horesta_text-explore-phases

May 12, 2021

1 Horesta - exploring textual content

In the following, various exploratory analysis are performed on posts from Horesta.

The data consists of all posts from the webpage last collected on March 25th 2021.

```
The data consists of 1351 posts
```

For each post the data contains the title, the URL of the post, the tags used, the publish date, the text of the post, the links in the text as well as the HTML code of the page.

A snippet of the data can be seen below:

```
                                                 url  accessed  \
0  https://horesta.dk/nyheder/2020/december/det-e…         1
1  https://horesta.dk/nyheder/2020/december/forsl…         1
2  https://horesta.dk/nyheder/2020/december/hores…         1
3  https://horesta.dk/nyheder/2020/december/webin…         1
4  https://horesta.dk/nyheder/2020/december/sidst…         1

                                               title  \
0             Det er tid til at få det lange lys på
1                           Forslag om lavere moms
2  HORESTA: Feriepenge bør følges op med et oplev…
3  Webinar med Danske Bank: Her bruger danskerne …
4  Sidste chance for finansiering - hør mere på w…

                                                tags  \
0  [coronakrise, horesta, kirsten munch, vaccinat…
1  [small danish hotels, jørgen christensen, moms…
2  [feriepenge, hjælpepakker, turisme, dansk turi…
3  [webinar, forbrug, danske bank, louise aggerst…
4          [vækstfonden, finansiering, coronakrise]

                                               links  publish_date  \
0                                                 []  04-12 - 2020
1                                                 []  04-12 - 2020
2                                                 []  03-12 - 2020
3  [/webinar-med-danske-bank, /cdn-cgi/l/email-pr…  03-12 - 2020
4  [/events/2020/december/webinar-med-vaekstfonde…  02-12 - 2020
```

```
  access_date                                                 text  \
0  2020-12-04  \n\n\n\n\n\n\n\nFra næste måned begynder udrul…
1  2020-12-04  \n\n\n\n\n\n\n\nDer er akut behov for at få st…
2  2020-12-04  \n\n\n\n\n\n\n\nHORESTA tager positivt imod, a…
3  2020-12-04  \n\n\n\n\n\n\n\nDanskernes forbrugsvaner har æ…
4  2020-12-04  \n\n\n\n\n\n\n\nHORESTA inviterer til webinar …

                                                 html
0  \n<!DOCTYPE html>\n\n<html lang="da">\n<head>\…
1  \n<!DOCTYPE html>\n\n<html lang="da">\n<head>\…
2  \n<!DOCTYPE html>\n\n<html lang="da">\n<head>\…
3  \n<!DOCTYPE html>\n\n<html lang="da">\n<head>\…
4  \n<!DOCTYPE html>\n\n<html lang="da">\n<head>\…
```

1.1 Post activity

```
The first post on Horesta is from 2018-04-20 00:00:00.
```

The data can be roughly split into a "pre-COVID" period and a "during COVID" period.

The "pre-COVID" period runs from 2018-2019.

The "during COVID" period runs from 2020-2021.

The post count in the two periods can be seen below:

```
                 Post count
Pre-COVID-19            633
During COVID-19         718
```

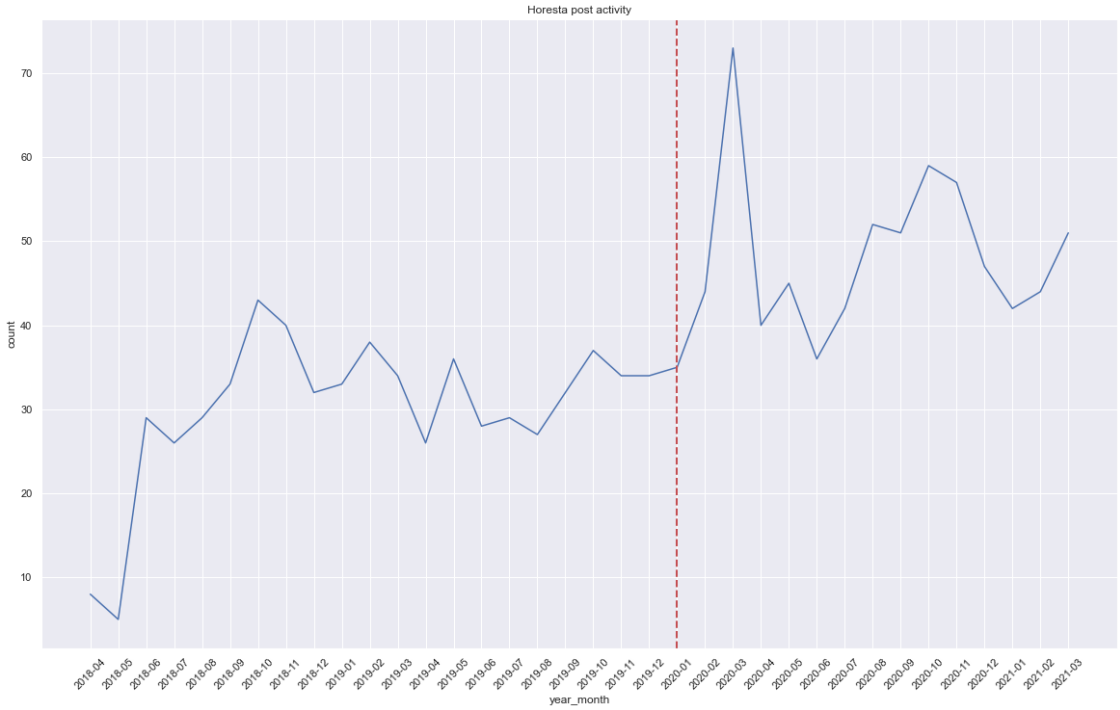
The data can be further split into different phases (see documentation). The post count for each phase can be seen below ("Phase 0" is pre-COVID):

```
         Post count
Phase 0         633
Phase 1         121
Phase 2          84
Phase 3          83
Phase 4          89
Phase 5          96
Phase 6         104
Phase 7         141
```

```
The first post mentioning coronavirus was titled "Corona-virus: Her er rådene
fra Sundhedsstyrelsen" and was published on 2020-02-05 00:00:00.
Link https://horesta.dk/nyheder/2020/februar/corona-virus-her-er-raadene-fra-
sundhedsstyrelsen/?_requestTag=
```

1.1.1 Visualizing post activity

The post activity can be visualized. The graph below shows the post count per month. The vertical line marks the division between the "pre-COVID" and "during COVID" period:



1.2 Use of tags

Each post is given a series of tags by Horesta. These indicate the topics that the post is addressing.

Combined with the date information, use of tags can be counted in each phase (see documentation) and pre-/during COVID.

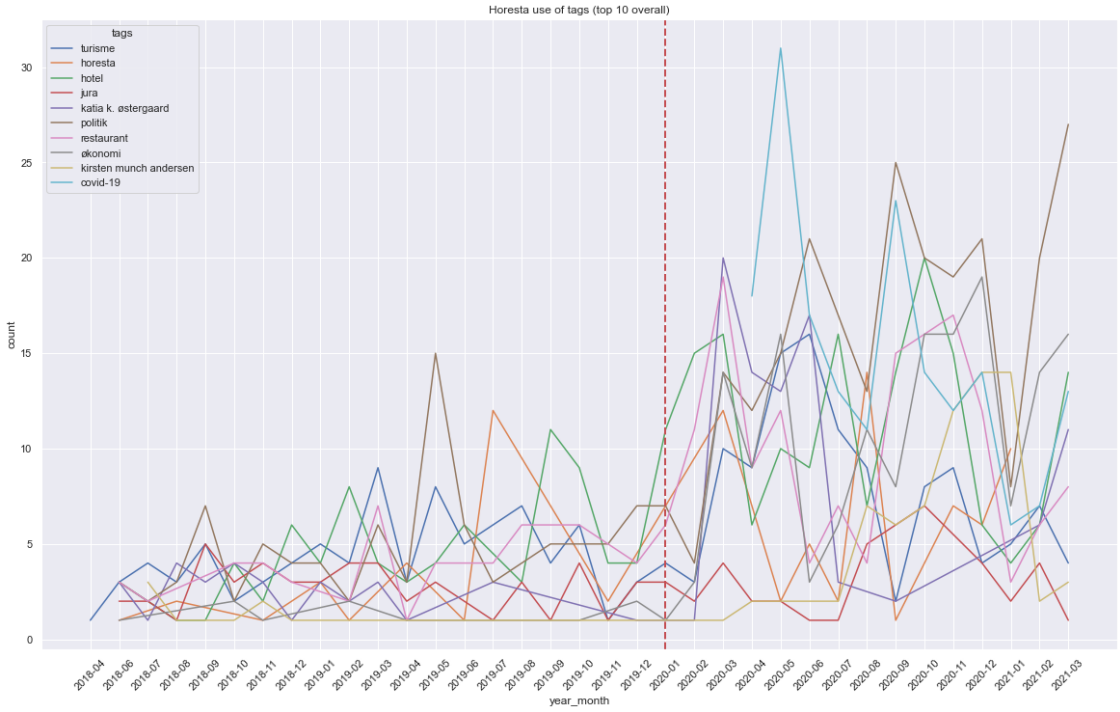
To make sure that the tag count does not just reflect the post activity, the use of tags has been calculated as a percentage for each phase.

The two enclosed Excel spreadsheets show the use of tags in each phase and the use of tags pre-/during COVID respectively.

```
C:\programs\Anaconda3\envs\lda\lib\site-packages\ipykernel_launcher.py:14:
DeprecationWarning: The default dtype for empty Series will be 'object' instead
of 'float64' in a future version. Specify a dtype explicitly to silence this
warning.
```

1.2.1 Visualizing the use of tags

In the graph below, the use of the top 10 tags overall is visualized (count per month). The vertical line marks the division between the "pre-COVID" and "during COVID" period:



1.3 Use of words

Common practice in text mining is a "keyword analysis". This usually involves identifying key terms in the text material.

The post data has been tokenized, meaning that each text is converted to a list of words, where only words with some semantic relevance is kept. In this case, only nouns and adjectives are kept.

The use of words (/tokens) is counted in each phase. To make sure that the use of words does not just reflect post activity, the "term frequency-inverse document frequency" (TF-IDF) is calculated for each word in each phase.

This metric takes into account both how often the word is used and how common the word is across texts. A word that is very common for a phase (meaning that many posts in that phase contains the word) weighs less than a word that is more unique for specific texts.

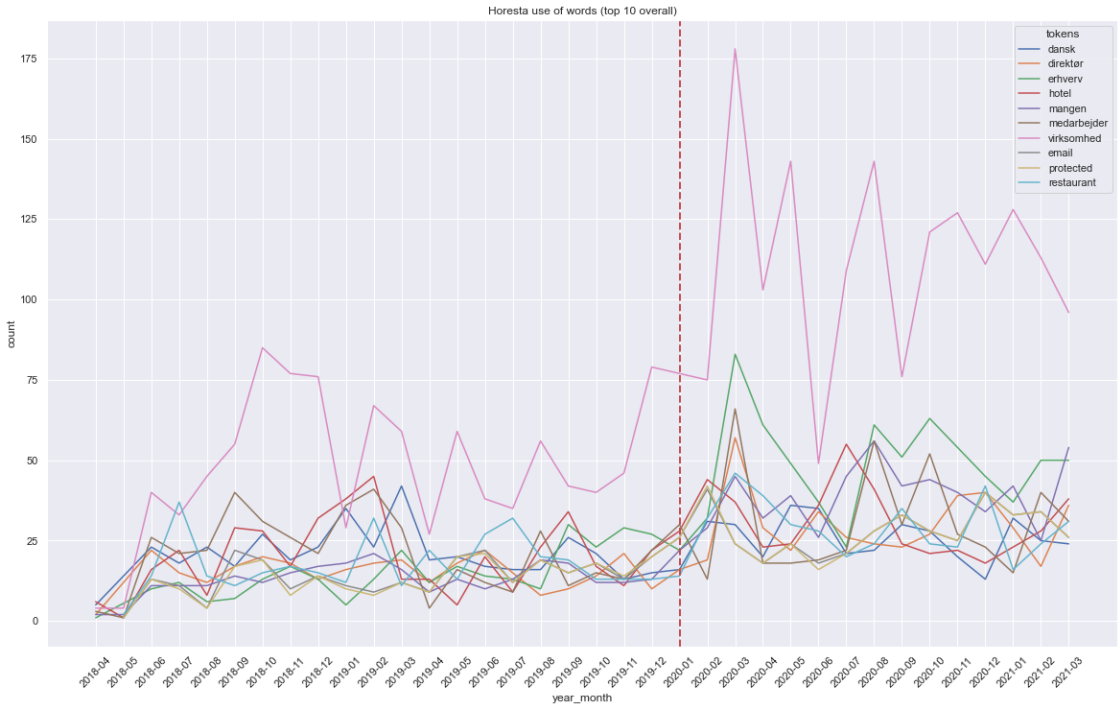
1.3.1 Top words in phases

The two enclosed Excel spreadsheets include the top 100 words overall as well as counts and tf-idf metric for each phase and for pre-/during COVID respectively.

```
C:\programs\Anaconda3\envs\lda\lib\site-packages\ipykernel_launcher.py:18:
DeprecationWarning: The default dtype for empty Series will be 'object' instead
of 'float64' in a future version. Specify a dtype explicitly to silence this
warning.
```

1.3.2 Visualizing the use of words

The graph below shows the use of the top 10 most used words (nouns and adjectives) overall and how it has developed over time (counts per month). The vertical line marks the division between the "pre-COVID" and "during COVID" period.



1.3.3 Visualizing the use of keywords

The use of a specific set of words can also be shown. The graph below shows how the use of the following keywords has developed over time:

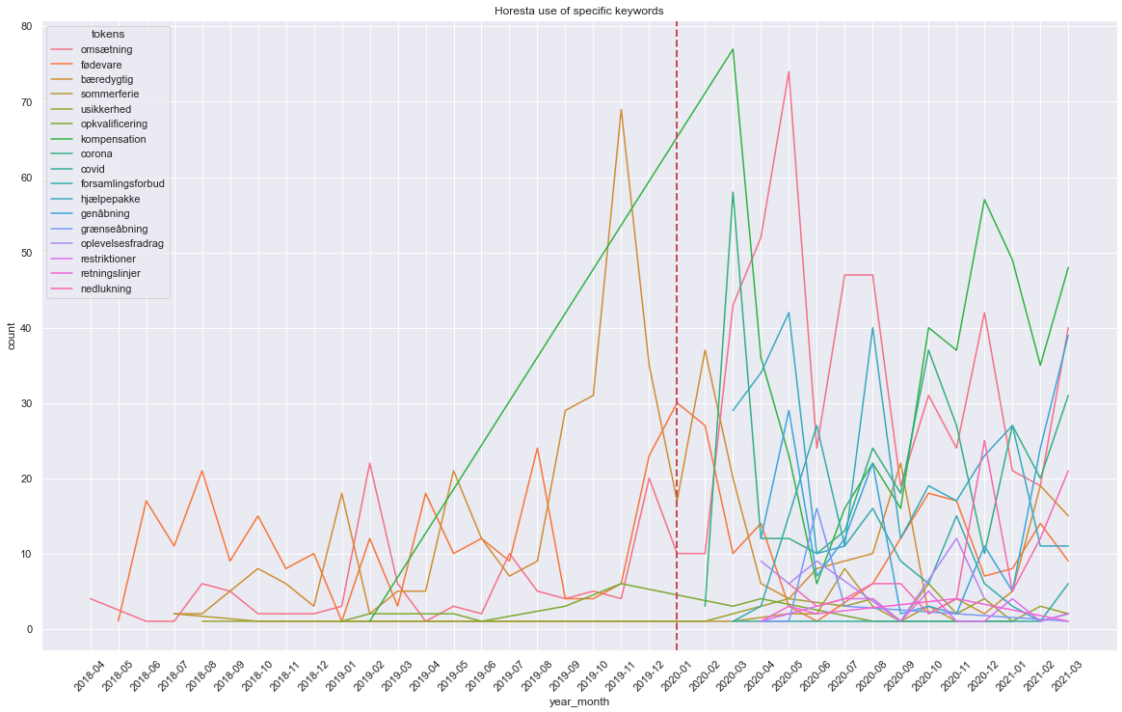
- 'bæredygtig'
- 'fødevare'
- 'corona'
- 'covid'
- 'hjælpepakke'
- 'restriktioner'
- 'kompensation'
- 'omsætning'
- 'opkvalificering'
- 'sommerferie'
- 'genåbning'
- 'oplevelsesfradrag'
- 'usikkerhed'
- 'retningslinjer'
- 'grænseåbning'

- ‘restriktioner’
- ‘nedlukning’

Tokens are stemmed so that words with the same stem like “fødevare”, “fødevarechef”, “fødevare-området”, “fødevaresikkerhed” are all counted as “fødevare”.

```
C:\programs\Anaconda3\envs\lda\lib\site-packages\pandas\core\indexing.py:1720:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_column(loc, value, pi)
```



1.4 Keywords in context as networks

One way to explore keywords in context is to visualize words co-occurring with other words as a network.

In the following, various network graphs are produced for the following keywords:

- ‘bæredygtig’
- ‘fødevare’
- ‘corona’
- ‘covid’
- ‘hjælpepakke’

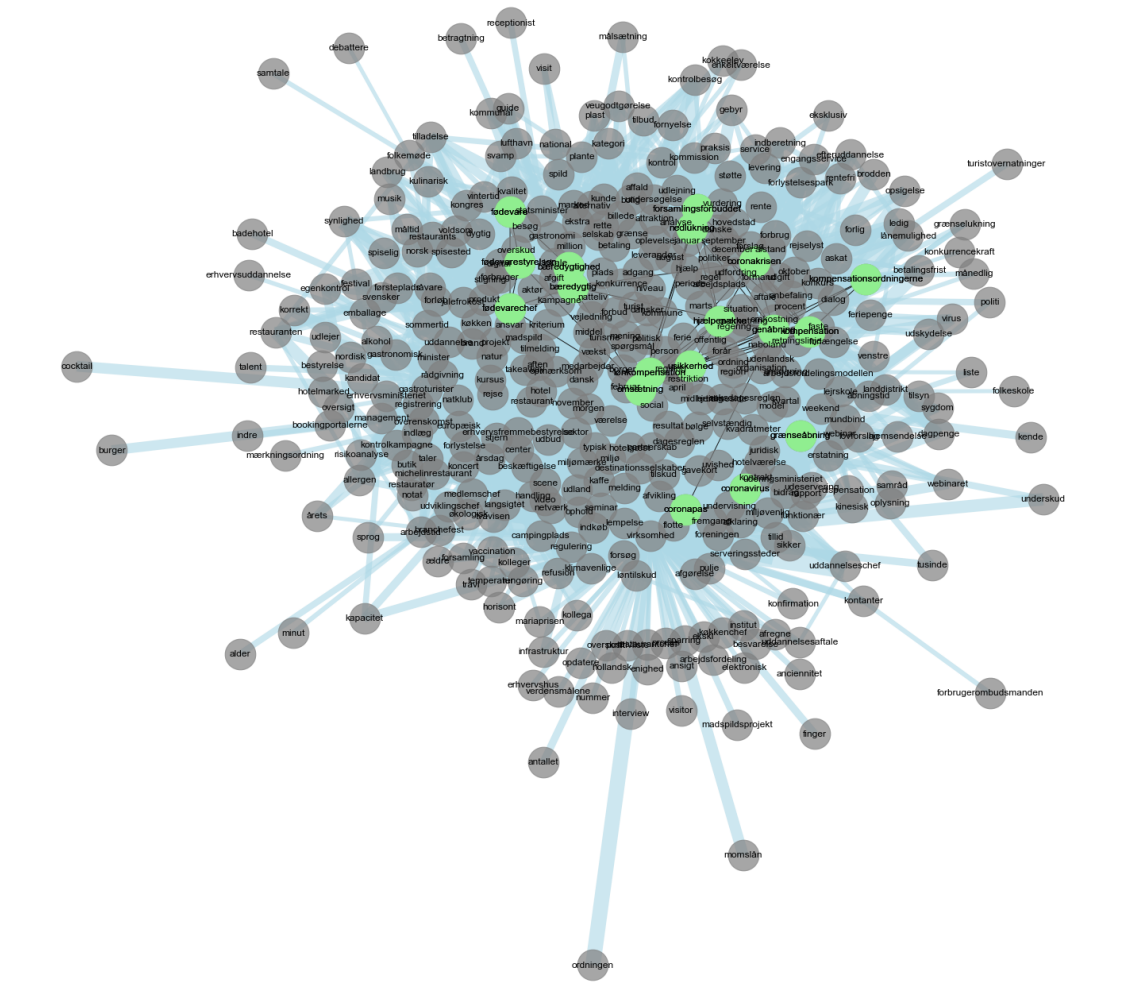
- ‘restriktioner’
- ‘kompensation’
- ‘omsætning’
- ‘opkvalificering’
- ‘sommerferie’
- ‘genåbning’
- ‘oplevelsesfradrag’
- ‘usikkerhed’
- ‘retningslinjer’
- ‘grænseåbning’
- ‘restriktioner’
- ‘nedlukning’

The graphs show which other words also occur in texts with those keywords. The thicker the line between the words, the more often they co-occur in a text.

1.4.1 Keywords in context - all data

The graph below shows keywords in context as a network across all data.

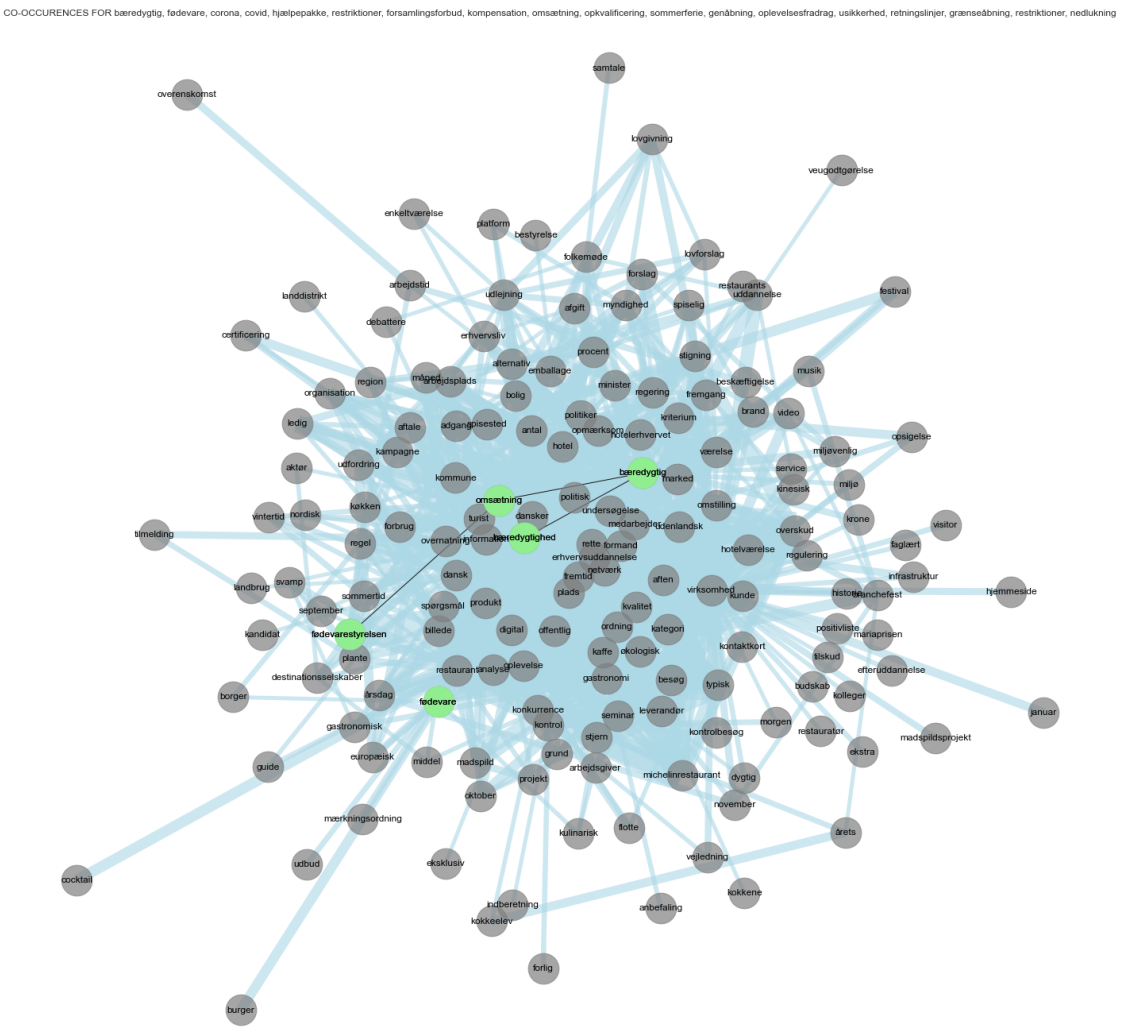
The data is filtered so that the word must occur a minimum of 10 times total and the word must have a normalized TF-IDF score above 0.4 (maximum is 1).



1.4.2 Keywords in context - pre-COVID

The graph below shows keywords in context as a network for the pre-COVID period.

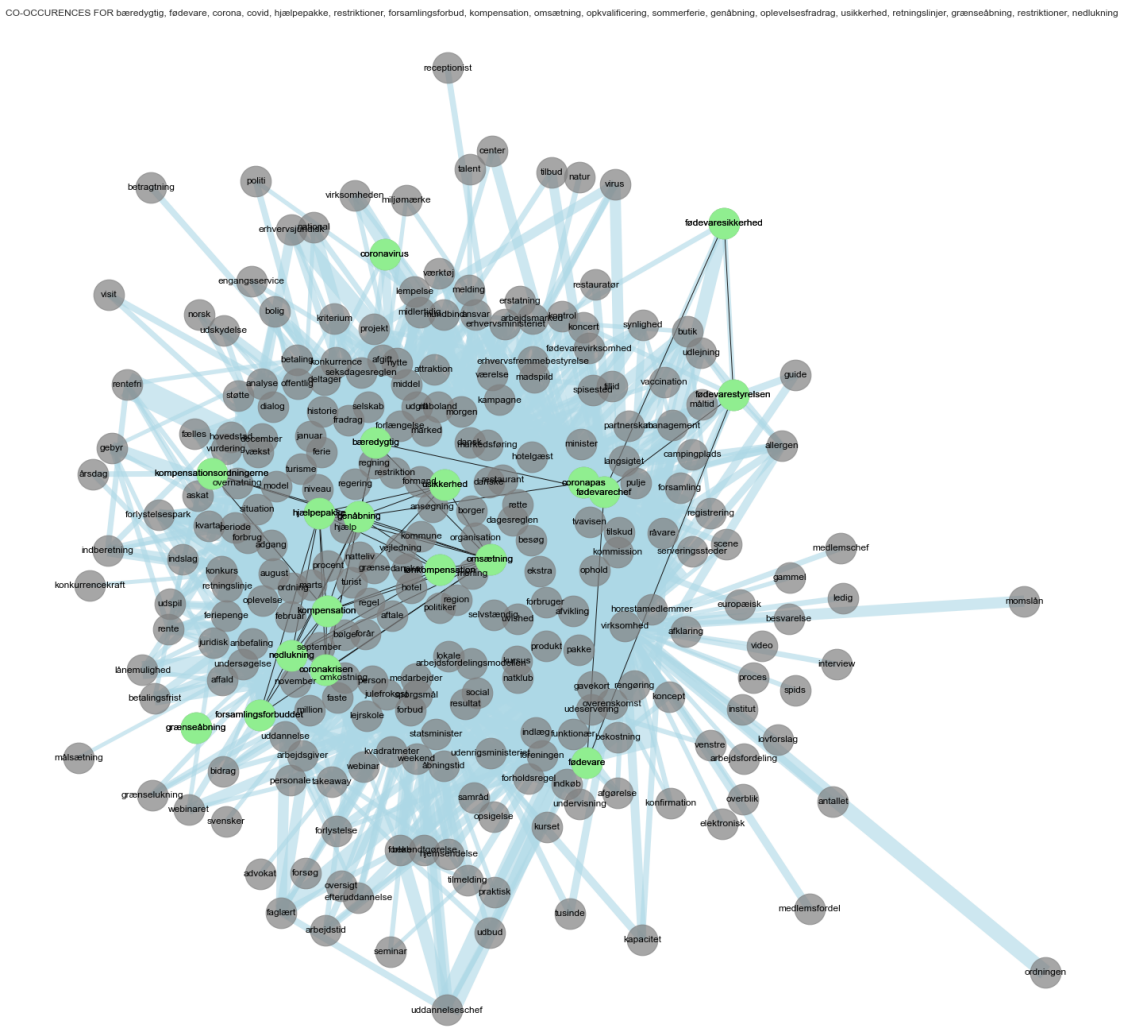
The data is filtered so that the word must occur a minimum of 10 times in the period and the word must have a normalized TF-IDF score above 0.4 (maximum is 1).



1.4.3 Keywords in context - during COVID

The graph below shows keywords in context as a network for the during COVID period.

The data is filtered so that the word must occur a minimum of 10 times in the period and the word must have a normalized TF-IDF score above 0.4 (maximum is 1).

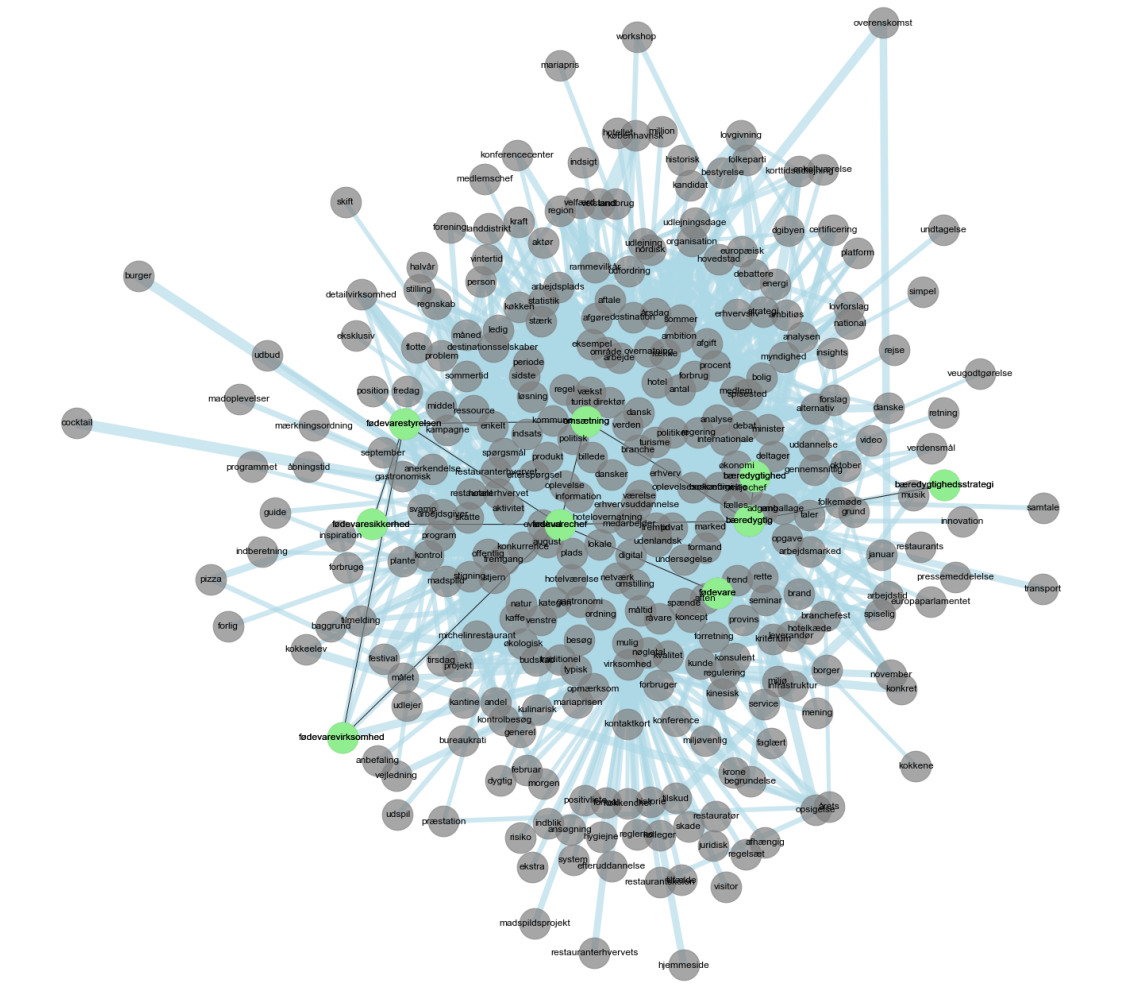


1.4.4 Keywords in context - Phases

The graph below shows keywords in context as a network for each phase (see documentation).

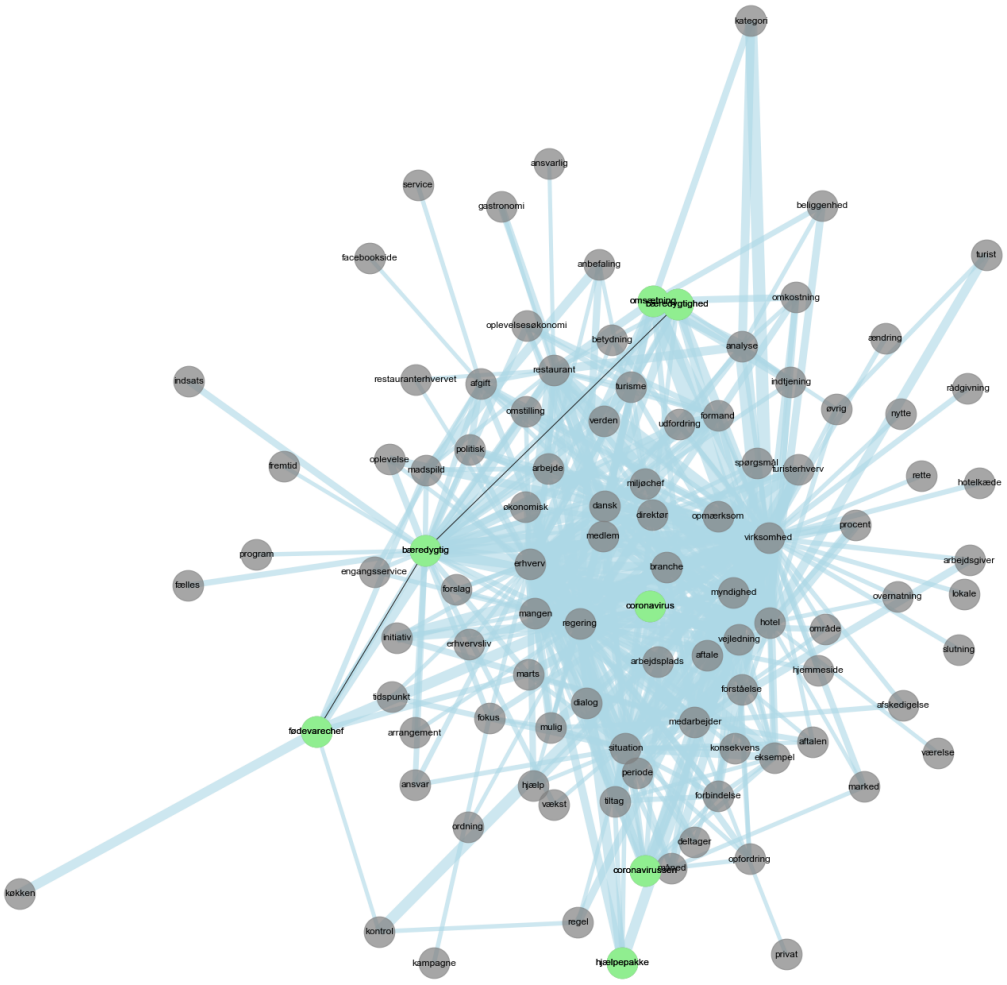
The data is filtered so that the word must occur a minimum of 10 times in the phase and the word must have a normalized TF-IDF score above 0.3 (maximum is 1) within the phase.

KEYWORDS IN CONTEXT FOR PHASE 0



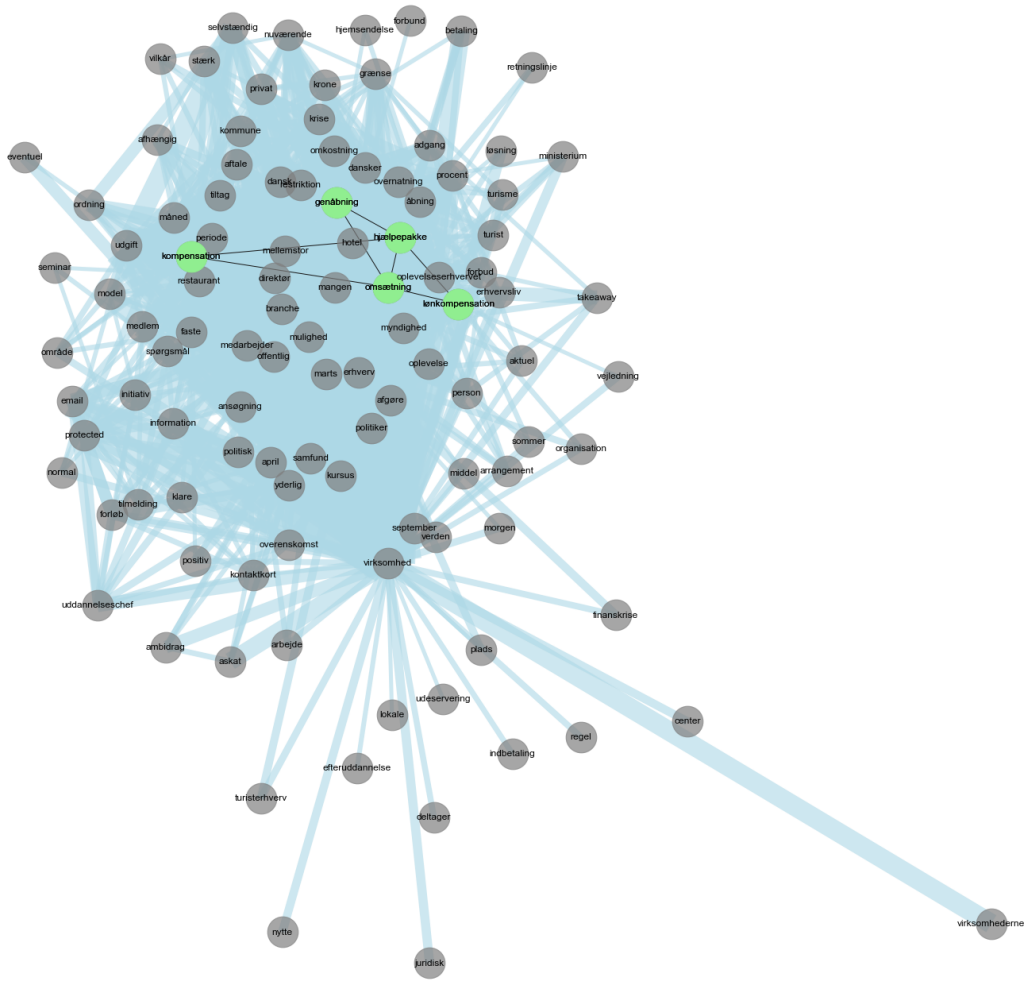
KEYWORDS IN CONTEXT FOR PHASE 1

CO-OCCURENCES FOR bæredygtig, fødevare, corona, covid, hjælpepakke, restriktioner, forsamlingsforbud, kompensation, omsætning, opkvalificering, sommerferie, genåbning, oplevelsesfradrag, usikkerhed, retningslinjer, grænseåbning, restriktioner, nedlukning

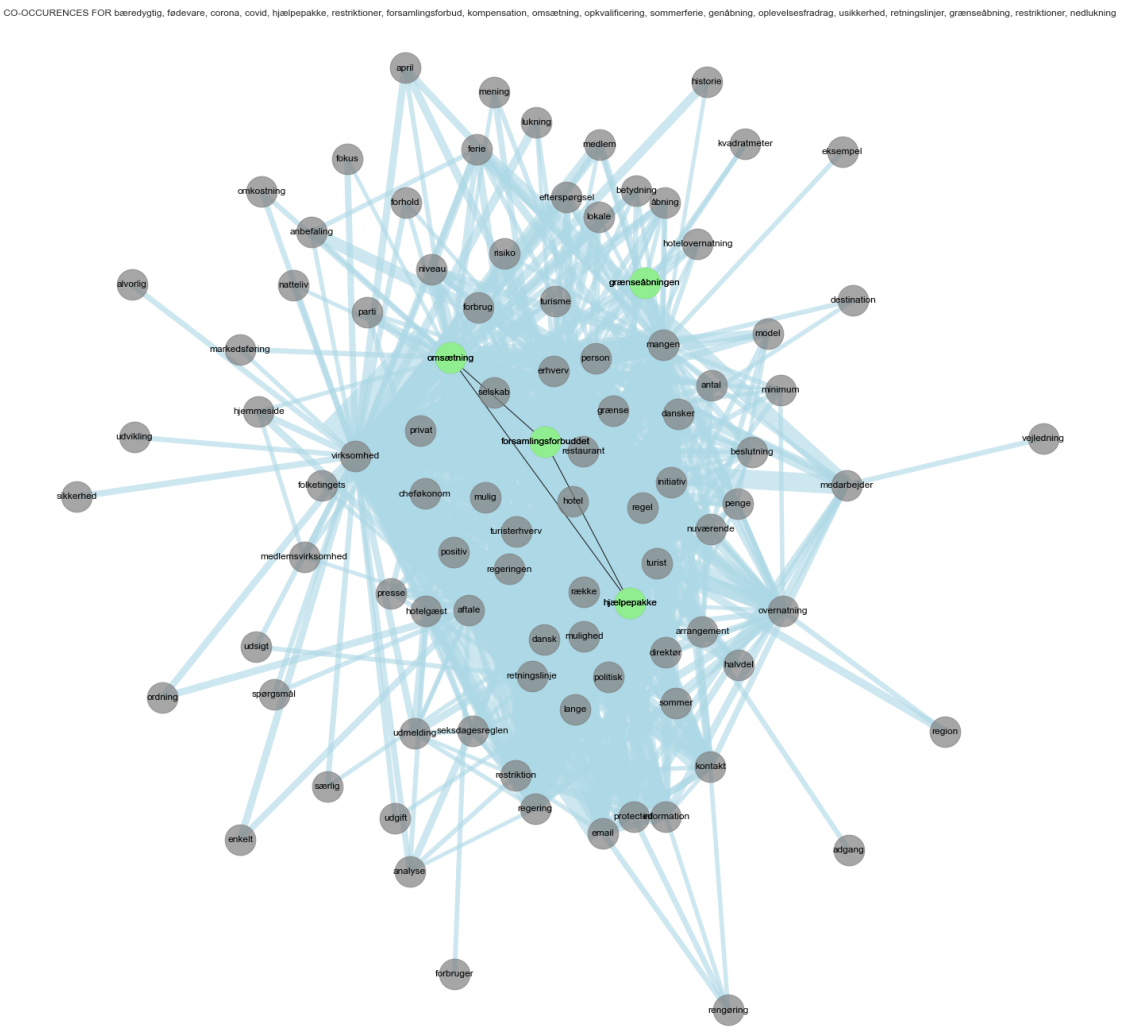


KEYWORDS IN CONTEXT FOR PHASE 2

CO-OCCURENCES FOR bæredygtig, fødevare, corona, covid, hjælpepakke, restriktioner, forsamlingsforbud, kompensation, omsætning, opkvalificering, sommerferie, genåbning, oplevelsesfradrag, usikkerhed, retningslinjer, grænseåbning, restriktioner, nedlukning

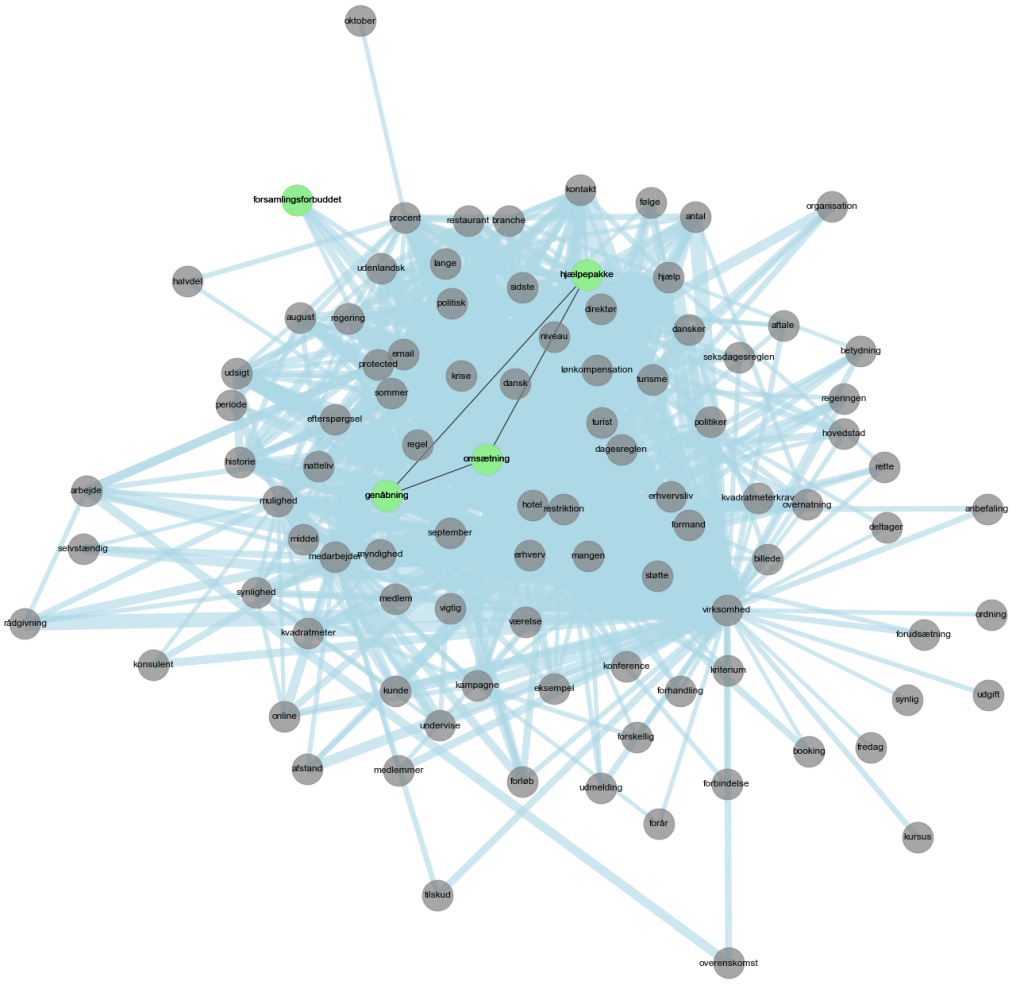


KEYWORDS IN CONTEXT FOR PHASE 3



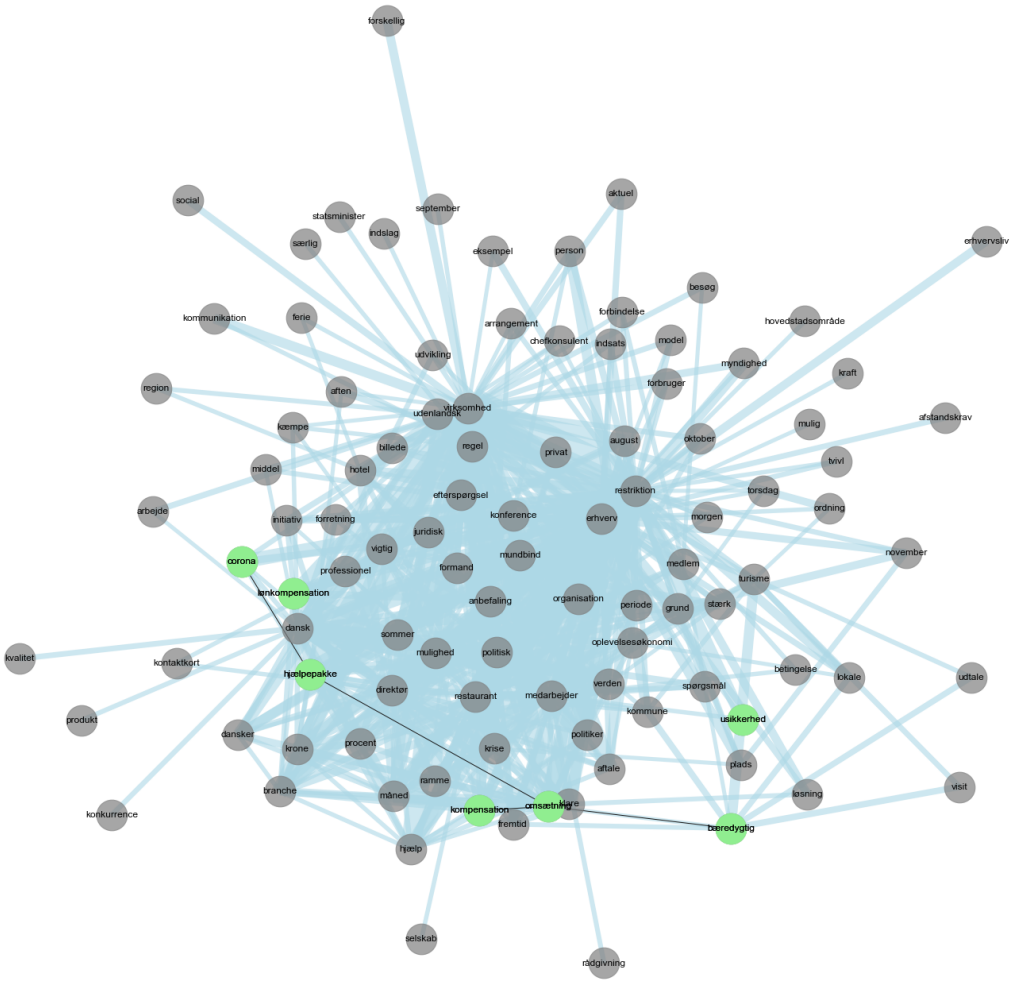
KEYWORDS IN CONTEXT FOR PHASE 4

CO-OCCURENCES FOR bæredygtig, fødevare, corona, covid, hjælpepakke, restriktioner, forsamlingsforbud, kompensation, omsætning, opkvalificering, sommerferie, genåbning, oplevelsesfradrag, usikkerhed, retningslinjer, grænseåbning, restriktioner, nedlukning



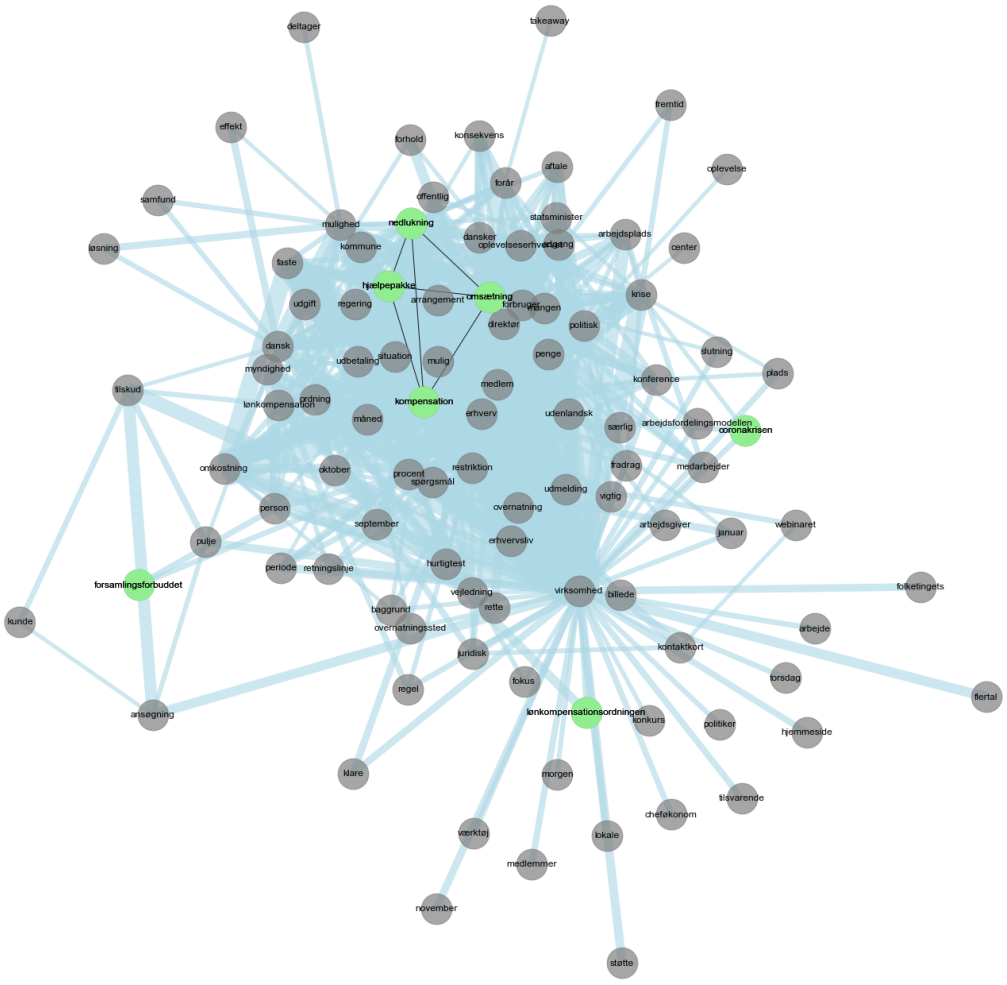
KEYWORDS IN CONTEXT FOR PHASE 5

CO-OCCURENCES FOR bæredygtig, fødevare, corona, covid, hjælpepakke, restriktioner, forsamlingsforbud, kompensation, omsætning, opkvalificering, sommerferie, genåbning, oplevelsesfradrag, usikkerhed, retningslinjer, grænseåbning, restriktioner, nedlukning



KEYWORDS IN CONTEXT FOR PHASE 6

CO-OCCURENCES FOR bæredygtig, fødevare, corona, covid, hjælpepakke, restriktioner, forsamlingsforbud, kompensation, omsætning, opkvalificering, sommerferie, genåbning, oplevelsesfradrag, usikkerhed, retningslinjer, grænseåbning, restriktioner, nedlukning



KEYWORDS IN CONTEXT FOR PHASE 7

CO-OCCURENCES FOR bæredygtig, fødevare, corona, covid, hjælpepakke, restriktioner, forsamlingsforbud, kompensation, omsætning, opkvalificering, sommerferie, genåbning, oplevelsesfradrag, usikkerhed, retningslinjer, grænseåbning, restriktioner, nedlukning