Union of Crystallography	3.339957
Medknow Publications & Media Pvt Ltd	3.112865
Oxford University Press	2.922084
Nature Publishing Group	2.782889
Medknow Publications	2.122074
Dove Medical Press	1.774231

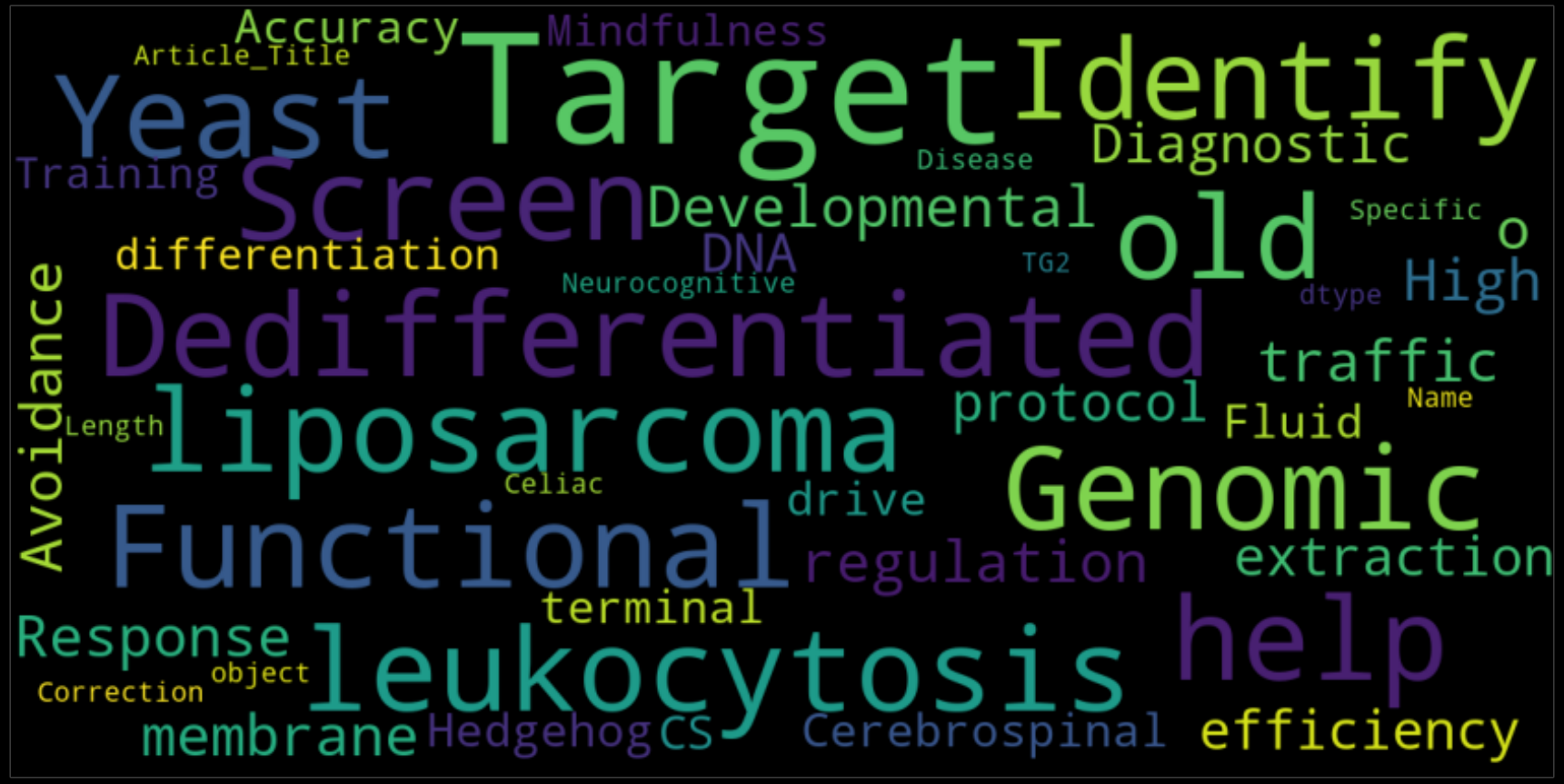
Categorized article titles and the number of articles in each category published in 2014?



- Highest published articles were related to ”Cancer” (35323)

Word cloud

- Most frequently used words within article titles in 2014



Thank you