R Markdown (Thematic Analysis)

Emma Chapman-Banks

2025-01-17

Part 1: Thematic Analysis of Topics

Load the necessary packages

```
library(tidyr)
library(tidytext)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Load Overton Data

```
overton_data <- read.csv("data/overton_results_expanded.csv")
```

From inspecting the data, topics quite often include multiple keywords. Therefore, we must separate each these words into their own separate rows to ensure that proper thematic analysis can be run

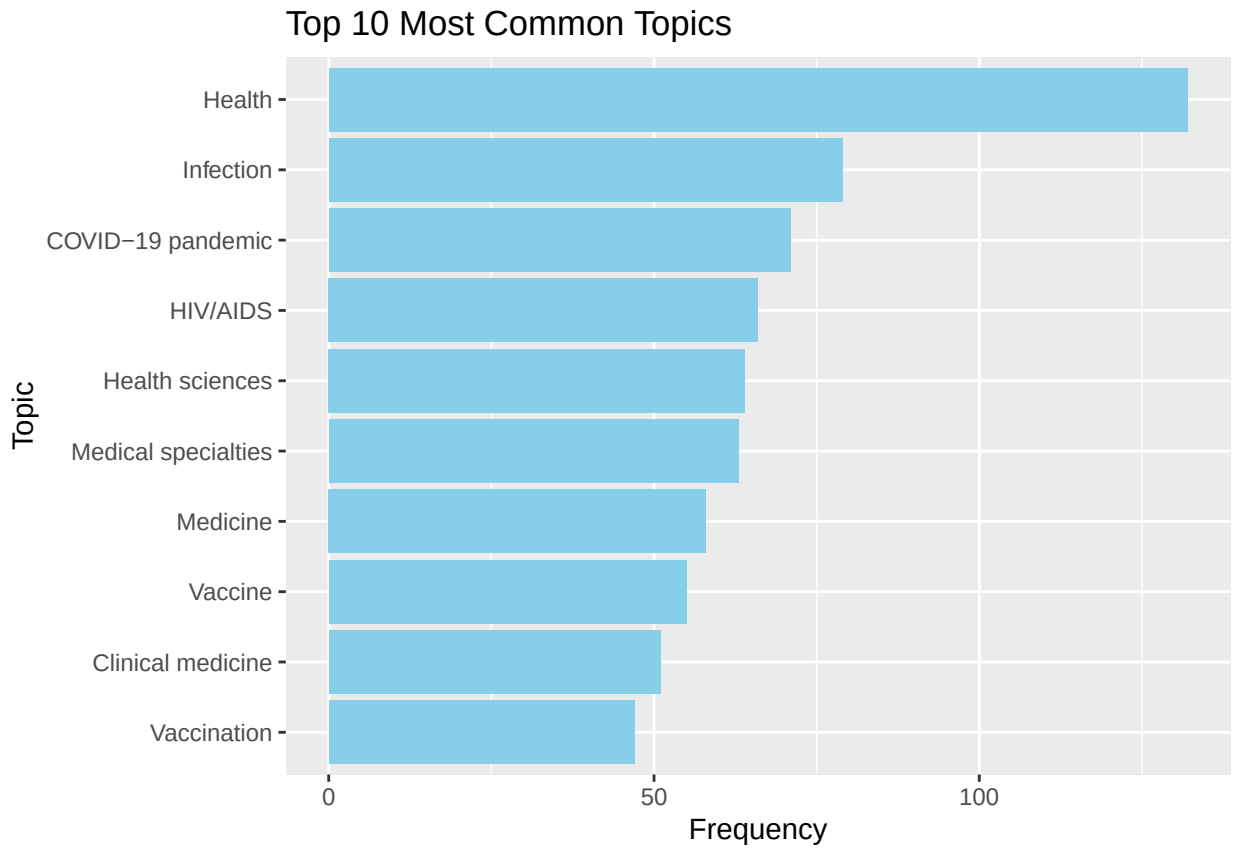
```
clean_data <- overton_data %>%
  mutate(topics = strsplit(as.character(topics), ";")) %>%
  unnest_longer(topics)%>%
  mutate(topics = trimws(topics)) %>%
  # Remove common stopwords (eg. and)
  anti_join(stop_words, by = c("topics" = "word"))
#view(clean_data) again, the data might be too large to view on a markdown
```

With over 3,000+ rows, lets find a way to count the frequency of each key word and see how often they appear.

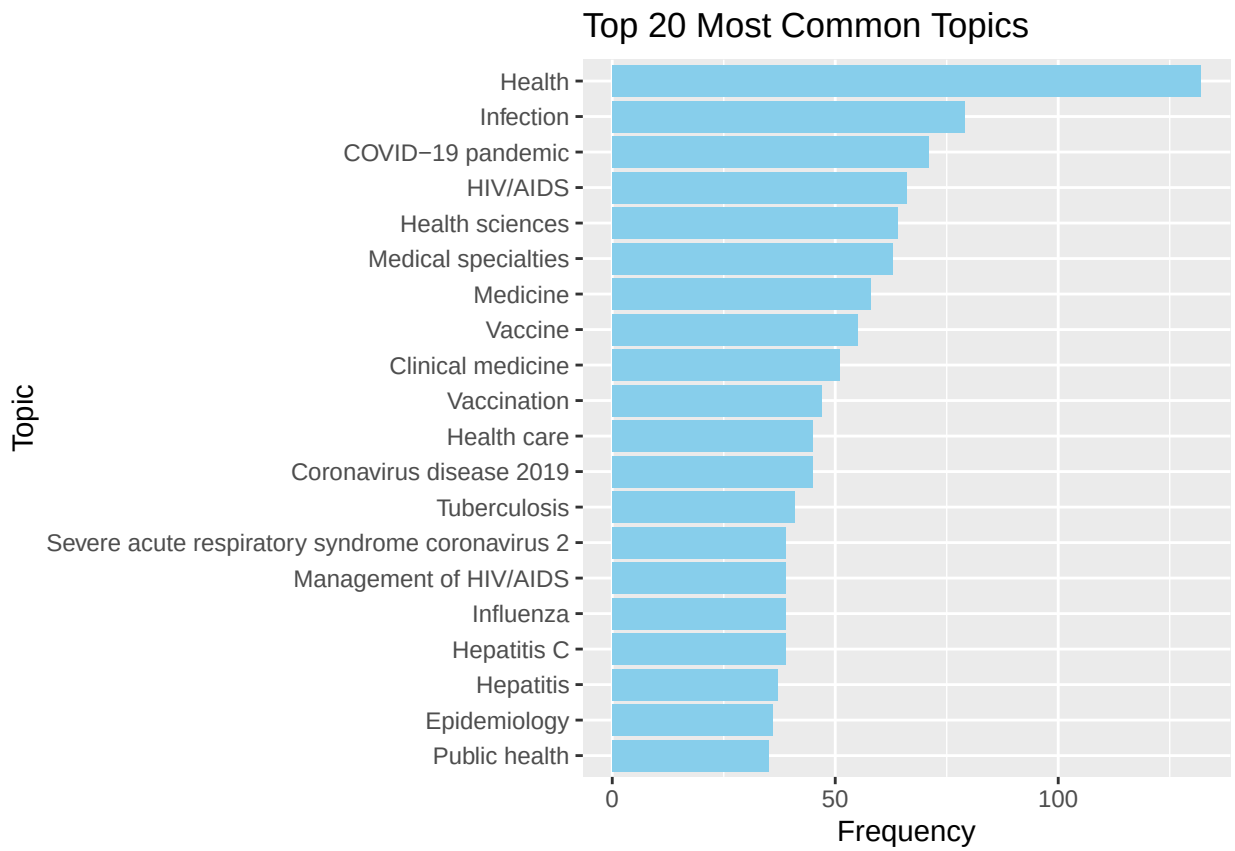
```r
frequency_counts <- clean_data %>%
  count(topics, sort = TRUE)
```

Once we have a table of frequencies, we can plot this to visualise the top 10, 20, or 30 key words appearing in these pieces of literature.

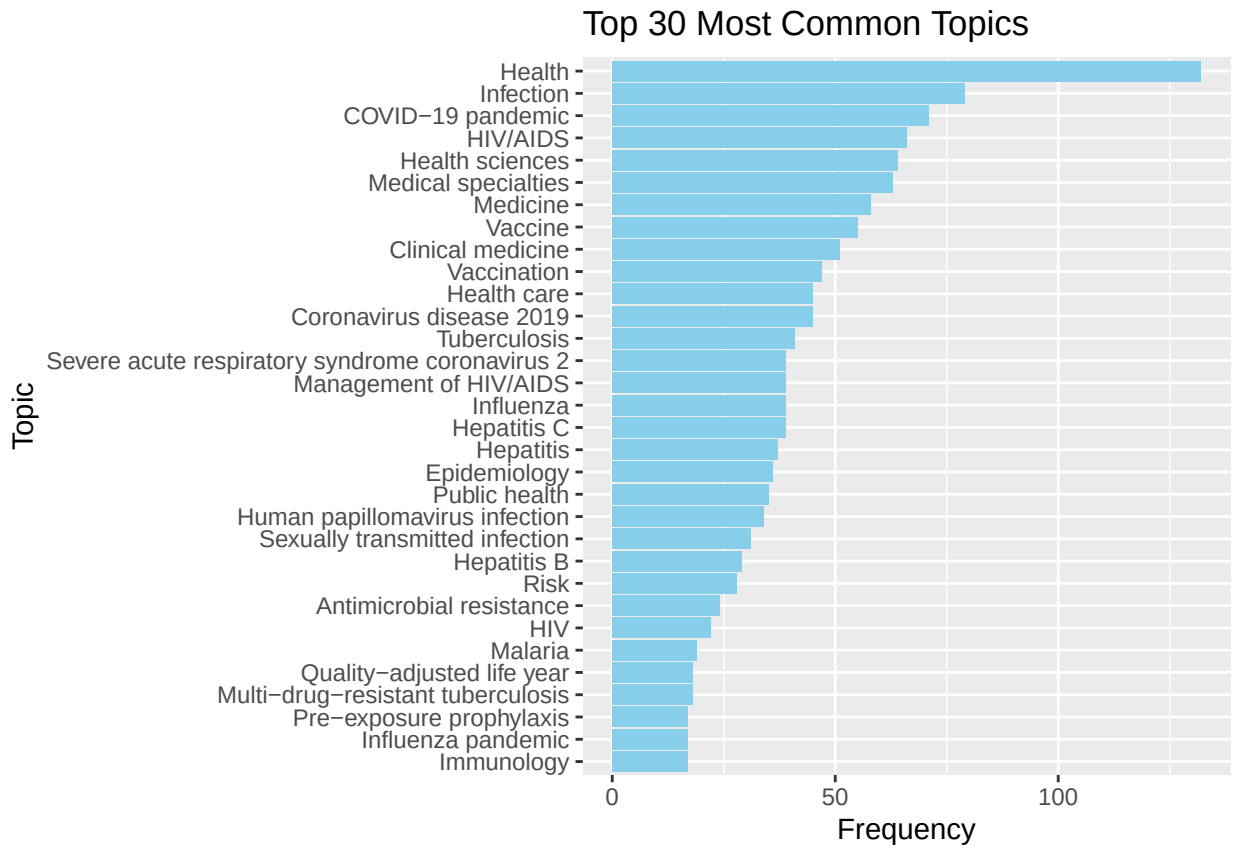
```r
#For Top 10
frequency_counts %>%
  top_n(10, n) %>%
  ggplot(aes(x = reorder(topics, n), y = n)) +
  geom_col(fill = "skyblue") +
  coord_flip() +
  labs(title = "Top 10 Most Common Topics", x = "Topic", y = "Frequency")
```




```r
#For Top 20
frequency_counts %>%
  top_n(20, n) %>%
  ggplot(aes(x = reorder(topics, n), y = n)) +
  geom_col(fill = "skyblue") +
  coord_flip() +
  labs(title = "Top 20 Most Common Topics", x = "Topic", y = "Frequency")
```




```r
#For Top 30
frequency_counts %>%
  top_n(30, n) %>%
  ggplot(aes(x = reorder(topics, n), y = n)) +
  geom_col(fill = "skyblue") +
  coord_flip() +
  labs(title = "Top 30 Most Common Topics", x = "Topic", y = "Frequency")
```



Part 2: Thematic Analysis of Classifications

Now, lets analyse the classifications column, which (from my understanding), represents hierarchical levels assigned to each document with each level providing more detail. Example: health -> disease and conditions -> communicable diseases

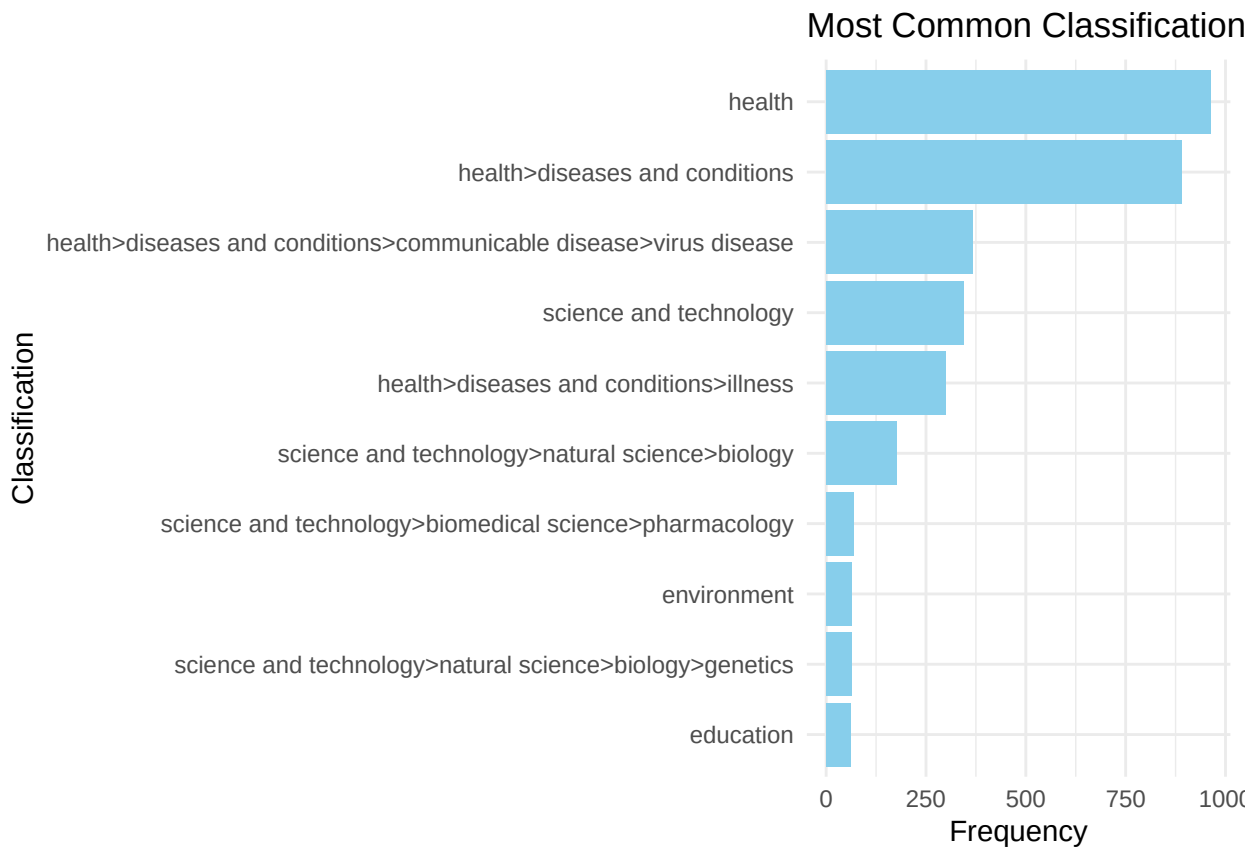
```
# Split classifications into individual categories
clean_classifications <- overton_data %>%
  mutate(classifications = strsplit(as.character(classifications), ";")) %>%
  unnest(classifications) %>%
  mutate(classifications = trimws(classifications))

# Count frequency of classifications
classification_counts <- clean_classifications %>%
  count(classifications, sort = TRUE)

# View the most frequent classifications
print(classification_counts)
```

```
## # A tibble: 90 x 2
##    classifications                                                      n
##    <chr>                                                            <int>
##  1 health                                                             964
##  2 health>diseases and conditions                                     890
##  3 health>diseases and conditions>communicable disease>virus disease  366
##  4 science and technology                                             345
##  5 health>diseases and conditions>illness                             298
```

```
##  6 science and technology>natural science>biology                    176
##  7 science and technology>biomedical science>pharmacology            69
##  8 environment                                                       65
##  9 science and technology>natural science>biology>genetics           63
## 10 education                                                         62
## # i 80 more rows
```

Let's construct a frequency table too.

```r
# Visualise the top 10 most common classifications
classification_counts %>%
  top_n(10, n) %>%  # Select top 10 by frequency
  ggplot(aes(x = reorder(classifications, n), y = n)) +  # Reorder topics for better visualization
  geom_col(fill = "skyblue") +  # Create bar chart
  coord_flip() +  # Flip coordinates for horizontal bars
  labs(
    title = "Most Common Classifications",
    x = "Classification",
    y = "Frequency"
  ) +
  theme_minimal()  # Apply a clean theme
```



Part 3: Analysing "base classifications"

Now, we will take a closer look and dissect the subcategories of each base classification. Use function extract_subcategories for each category.

```
# Extract the base classifications
base_classifications <- classification_counts %>%
  mutate(base = sapply(strsplit(as.character(classifications), ">"), `[`, 1)) %>%
  distinct(base)

# View the base classifications
print(base_classifications)
```

```
## # A tibble: 13 x 1
##    base
##    <chr>
##  1 health
##  2 science and technology
##  3 environment
##  4 education
##  5 economy, business and finance
##  6 lifestyle and leisure
##  7 weather
##  8 politics
##  9 disaster, accident and emergency incident
## 10 labour
## 11 society
## 12 crime, law and justice
## 13 conflicts, war and peace
```

```
# Function to Extract Subcategories for a Specific Base
extract_subcategories <- function(base_name, data) {
  data %>%
    filter(grepl(paste0("^", base_name, ">"), classifications)) %>%
    mutate(subcategories = gsub(paste0("^", base_name, ">"), "", classifications)) %>%
    select(subcategories)
}
```

3.1: "Health"

```
health_subcategories <- extract_subcategories("health", classification_counts)
print(health_subcategories)
```

```
## # A tibble: 16 x 1
##    subcategories
##    <chr>
##  1 diseases and conditions
##  2 diseases and conditions>communicable disease>virus disease
##  3 diseases and conditions>illness
##  4 health treatment
##  5 health treatment>medicine
##  6 diseases and conditions>communicable disease>epidemic
```

```
##  7 diseases and conditions>communicable disease
##  8 health treatment>preventative medicine>vaccines
##  9 diseases and conditions>cancer
## 10 health treatment>preventative medicine
## 11 diseases and conditions>communicable disease>virus disease>retrovirus
## 12 health treatment>diet
## 13 diseases and conditions>communicable disease>virus disease>AIDS
## 14 health treatment>medical procedure/test
## 15 diseases and conditions>injury
## 16 healthcare policy>government health care
```

3.2: "Science and Technology"

```
st_subcategories <- extract_subcategories("science and technology", classification_counts)
print(st_subcategories)
```

```
## # A tibble: 13 x 1
##    subcategories
##    <chr>
##  1 natural science>biology
##  2 biomedical science>pharmacology
##  3 natural science>biology>genetics
##  4 mathematics
##  5 social sciences
##  6 natural science>biology>physiology
##  7 natural science>physics
##  8 social sciences>economics
##  9 biomedical science
## 10 biomedical science>dentistry
## 11 natural science>chemistry
## 12 natural science>meteorology
## 13 social sciences>information science
```

3.3: "Society"

```
society_subcategories <- extract_subcategories("society", classification_counts)
print(society_subcategories)
```

```
## # A tibble: 4 x 1
##   subcategories
##   <chr>
## 1 social problem>addiction
## 2 social condition>poverty
## 3 values>death and dying>suicide
## 4 values>ethics
```

3.4: "Education"

```
education_subcategories <- extract_subcategories("education", classification_counts)
print(education_subcategories)
```

```
## # A tibble: 2 x 1
##   subcategories
##   <chr>
## 1 school
## 2 school>further education
```

3.5: "Labour"

```
labour_subcategories <- extract_subcategories("labour", classification_counts)
print(labour_subcategories)
```

```
## # A tibble: 2 x 1
##   subcategories
##   <chr>
## 1 employment
## 2 unemployment
```

3.6: "Environment"

```
environment_subcategories <- extract_subcategories("environment", classification_counts)
print(environment_subcategories)
```

```
## # A tibble: 5 x 1
##   subcategories
##   <chr>
## 1 natural resources>water
## 2 environmental pollution
## 3 nature
## 4 natural resources
## 5 climate change
```

3.7: "Economy, Business, and Finance"

```
ebf_subcategories <- extract_subcategories("economy, business and finance", classification_counts)
print(ebf_subcategories)
```

```
## # A tibble: 23 x 1
##    subcategories
##    <chr>
##  1 economic sector>computing and information technology
##  2 economy
##  3 economic sector>consumer goods>food
##  4 economic sector>transport
##  5 economic sector>energy and resource
##  6 economic sector>chemicals
##  7 economy>macro economics>inflation
##  8 economic sector>computing and information technology>software
##  9 economy>macro economics>recession
## 10 economic sector>agriculture
## # i 13 more rows
```

3.8: "Disaster, Accident and Emergency Incident"

```
dae_subcategories <- extract_subcategories("disaster, accident and emergency incident", classification_
print(st_subcategories)
```

```
## # A tibble: 13 x 1
##    subcategories
##    <chr>
##  1 natural science>biology
##  2 biomedical science>pharmacology
##  3 natural science>biology>genetics
##  4 mathematics
##  5 social sciences
##  6 natural science>biology>physiology
##  7 natural science>physics
##  8 social sciences>economics
##  9 biomedical science
## 10 biomedical science>dentistry
## 11 natural science>chemistry
## 12 natural science>meteorology
## 13 social sciences>information science
```

3.9: "Politics"

```
politics_subcategories <- extract_subcategories("politics", classification_counts)
print(politics_subcategories)
```

```
## # A tibble: 3 x 1
##   subcategories
##   <chr>
## 1 government
## 2 government policy>regulatory of industry>food and drink regulations
## 3 government>parliament
```

3.10: "Weather"

```
weather_subcategories <- extract_subcategories("weather", classification_counts)
print(weather_subcategories)
```

```
## # A tibble: 0 x 1
## # i 1 variable: subcategories <chr>
```

3.11: "Lifestyle and Leisure"

```
ll_subcategories <- extract_subcategories("lifestyle and leisure", classification_counts)
print(ll_subcategories)
```

```
## # A tibble: 3 x 1
##   subcategories
##   <chr>
## 1 lifestyle>food and drink
## 2 leisure>leisure venue>restaurant
## 3 leisure>recreational activities>fishing
```

3.12: "Conflicts, War and Peace"

```
cwp_subcategories <- extract_subcategories("conflicts, war and peace", classification_counts)
print(cwp_subcategories)
```

```
## # A tibble: 3 x 1
##   subcategories
##   <chr>
## 1 act of terror
## 2 armed conflict
## 3 armed conflict>war
```

3.13: "Crime, Law, and Justice"

```
clj_subcategories <- extract_subcategories("crime, law and justice", classification_counts)
print(clj_subcategories)
```

```
## # A tibble: 4 x 1
##   subcategories
##   <chr>
## 1 crime
## 2 crime>drug related crimes
## 3 law
## 4 law enforcement
```

Part 4: Constructing a Flow Chart

Ignore code for now. This is an example of how we can visualise the classifications and their respective categories.

```
# Install necessary packages
if (!require(igraph)) install.packages("igraph")
```

```
## Loading required package: igraph
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following object is masked from 'package:tidyr':
##
##     crossing
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```r
library(igraph)
library(dplyr)

# Example: Updated health_subcategories data
health_subcategories <- data.frame(
  subcategories = c(
    "diseases and conditions",
    "diseases and conditions>communicable disease>virus disease",
    "diseases and conditions>communicable disease>virus disease>retrovirus",
    "diseases and conditions>communicable disease>virus disease>AIDS",
    "diseases and conditions>illness",
    "health treatment",
    "health treatment>medicine",
    "health treatment>diet",
    "health treatment>medical procedure/test",
    "diseases and conditions>communicable disease>epidemic",
    "diseases and conditions>communicable disease",
    "health treatment>preventative medicine>vaccines",
    "diseases and conditions>cancer",
    "diseases and conditions>injury",
    "health treatment>preventative medicine",
    "healthcare policy>government health care"
  ),
  stringsAsFactors = FALSE
)

# Add "health" as the root node
health_subcategories <- health_subcategories %>%
  mutate(subcategories = paste("health", subcategories, sep = ">"))

# Prepare edges dynamically by splitting subcategories
edges <- strsplit(health_subcategories$subcategories, ">") %>%
  lapply(function(path) {
    if (length(path) > 1) {
      # Create parent-child pairs dynamically
      data.frame(from = head(path, -1), to = tail(path, -1), stringsAsFactors = FALSE)
    } else {
      NULL
    }
  }) %>%
  do.call(rbind, .)

# Create the graph from edges
g <- graph_from_data_frame(edges, directed = TRUE)

# Generate a tree-like layout
layout <- layout_as_tree(g, root = "health")

# Calculate custom spacing
layout[, 1] <- layout[, 1] * 10  # Increase horizontal spacing significantly
```

```r
layout[, 2] <- layout[, 2] * 5   # Increase vertical spacing significantly

# Manually jitter nodes horizontally for better separation (optional)
layout[, 1] <- jitter(layout[, 1], amount = 2)  # Add randomness to avoid overlap

# Plot the flow chart with adjusted layout
plot(
  g,
  layout = layout,            # Use the adjusted layout
  vertex.color = "lightblue",
  vertex.size = 20,
  vertex.label.cex = 0.8,     # Adjust label size
  edge.arrow.size = 0.5,      # Adjust arrow size
  main = "Flow Chart for Health Subcategories"
)
```