

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. Sabin

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.

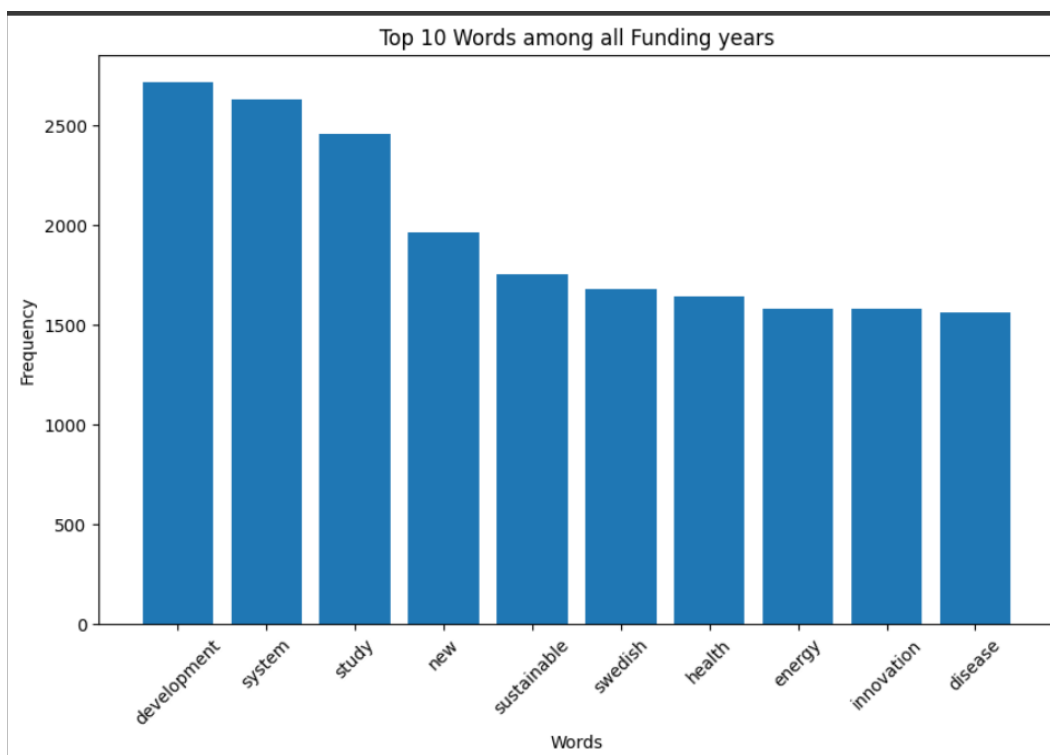
Results

[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']

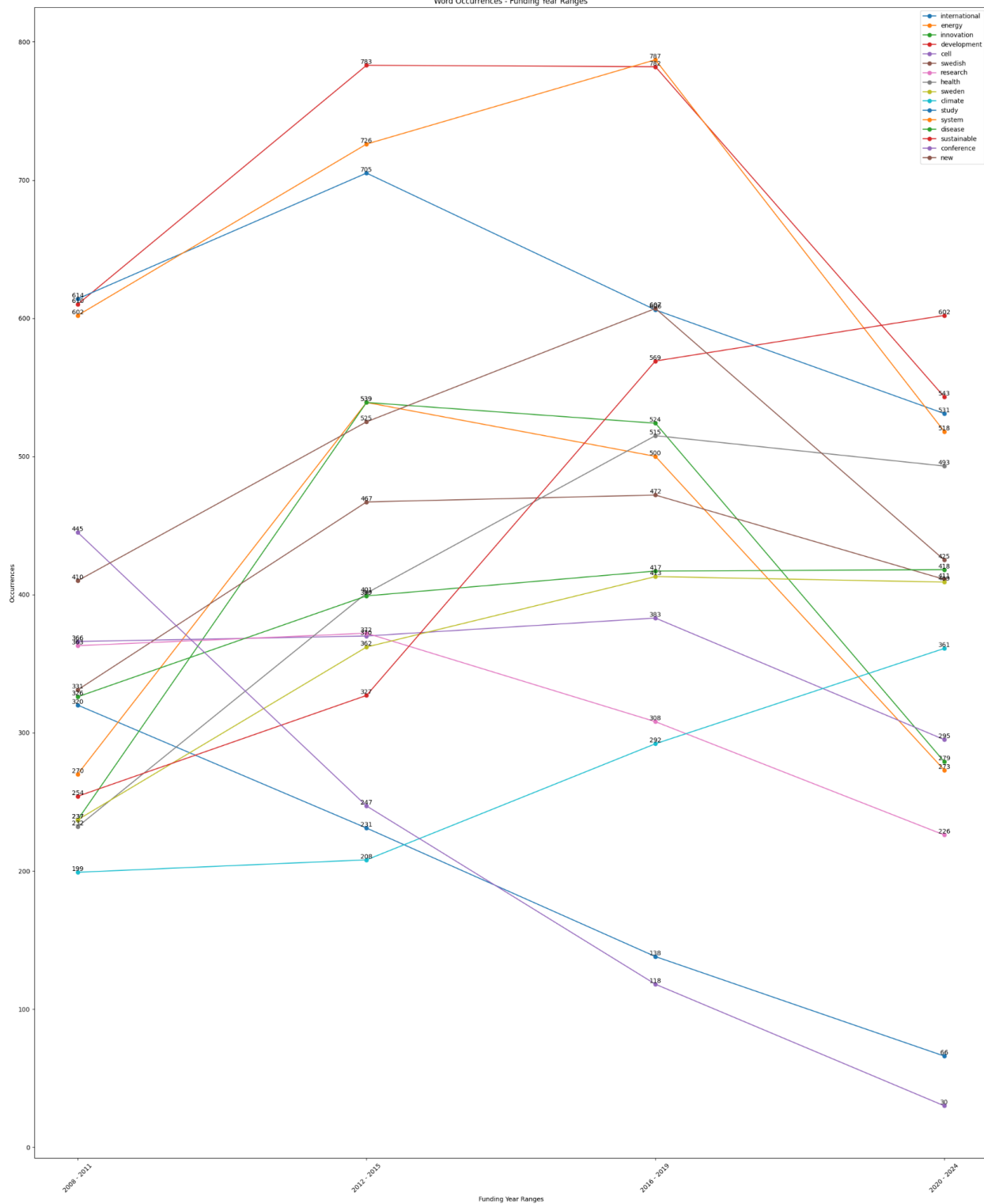
📁 Funding Year Range: 2008 - 2011
Top 10 Words: ['study', 'development', 'system', 'conference', 'new', 'cell', 'research', 'swedish', 'disease', 'international']

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

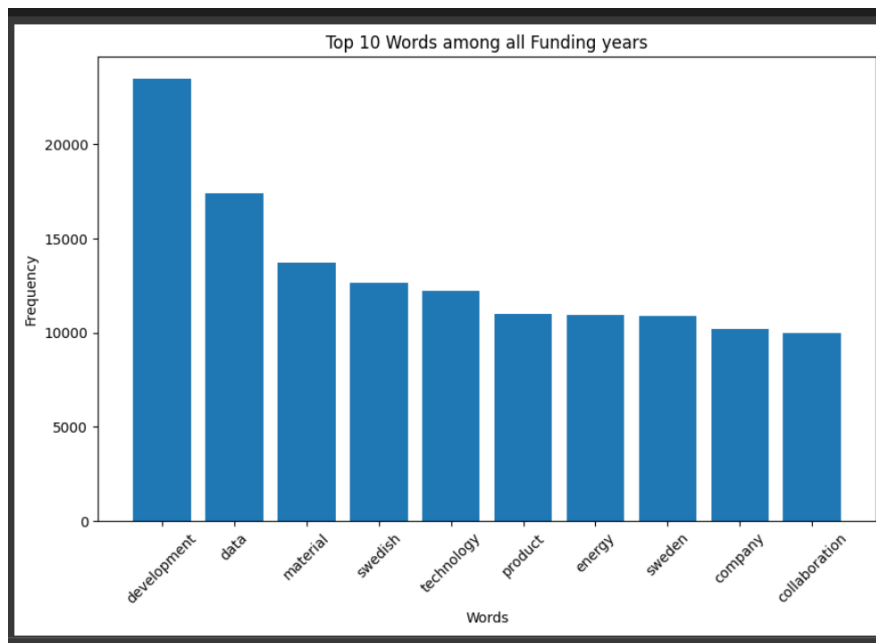


2. Abstract Analysis

```
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: tøjbara         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']

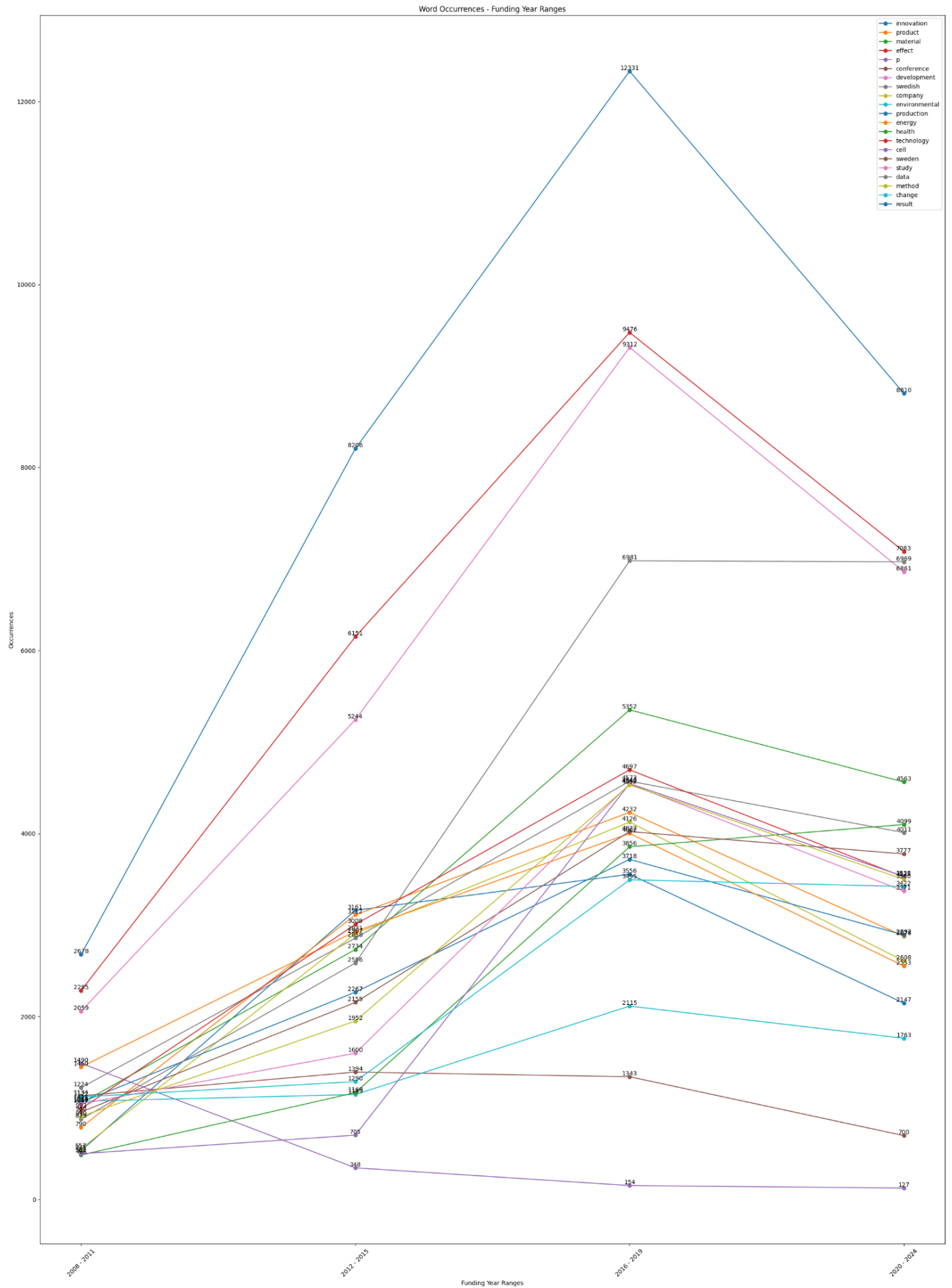
Funding Year: 2019
Top 10 Words: ['result', 'effect', 'development', 'data', 'material', 'swedish', 'technology', 'study', 'collaboration', 'sweden']
```

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

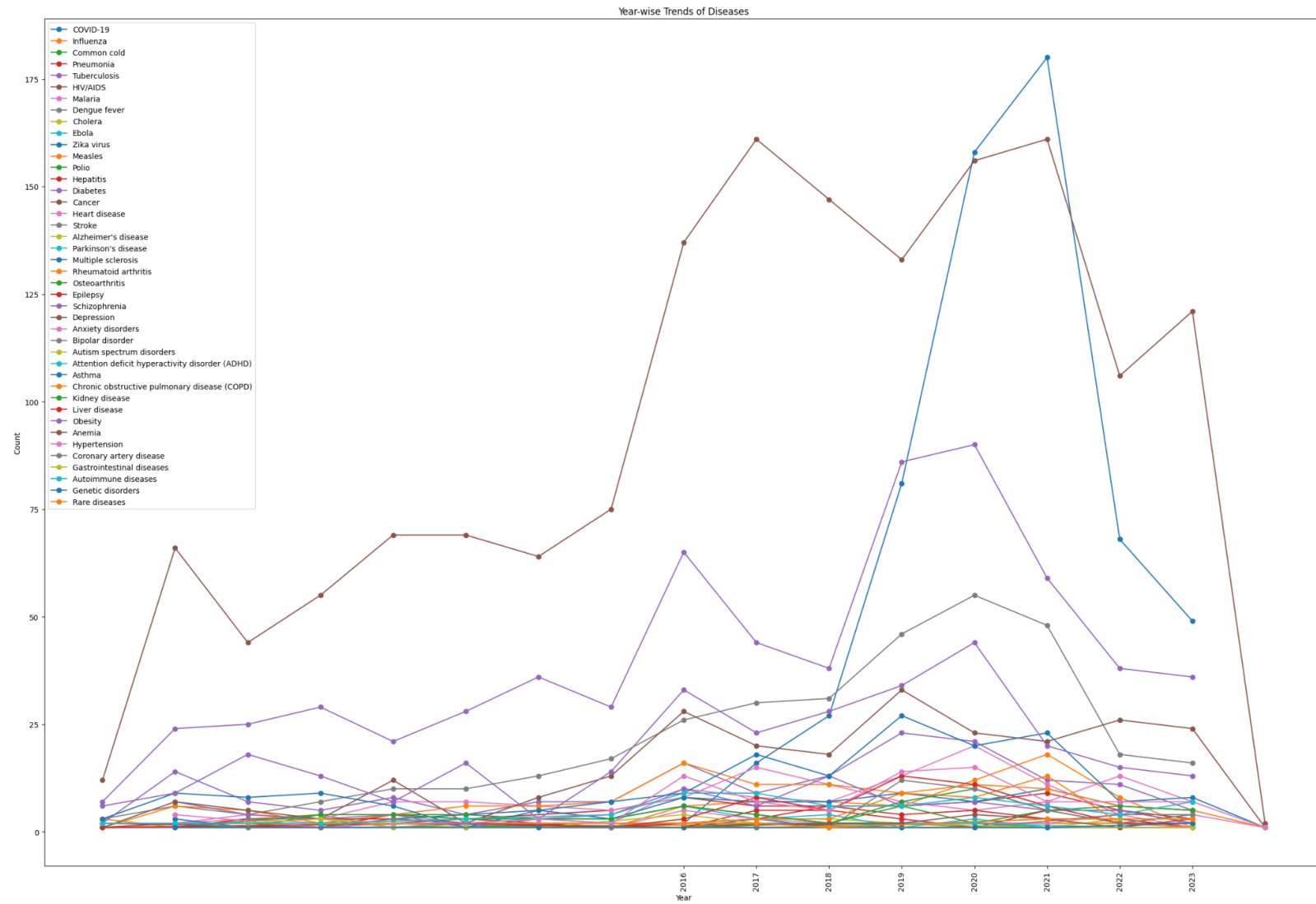
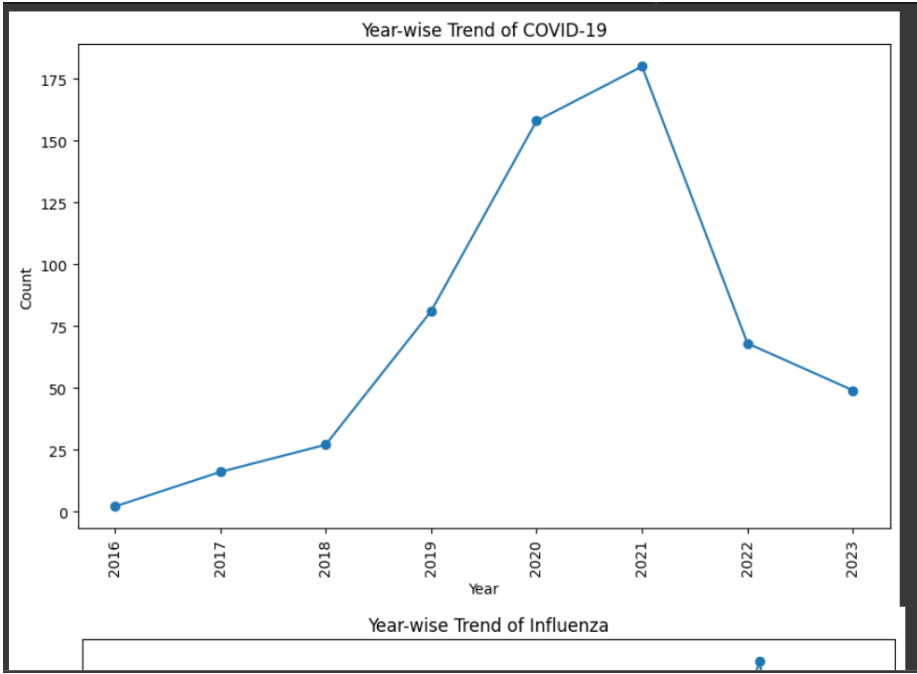
Funding Year Range: 2012 - 2015
Top 10 Words: ['result', 'effect', 'development', 'innovation', 'product', 'technology', 'energy', 'company', 'swedish', 'material']

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']
```



3. Disease Analysis




```

Words before 'disease':
Year: 2008 Words: {'coeliac', 'metabolic', 'celiac', 'genes', 'pulmonary', 'subjected', 'alzheimers', 'association', 'cardiovascular', 'liver', 'bacterial', 'autoimmune', 'foliar', 'new'}
Year: 2009 Words: {'used', 'storage', 'neurological', 'common', 'carrier', 'multi-factorial', 'elm', 'blight', 'enteric', 'heart', 'infectious', 'distinct', 'dynamics', 'follow', 'performance', 'air'}
Year: 2010 Words: {'spread', 'crustacean', 'cardiovascular', 'bowel', 'increase', 'antagonism', 'plant', 'forest', 'interventions', 'diseases', 'control', 'human', 'chagas', 'following', 'heart', 'b'}
Year: 2011 Words: {'alzheimer's', 'responsible', 'pest', 'cardiovascular', 'lung', 'skin', 'linkage', 'rot', 'asthma/respiratory', 'wasting', 'forest', 'metabolic', 'treatments', 'brood', 'add'}
Year: 2012 Words: {'pests', 'alzheimer's', 'healthcare', 'treatment', 'cardiovascular', 'wildlife', 'organs', 'lifestyle', 'differentiate', 'insect', 'forest', 'introductions', 'pops', 'zoonotic', 'v'}
Year: 2013 Words: {'markers', 'spread', 'alzheimer's', 'cardiovascular', 'function', 'bowel', 'common', 'detect', 'forest', 'milk', 'liver', 'alzheimers', 'elm', 'vw', 'metastatic', 'economic', 'var'}
Year: 2014 Words: {'endoparasitic', 'psychiatric', 'used', 'alzheimer's', 'due', 'cardiovascular', 'chronic', 'lung', 'globally', 'vivo', 'plant', 'lifestyle', 'questions', 'cardiopulmonary', 'growth'}
Year: 2015 Words: {'spraing', 'trees', 'creutzfeld-jacob', 'alzheimer's', 'asp-1-like', 'pulmonary', 'local', 'relevant', 'cardiovascular', 'chronic', 'mosquito', 'evaluation', 'gi', 'inflammation'}
Year: 2016 Words: {'contributor', 'died', 'known', 'detection', 'mdd', 'surveys', 'due', 'chronic', 'integrity-related', 'host', 'bowel', 'connected', 'reverse', 'skin', 'residual', 'alzheimers's', 'ca'}
Year: 2017 Words: {'asthmatic', 'function', 'chronic', 'hemolytic', 'bowel', 'lung', 'reverse', 'risk-factor', 'initiation', 'global', 'biggest', 'characteristics', 'gallstone', 'risk', 'liver', 'ca'}
Year: 2018 Words: {'physiology', 'chronic', 'feeding', 'transmit', 'bowel', 'bipolar', 'increase', 'upland', 'residual', 'cardiovascular', 'active', 'mechanisms', 'broad-spectrum', 'cancer', 'h'}
Year: 2019 Words: {'dysfunctional', 'predispose', 'chronic', 'function', 'successful', 'bowel', 'lung', 'host', 'explore', 'pathogens', 'active', 'mechanisms', 'suffer', 'bronchial', 'mutations', 'h'}
Year: 2020 Words: {'benign', 'musculoskeletal', 'chronic', 'explore', 'function', 'bowel', 'transmit', 'lung', 'genome-wide', 'skin', 'global', 'sorted', 'player', 'wasting', 'gallstone', 'factor', 'fa'}
Year: 2021 Words: {'mitochondrial', 'unexplored', 'due', 'chronic', 'neurological', 'bowel', 'lung', 'global', 'active', '3-year', 'mechanisms', 'sprout', 'malfunction', 'pathological', 'liver', 'fa'}
Year: 2023 Words: {'pests', 'architecture', 'detection', 'atherosclerotic', 'protection', 'chronic', 'explore', 'site', 'recently', 'residual', 'active', 'remission', 'hematologic', 'spot', 'cancer', 'cancer'}
Year: 2022 Words: {'died', 'dysfunctional', 'predictor', 'mdd', 'manage', 'nefarious', 'chronic', 'lesions', 'neurological', 'within', '20', 'large-scale', 'processes', 'factor', 'cancer', 'liver', 'liver'}
Year: 2024 Words: {'mitigate', 'cardiovascular'}

Words before 'virus':
Year: 2008 Words: {'management', 'influenza', 'eradicate', 'hepatitis', 'dwarf', 'blue-tongue'}
Year: 2009 Words: {'detekterade', 'bluetongue', 'influenza', 'field', 'binds', 'measure', 'new', 'one'}
Year: 2010 Words: {'influenza', 'sendai', 'monocytogenes', 'mop-top', 'c', 'e'}
Year: 2011 Words: {'influenza', 'inhibit', 'usa', 'encephalitis', 'rotavirus', 'identification', 'sindbis', 'purified'}
Year: 2013 Words: {'pathogenic', 'herpes', 'schmallenberg', 'pathogens'}
Year: 2012 Words: {'fever', 'puumala'}
Year: 2014 Words: {'influenza'}
Year: 2015 Words: {'reservoirs', 'teno', '3', 'cleaning'}
Year: 2016 Words: {'simplex', 'influenza', 'dengue', 'attenuated', 'syncytial', 'fungus', 'image', 'integrating', 'fever', 'incidence', 'immunodeficiency', 'individual', 'hiv-1', 'changes', 'one'}
Year: 2017 Words: {'rna', 'influenza', 'syncytial', 'control', 'encephalitis', 'respiratory', 'latent', 'two', 'zika', 'large', 'b', 'simplex', 'bronchitis', 'animal'}
Year: 2018 Words: {'crucial', 'rsams', 'viruseswhether', 'influenza', 'cycle', 'involved', 'puumala', 'mop-top', 'zika', 'affected', 'threat', 'langat', 'regardless', 'trans-synaptic', 'bind', 'plan'}
Year: 2019 Words: {'sequencing', 'influenza', 'dengue', 'syncytial', 'limits', 'encephalitis', 'dsrna', 'fever', 'eller', 'corona', 'simplex'}
Year: 2020 Words: {'spread', 'hiis', 'invading', 'corona', 'kostvanor', 'enough', 'handling', 'syncytial', 'forest', 'papilloma', 'fever', 'information', 'capture', 'new', 'amount', 'av', 'ports', 'v'}
Year: 2021 Words: {'locations', 'attacked', 'persistent', 'fullbordad', 'recent', 'herpes', 'infected', 'c', 'corona', 'role', 'common', 'region', 'airways', 'whole', 'steps', 'syncytial', 'encephal'}
Year: 2022 Words: {'resolve', 'sensitive', 'due', 'effect', 'corona', 'transport', 'syncytial', 'encephalitis', 'papilloma', 'rwanda', 'pathway', 'proof', 'enteric', 'similar', 'infectious', 'asf', 'asf'}
Year: 2023 Words: {'inactivate', 'proteins', 'influenza', 'spread', 'mutations', 'fever', 'surfaces', 'regulate', 'dominant', '19', 'zoonotic', 'large', 'zika', 'endocytosis', 'differently'}

```

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.