# Documentation

## NLP Based Analysis of SweCris to understand the dynamically changing trends in Research Fund distribution

**IP Summer Sem 2023**

**Guide: Dr. N Arul Murugan**

**Acknowledgement: K. Sibin**

**Students: Arha Samanta & Abhishek Acharya**

### Project Aim

Swecris is a nationwide database where we can view how the participating research funding organizations have allocated their funding to researchers in Sweden. The database includes information from 11 different financing organizations, both public and private. The Swedish Research Council manages Swecris on behalf of the government. We can look up Swedish research projects using the Swecris database, compare them, and generate statistics from them. The data is compiled in one location from several research funding organizations.

The aim of our Project is to interact with the Swecris API and get the current database. From this database, we utilize NLP techniques to gather useful information about the most popular topics across various domains. This will be extremely helpful for young researchers as they choose which study fields to focus on for their academic and research careers.

### Implementation

1. We import the necessary libraries and fetch the data from SweCris API.

2. After collecting the responses, we extract the headers from the first item of the response. Then we convert the response to a dataframe and drop columns ending with "Sv".

3. Then we specify the columns like “ProjectID”,”ProjectTitleEn” etc. and Fetch the desired columns.

4. Then we download the necessary resources and fetch the necessary libraries for removing the stopwords, tokenize and lemmatize the whole text.

5. Then we do the **Project Title Analysis** and found the top 10 words among all the funding years. After that we plot the graph between top 10 words among all the funding years and their frequencies.

6. Then we found the top 10 words for each funding year from 2008 to 2024. Then we did the same for the various funding year ranges and plotted the graph.

7. We repeat step 5 and 6 for **Project Abstract Analysis** as well. The only modification is we first found the top 100 most common words, and removed the unnecessary ones, to finally get the top 10 relevant words.

8. For the Disease Analysis, we take the most common diseases, take their count of occurrence in each year and plot their respective graphs.

9. Further, we take words before and after 'disease', 'virus' and 'infection', and try to find any pattern in them.

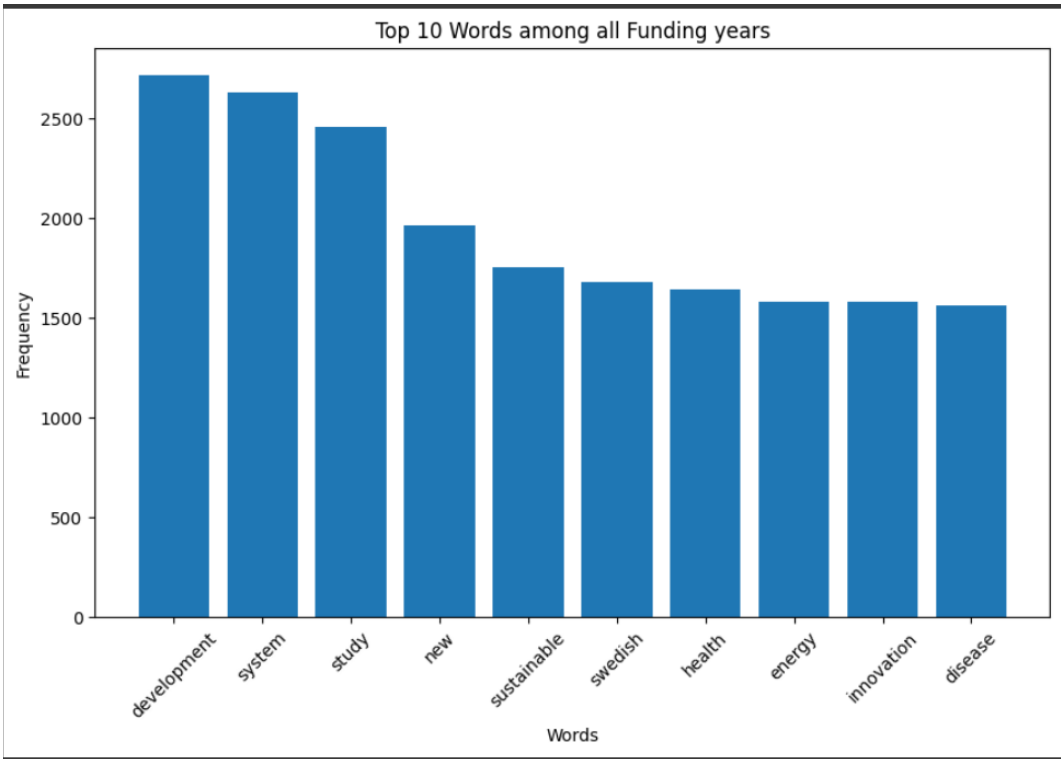
## Results

*[As some results are long, they have been clipped here. The full results are available in the Jupyter Notebook itself.]*

We get the following results:-

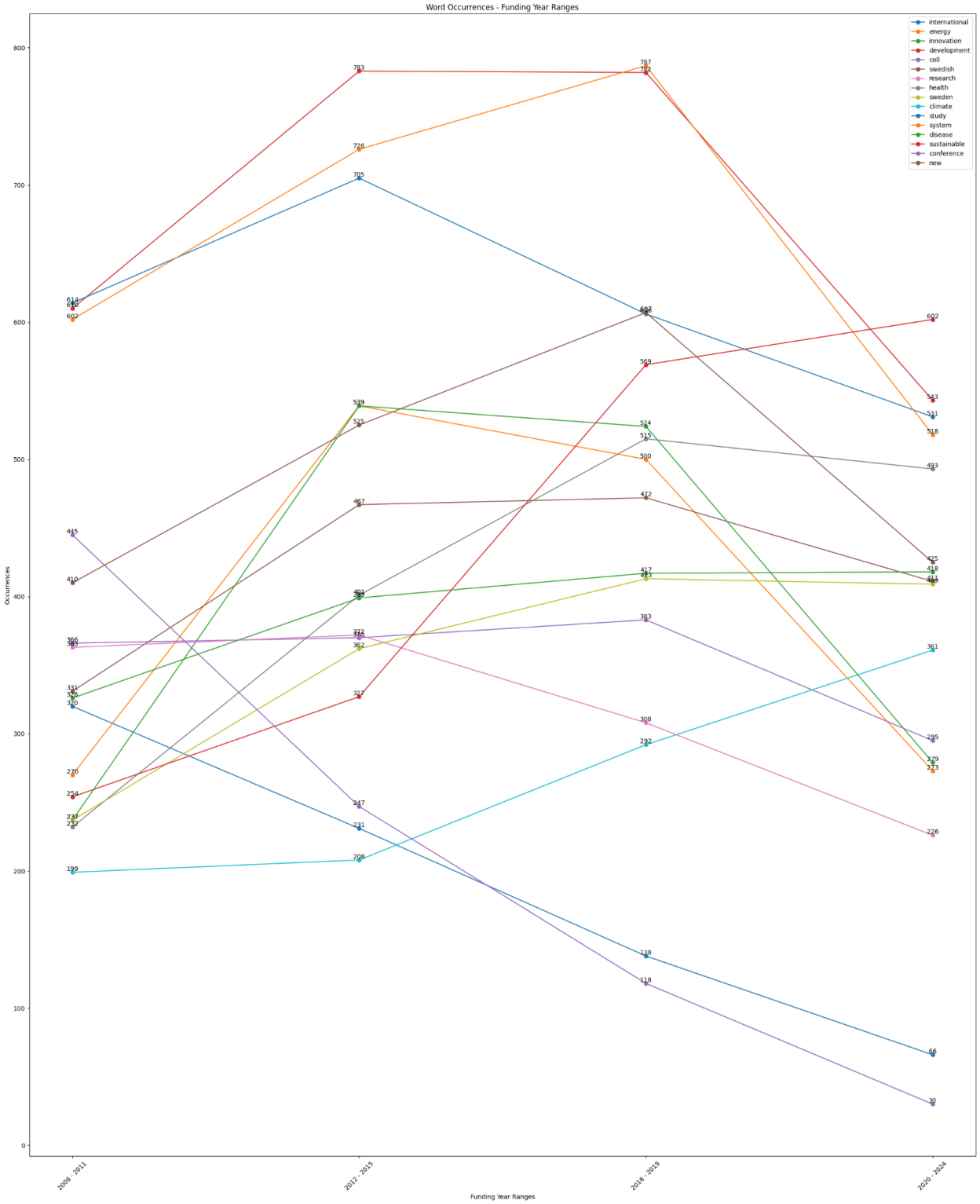
1. From Title Analysis

```
Top 10 Words among all Funding years: ['development', 'system', 'study', 'new', 'sustainable', 'swedish', 'health', 'energy', 'innovation', 'disease']
Word: development               Count: 2718
Word: system            Count: 2633
Word: study             Count: 2456
Word: new               Count: 1967
Word: sustainable               Count: 1752
Word: swedish           Count: 1681
Word: health            Count: 1641
Word: energy            Count: 1582
Word: innovation                Count: 1579
Word: disease           Count: 1560
```



```
Funding Year: 2008
Top 10 Words: ['development', 'system', 'research', 'conference', 'innovation', 'new', 'swedish', 'international', 'sustainable', 'study']

Funding Year: 2009
Top 10 Words: ['study', 'development', 'system', 'cell', 'conference', 'research', 'new', 'disease', 'molecular', 'mechanism']

Funding Year: 2010
Top 10 Words: ['study', 'system', 'development', 'conference', 'new', 'disease', 'research', 'swedish', 'international', 'cell']

Funding Year: 2011
Top 10 Words: ['system', 'study', 'development', 'conference', 'new', 'cell', 'disease', 'swedish', 'energy', 'international']

Funding Year: 2012
Top 10 Words: ['study', 'development', 'system', 'swedish', 'new', 'innovation', 'energy', 'disease', 'conference', 'research']

Funding Year: 2013
Top 10 Words: ['development', 'system', 'innovation', 'study', 'new', 'energy', 'swedish', 'sweden', 'disease', 'cell']

Funding Year: 2014
Top 10 Words: ['development', 'study', 'system', 'innovation', 'new', 'energy', 'swedish', 'research', 'health', 'disease']

Funding Year: 2015
Top 10 Words: ['system', 'development', 'energy', 'study', 'swedish', 'new', 'health', 'innovation', 'sustainable', 'disease']

Funding Year: 2016
Top 10 Words: ['system', 'development', 'study', 'energy', 'new', 'innovation', 'health', 'swedish', 'sustainable', 'disease']

Funding Year: 2017
Top 10 Words: ['development', 'system', 'new', 'study', 'sustainable', 'health', 'swedish', 'innovation', 'production', 'energy']

Funding Year: 2018
Top 10 Words: ['development', 'system', 'sustainable', 'study', 'new', 'health', 'innovation', 'energy', 'swedish', 'digital']

Funding Year: 2019
Top 10 Words: ['system', 'development', 'sustainable', 'study', 'new', 'disease', 'sweden', 'energy', 'innovation', 'health']

Funding Year: 2020
Top 10 Words: ['study', 'system', 'development', 'sustainable', 'disease', 'new', 'health', 'energy', 'swedish', 'climate']
```

```
Funding Year Range: 2008 - 2011
Top 10 Words: ['study', 'development', 'system', 'conference', 'new', 'cell', 'research', 'swedish', 'disease', 'international']

Funding Year Range: 2012 - 2015
Top 10 Words: ['development', 'system', 'study', 'energy', 'innovation', 'new', 'swedish', 'health', 'disease', 'research']

Funding Year Range: 2016 - 2019
Top 10 Words: ['system', 'development', 'new', 'study', 'sustainable', 'innovation', 'health', 'energy', 'swedish', 'disease']

Funding Year Range: 2020 - 2024
Top 10 Words: ['sustainable', 'development', 'study', 'system', 'health', 'new', 'disease', 'swedish', 'sweden', 'climate']
```

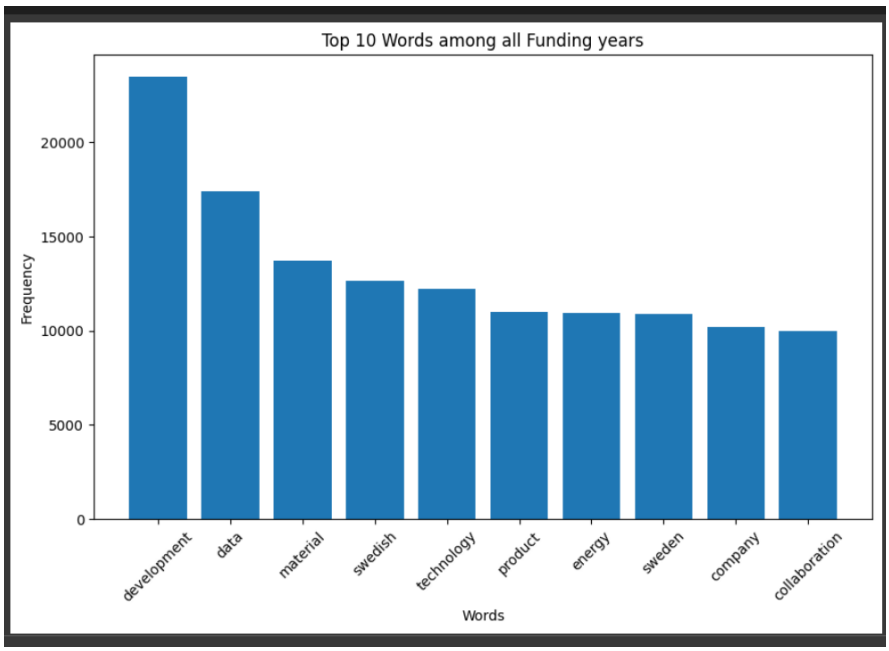

Word Occurrences - Funding Year Ranges
international
energy
innovation
development
cell
swedish
research
health
sweden
climate
study
system
disease
sustainable
conference
new
Occurrences
Funding Year Ranges
2008 - 2011
2012 - 2015
2016 - 2019
2020 - 2024
0
100
200
300
400
500
600
700
800

2. Abstract Analysis

```
Top 10 Words among all Funding years: ['development', 'data', 'material', 'swedish', 'technology', 'product', 'energy', 'sweden', 'company', 'collaboration']
Word: development               Count: 23476
Word: data              Count: 17415
Word: material          Count: 13704
Word: swedish           Count: 12666
Word: technology                Count: 12223
Word: product           Count: 11012
Word: energy            Count: 10936
Word: sweden            Count: 10912
Word: company           Count: 10198
Word: collaboration             Count: 9996
```

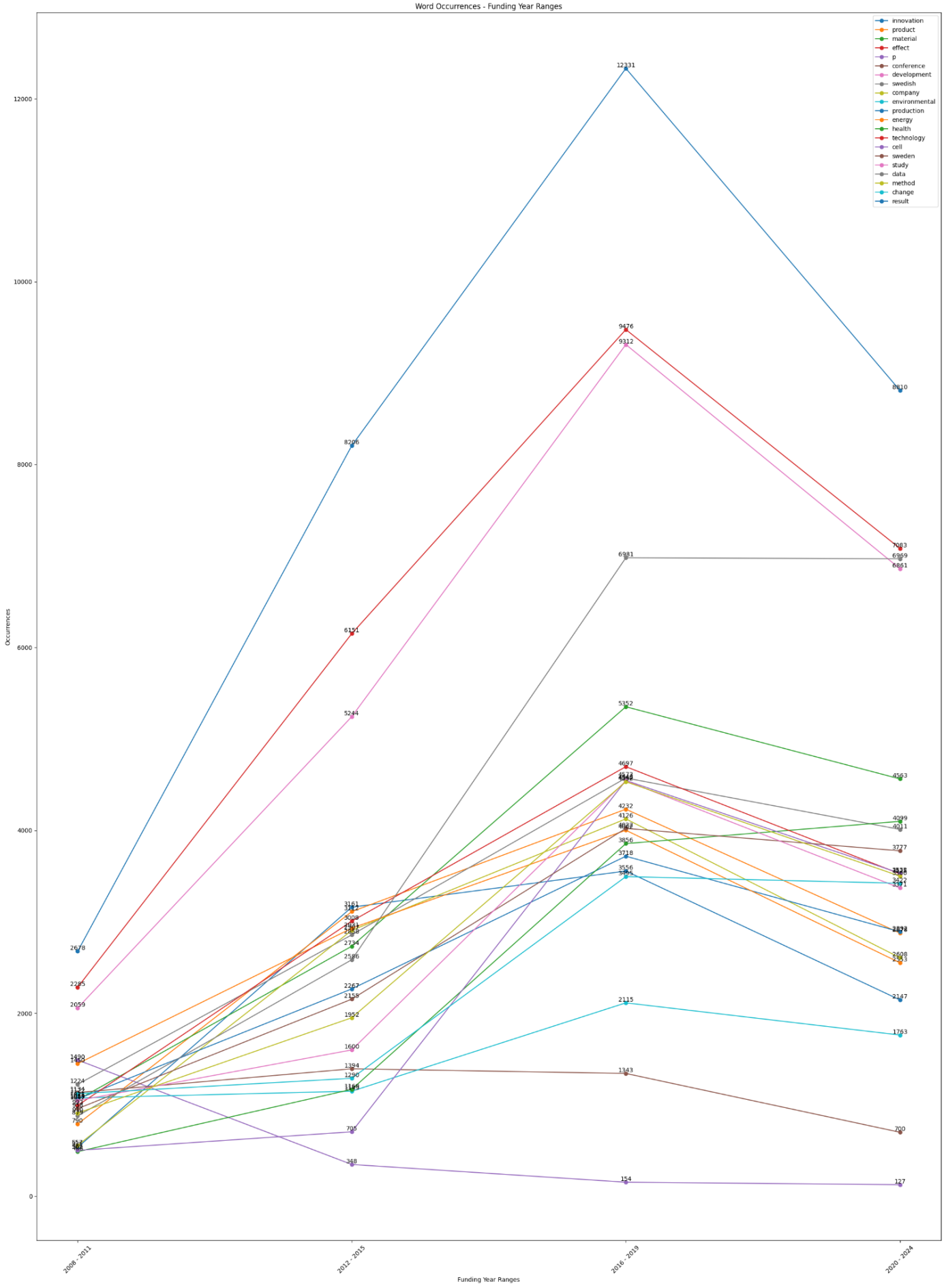
Below is the TF-IDF scores, but it doesn't give us relevant words. Hence we don't take them.

```
Word: plenty            TF-IDF Score: 763.7727440575835
Word: matrice           TF-IDF Score: 622.6404061488334
Word: processit         TF-IDF Score: 603.9216765253952
Word: vogue             TF-IDF Score: 569.5431320462849
Word: rationing         TF-IDF Score: 528.9778146611245
Word: kvarstående               TF-IDF Score: 516.096243172444
Word: attenuated                TF-IDF Score: 504.7958978079367
Word: hypertensive              TF-IDF Score: 504.0797592064767
Word: töjbara           TF-IDF Score: 504.04641755420124
Word: graviditeten              TF-IDF Score: 449.58934903090795
```

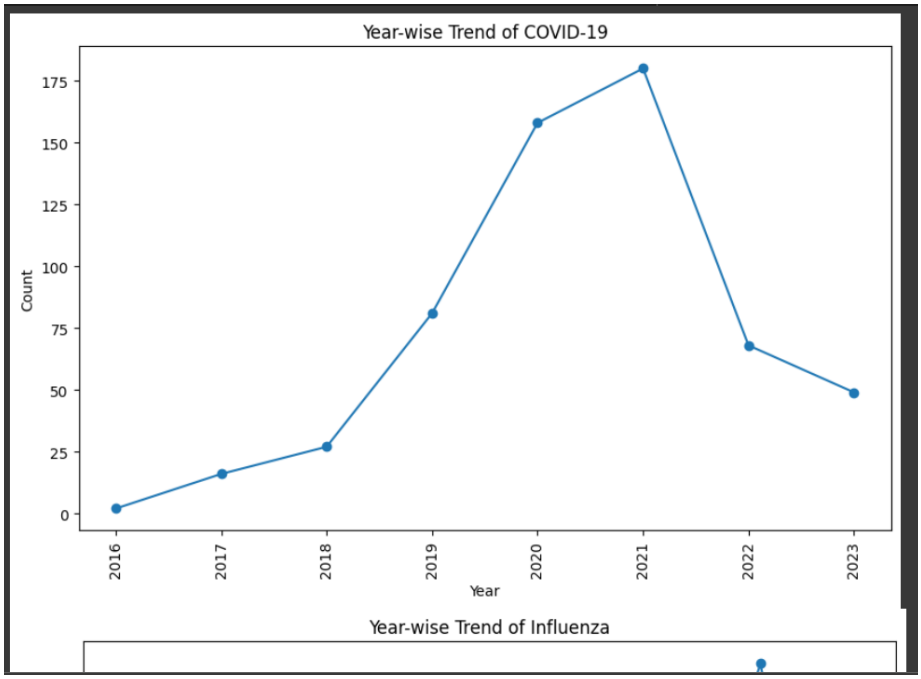


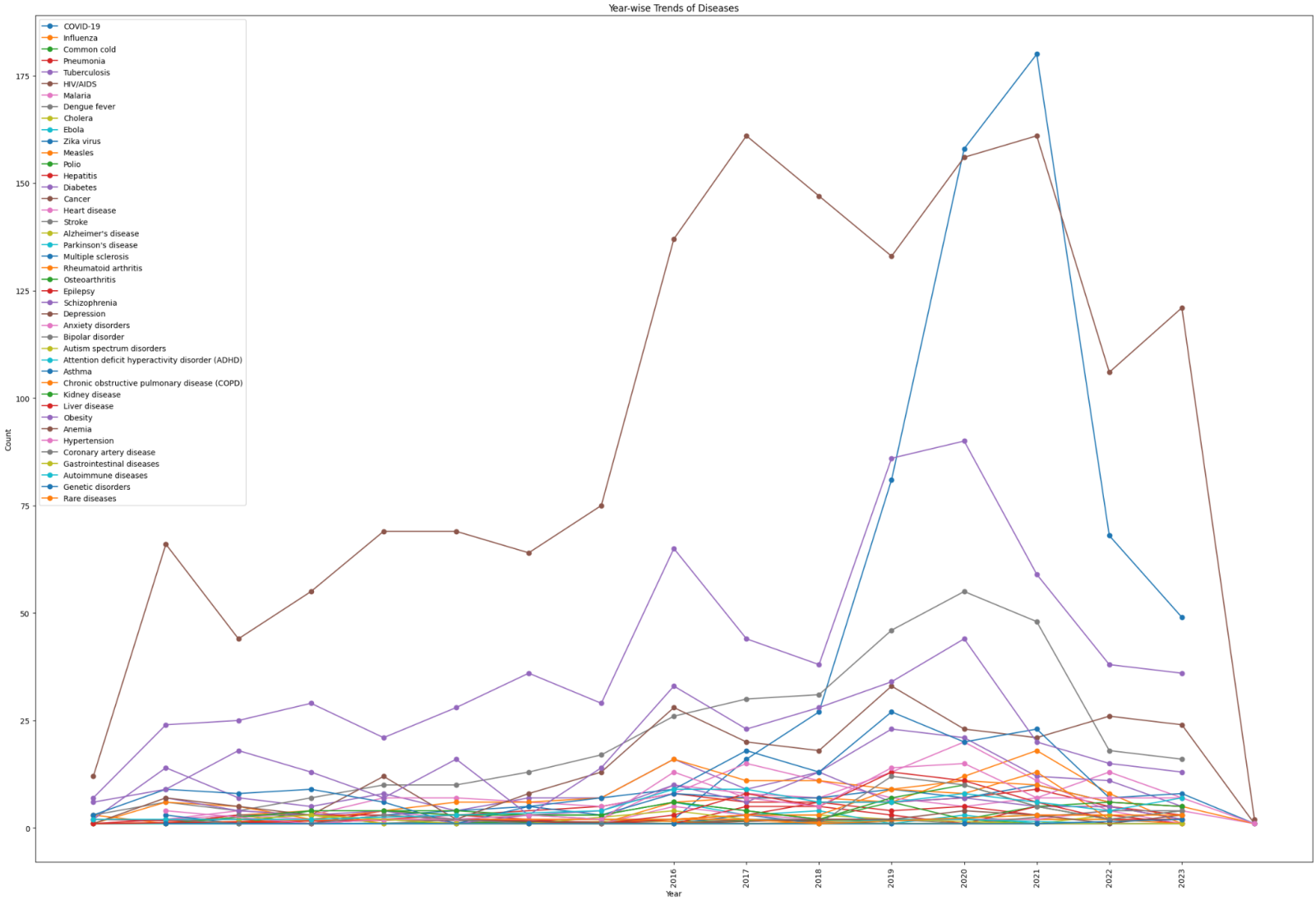
```
Funding Year: 2008
Top 10 Words: ['development', 'p', 'result', 'effect', 'change', 'environmental', 'study', 'swedish', 'method', 'climate']

Funding Year: 2009
Top 10 Words: ['result', 'development', 'effect', 'p', 'energy', 'change', 'material', 'production', 'study', 'conference']

Funding Year: 2010
Top 10 Words: ['result', 'effect', 'development', 'energy', 'swedish', 'conference', 'production', 'change', 'application', 'sweden']

Funding Year: 2011
Top 10 Words: ['result', 'effect', 'development', 'p', 'swedish', 'energy', 'conference', 'technology', 'material', 'product']

Funding Year: 2012
Top 10 Words: ['result', 'effect', 'development', 'company', 'product', 'innovation', 'swedish', 'technology', 'energy', 'industry']

Funding Year: 2013
Top 10 Words: ['result', 'effect', 'development', 'innovation', 'product', 'company', 'technology', 'energy', 'swedish', 'application']

Funding Year: 2014
Top 10 Words: ['result', 'effect', 'development', 'innovation', 'product', 'technology', 'swedish', 'energy', 'application', 'industry']

Funding Year: 2015
Top 10 Words: ['result', 'effect', 'development', 'material', 'energy', 'technology', 'data', 'product', 'swedish', 'company']

Funding Year: 2016
Top 10 Words: ['result', 'development', 'effect', 'data', 'cell', 'study', 'material', 'method', 'swedish', 'energy']

Funding Year: 2017
Top 10 Words: ['result', 'effect', 'development', 'data', 'material', 'technology', 'product', 'cell', 'method', 'swedish']

Funding Year: 2018
Top 10 Words: ['result', 'effect', 'development', 'data', 'material', 'technology', 'method', 'study', 'product', 'swedish']

Funding Year: 2019
Top 10 Words: ['result', 'effect', 'development', 'data', 'material', 'swedish', 'technology', 'study', 'collaboration', 'sweden']
```

```
Funding Year Range: 2008 - 2011
Top 10 Words: ['result', 'effect', 'development', 'p', 'energy', 'swedish', 'conference', 'change', 'environmental', 'production']

Funding Year Range: 2012 - 2015
Top 10 Words: ['result', 'effect', 'development', 'innovation', 'product', 'technology', 'energy', 'company', 'swedish', 'material']

Funding Year Range: 2016 - 2019
Top 10 Words: ['result', 'effect', 'development', 'data', 'material', 'technology', 'swedish', 'cell', 'study', 'method']

Funding Year Range: 2020 - 2024
Top 10 Words: ['result', 'effect', 'data', 'development', 'material', 'health', 'swedish', 'sweden', 'cell', 'technology']
```


Word Occurrences - Funding Year Ranges
innovation
product
material
effect
p
conference
development
swedish
company
environmental
production
energy
health
technology
cell
sweden
study
data
method
change
result
Occurrences
Funding Year Ranges
2008 - 2011
2012 - 2015
2016 - 2019
2020 - 2024
0
2000
4000
6000
8000
10000
12000
12331
8810
9476
9312
8206
7083
6969
6981
6861
6151
5244
5352
4563
4697
4232
4126
4099
4011
3856
3727
3718
3556
3522
3371
3161
3008
2734
2586
2678
2285
2267
2155
2059
2147
2115
1952
1763
1600
1420
1394
1343
1290
1224
700
705
348
154
127

3. Disease Analysis




**Performance Comparison**

After extending the work of K. Sibin, we found the top 10 most relevant words from the Title as well as the Abstract. We found the analysis of Title to be more helpful and revealed more conclusive results. The Abstract had a lot of common words, which had to be removed manually before getting the relevant words. The exact words and results are in the screenshots above.

Further, we analysed the different most common diseases over the years, by counting their occurrences to reveal their trends. Lastly, the words before and after 'disease', 'virus' and 'infection' were extracted, that can be used for any further processing.