

TEACHING ABOUT TEXT AND TEXT ANALYSIS IN THE ERA OF GENERATIVE AI

Keyao “Kyle” Pan
Florida International University
kpan@fiu.edu

CENTRAL QUESTIONS

1. How does someone who knows the basic but have very little coding skills start with a text mining/analysis project?
 1. About me: some quantitative background (college), trained as a historian (Ph.D.)
 2. My constraints: self-taught some skills/literature but no systematic training
 1. Took FreeCodeCamp Python class but haven't gotten the certificate yet; Learned Cypher with Neo4j resources but haven't obtained the certificate yet
 2. DH classes/workshops with Constellate, at Harvard etc.
 3. What I know: basic terminologies, how to describe a goal/problem
 4. Difficulties: non-European source language, no equivalents of tools in the previous presentations available yet
2. Change: the popularization of generative AIs

ROAD MAP



Reading and Writing in the
Age of Generative AI



Enhancing Text Analysis
Services and Learning with
AI



Text Mining with the Help of
AI

EXPLAINING DIFFICULT TEXTS



Based on the passage provided, it appears that "colonial translation" refers to the process by

which a colonial system, such as the Japanese state, rewrites and represents social difference or heterogeneity in order to justify the necessity for its own rule and control. This process involves the suppression of the language and culture of the colonized, the promotion of the colonizers' own language and culture, and the use of ideologies such as historicism and ethnocentrism to justify the colonization and reinterpretation of the colonized people's history. The article suggests that this strategy of control serves to obscure the contingent and hazardous process through which capitalist society and the nation-state were produced, and operate as a form of "metaphysics" that presumes the existence of certain immutable forms.


K


Great! Now what does the author mean by "historicism" and "ethnocentrism"?




According to the passage provided, historicism is an ideology that is linked to an understanding of time in which all human societies, despite their historical differences and diversities, follow the exact same linear path of progress. Ethnocentrism, on the other hand, is an ideology linked to a spatial understanding in which people who share the same cultural origins – language, history, religion, customs – form the nucleus of the community as a specifically national community. Both of these ideologies are presented as having been instrumental in allowing the Japanese state to justify its appropriation and rule of the territory of Hokkaido, formerly inhabited by the Ainu people, by portraying the Ainu as an "uncivilized" people in need of enlightenment and assimilation by the Japanese, who were seen as a superior culture. The article suggests that these ideologies serve to obscure the contingent and hazardous process through which capitalist society and the nation-state were produced, and operate as a form of "metaphysics" that presumes the existence of certain immutable forms.


WRITING ADVICE






EditorBotGPT

By  @keyaopan

 Following bot · 1

 Share bot







Hide prompt

and double-check the answer to ensure it meets the requirements of the question. I will give you a paragraph I wrote. In order to make it a better academic paper, please list any grammatical mistakes and places that I can improve in bullet points, and any suggestions in bullet points too.

Edit bot

that can impact how issues are resolved, such as the relationship between a former colony and its former colonizer.

Here are some grammatical mistakes and places for improvement in your paragraph:

    React  Reply 



Claude APP 3:59 PM

Here is my feedback in bullet points:

•Grammar:

- Use "severing ties with a former colonizer are complex" or "the severance of ties with a former colonizer is complex" for subject-verb agreement

•Areas for improvement:

- Specify what kind of tensions (political, economic, social?) between the former colony and colonizer

- Expand on the examples of causes of tensions (reparations, natural resource exploitation, neocolonialism) with more details and examples

- Explain what "closure" means in this context and how the lack of it allows old wounds to reemerge

•Suggestions:

›Discuss how historical memory impacts the national identity and relationships with other nations

›Explain how unresolved issues from colonization can continue to shape global power dynamics between former colonies and colonizers.

›Analyze specific examples of former colonies and their struggles with decolonization to strengthen your arguments

›Discuss steps that can be taken to improve relationships and address systemic issues stemming from colonization

›You may also want to touch on new forms of colonization and imperialism to link with your topic of neocolonialism.

e" could be

nory" could

with a more

imples to

ften leads
lonizers.

disputes
id

have been
historical

ast traumas
nal

atriation of
a

d their

ENHANCING TEXT ANALYSIS LEARNING

Constellate

Classes & TutorialsBuilderHelpYour dashboard

Home / Events

Events

Webinar - Help Constellate Test the New Lab

May 1, 1-2 pm EDT. Help us test our new features in a class-like setting (and get a preview of what is coming)...

[Register](#)

Constellate Visualization

April 24 (1 week, 3 days). Help us test our new features in a class-like setting (and get a preview of what is coming)...

[Register](#)

Constellate Class - Pandas

March 27 (1 week, 3 day class). An introduction to data manipulation using Pandas...

[Past](#)

Constellate

February 21, 2023. An introduction to data manipulation using Pandas...

[Past](#)

Search or jump to...

PullsIssuesCodespacesMarketplaceExplore

xandaschofield / tapi-text-dataPublic

Watch0Fork5Starred3

forked from nkelber/constellate-python-workshop-template

<> Code

Pull requests

nkkelber/constellate-python-work...Public

Insights

main

A teaching template for creating a Constellate workshop

Jupyter Notebook5Updated on Jan 5, 2022

This branch is 16 commits ahead of nkelber:main.

xandaschofield | goofed up a bit of text fixing some err...on Jul 1, 202227

.gitignore

Create .gitignore

2 years ago

README.md

Create README.md

2 years ago

requirements.txt

Adding stuff for day 1

10 months ago

template.ipynb

Beginning populating templates for text cu...

last year

textcuration-1.ipynb

Final touches on lesson 1

10 months ago

textcuration-2.ipynb

Clearing lecture 2 outputs

10 months ago

textcuration-3.ipynb

I goofed up a bit of text fixing some errors ...

10 months ago

Readme

3 stars

0 watching

5 forks

Report repository

Releases

No releases published

ENHANCING TEXT ANALYSIS

We'll use our `defaultdict` to make a list appending each entry to the list for that author. We'll use the `random` library to grab a sample for any that are too large.

```
In [ ]: from collections import defaultdict
import random

# Using the poetry_metadata variable from
# we'll make a list for each author
metadata_by_author = defaultdict(list)
for meta_dict in poetry_metadata:
    metadata_by_author[meta_dict['author']]

# Iterate through each of the keys (authors)
# and add up to 10 poems to our filtered
max_per_author = 10
filtered_author_metadata = []
for author in metadata_by_author:
    if len(metadata_by_author[author]) > max_per_author:
        filtered_author_metadata += random.sample(metadata_by_author[author], max_per_author)
    else:
        filtered_author_metadata += metadata_by_author[author]

print("Length of original collection:", len(poetry_metadata))
print("Length of filtered collection:", len(filtered_author_metadata))
```



Claude APP 4:04 PM

Sure, here is a step by step explanation of the code:

```
max_per_author = 10
# This line sets a maximum number of 10 entries per author.

filtered_author_metadata = []
# This initializes an empty list to hold the filtered metadata.

for author in metadata_by_author:
    # This for loop iterates through each author in the metadata_by_author
    # dictionary.

    if len(metadata_by_author[author]) > 10:
        # This if check sees if there are greater than 10 entries for the current author.

        filtered_author_metadata += random.sample(metadata_by_author[author],
        max_per_author)
        # If so, it takes a random sample of 10 entries and adds it to the
        # filtered_author_metadata list.

    else:
        filtered_author_metadata += metadata_by_author[author]
        # If there are less than 10 entries for the current author, it adds all entries for
        # that author to the filtered list.
```

In summary, this code loops through metadata grouped by author, checks if any author has greater than 10 entries, and if so it takes a random sample of 10. Otherwise it keeps all entries for authors with less than 10. The end result is a list `filtered_author_metadata` with a maximum of 10 entries per author.



This code seems to be filtering a list of author metadata by selecting a maximum number of metadata items per author and randomly sampling the selected metadata items.

the code:

`max_per_author = 10`: This line assigns the value of `'10'` to the variable `max_per_author`, which represents the maximum number of metadata items to select per author.

`filtered_author_metadata = []`: This line initializes an empty list called `filtered_author_metadata`, which will store the filtered author metadata.

`for author in metadata_by_author:`: This line starts a loop that iterates through each author in the `metadata_by_author` dictionary, where `'metadata_by_author'` is presumably a dictionary with author names as keys and lists of metadata items as values.

`if len(metadata_by_author[author]) > 10:`: This line checks if the number of metadata items for the current author (accessed via `metadata_by_author[author]`) is greater than 10.

`filtered_author_metadata += random.sample(metadata_by_author[author], max_per_author)`: If the number of metadata items for the current author is greater than 10, this line takes a random sample of `max_per_author` metadata items from the list `metadata_by_author[author]` using the `random.sample` function, and appends them to the `filtered_author_metadata` list.

`else:` `filtered_author_metadata += metadata_by_author[author]`: If the number of metadata items for the current author is less than or equal to 10, this line simply appends the entire list of metadata items for that author to the `filtered_author_metadata` list, since no filtering is needed.

Finally, `filtered_author_metadata` will contain the filtered author metadata, ensuring that no author has more than `max_per_author` metadata items per author, with excess items being randomly selected if there were more than `max_per_author` items.

ENHANCING TEXT ANALYSIS

Programming Historian



EDITED BY
Adam Crymble

PUBLISHED | 2013-06-22 MODIFIED

Understanding Regular Expressions

Let's go further

Replace `^(\d+)`

with `$1$2$3$`

and repeat as

Now similarly

Replace `^(\d+)`

with `$1$2$3$`

K

make the following report into table:

STATISTICAL REPORTS OF MORBIDITY AND MORTALITY, STATES AND CITIES OF THE UNITED STATES UNTABULATED.

Arizona. — Quarter ended June 30, 1907. Estimated population, 122,931. Total number of deaths 292, including diphtheria 1, enteric fever 4, scarlet fever 11, smallpox 2, and 49 from tuberculosis.

Quarter ended September 30, 1907. Total number of deaths 402, including diphtheria 4, enteric fever 12, scarlet fever 2, smallpox 1.

K

for this table, the "total number of deaths" are consisted of deaths from different diseases. These diseases can be found after the word "including" in the original text. After "total number of deaths", add columns that represent each of the diseases and get their number from the original text



Location	Time Period	Estimated Population	Total Number of Deaths	Diphtheria	Enteric Fever	Scarlet Fever	Smallpox	Tuberculosis
Arizona	June 30, 1907	122,931	292	1	4	11	2	49
Arizona	Sept. 30, 1907	N/A	402	4	12	2	1	73

TEXT MINING AND CORPUS BUILDING

1. Corpus building (with copyright/user agreement issues in mind): resources through the library, corpus builders in previous presentations, OCR, scraping...
2. APIs: e.g., governmental archives, congress/parliament records
 1. My tentative research: Japanese Diet records
 2. Topic: the use of the term “human rights” in the parliament during the Occupation Period (1945-52)

The image shows two overlapping web pages. The background page is the 'Congress.gov API' page, which includes the 'Congress.gov' logo, a description of the API, and a list of endpoints such as 'bill', 'committee', and 'member'. The foreground page is the '国会会議録検索システム 検索用APIの仕様' (Japanese Diet Records Search System Search API Specification) page. It features a navigation menu with links like 'トップ', 'ヘルプ', and '検索用APIの仕様'. The main content area includes a table of contents with sections for '1. 検索用APIとは' (What is the Search API?) and '2. 概要' (Overview). The '1. 検索用APIとは' section explains that the system is an API for searching and retrieving data from the Japanese Diet Records Search System. The '2. 概要' section provides details about the search process, including the use of HTTP GET requests and the format of the returned data (XML or JSON).

Congress.gov API

Congress.gov shares its application programming interface (API) with the public to ingest the Congressional data. [Sign up for an API key](#) from api.data.gov that you can use to access web services provided by Congress.gov. To learn more, view our [GitHub repository](#).

bill Returns bill data from the

GET /bill Returns

GET /bill/{con Returns

GET /bill/{con Returns

GET /bill/{con Returns

Authorize

国会会議録検索システム 検索用APIの仕様

[トップ](#) → [ヘルプ \(使い方ガイド\)](#) → [検索用APIの仕様](#)
[シンプル表示トップ](#) → [ヘルプ \(使い方ガイド\)](#) → [検索用APIの仕様](#)

目次

- 1. [検索用APIとは](#)
- 2. [概要](#)
- 3. [利用例](#)
- 4. [利用条件・免責事項](#)
- 5. [詳細仕様](#)

1. 検索用APIとは

国会会議録検索システムに登録されているデータを検索し、取得するための外部提供インターフェイス（API : Application Programming Interface）です。

[目次へ戻る](#)

2. 概要

- 国会会議録検索システム（ウェブサイト）での検索と同等の検索、返戻機能を有しています。
- HTTPのGetメソッドで送信された検索リクエストに対し、XML形式又はJSON形式でデータを返戻します。
- 次の3種類の検索APIがあります。検索リクエストの指定方法は同じですが、データの返戻形式が異なります。

- (1) 会議単位簡易出力では、指定した検索条件でヒットした会議録の情報（回次、院、会議名、号、開催日、ID、URL等）を、1リクエストに対し最大100件まで、XML形式又はJSON形式で返戻します。
 - アクセスURLは https://kokkai.ndl.go.jp/api/meeting_list?{検索条件} です。
 - 発言を対象に検索した場合には、会議録中の該当する発言の情報（発言者名、発言順、ID、URL）も合わせて返戻します。
 - 本文のテキストデータは返戻しません。

NO/LOW CODE TEXT-MINING WITH AI ASSISTANCE



According to the API documentation given here (<https://kokkai.ndl.go.jp/api.html>), write a python code that extracts (over 100) records with the search condition 'any=人権&from=1945-08-14&until=1952-04-28&recordPacking=json'



Sure, here is the Python code from=1945-08-14&until=1952-

Python

```
import requests

# Set the search conditions
search_conditions = {
    'any': '人権',
    'from': '1945-08-14',
    'until': '1952-04-28',
    'recordPacking': 'json'
}

# Make the API request
response = requests.get('https://kokkai.ndl.go.jp/api.html?any=人権&from=1945-08-14&until=1952-04-28&recordPacking=json')
```



the documentation says that one request will return a maximum of 100 records. How do I change the code to retrieve more than 100 records?



Sure, here is how you can change the code to retrieve more than 100 records:

Python

```
import requests

# Set the search conditions
search_conditions = {
    'any': '人権',
    'from': '1945-08-14',
    'until': '1952-04-28',
    'recordPacking': 'json'
}
```



terminal returns the error "line 21, in <module>
total_records = records['total']
KeyError: 'total'"



You are correct, the API does not return a `total` key. Here is the corrected code:

Python

```
import requests
```

TEXT ANALYSIS TOOL: KH CODER

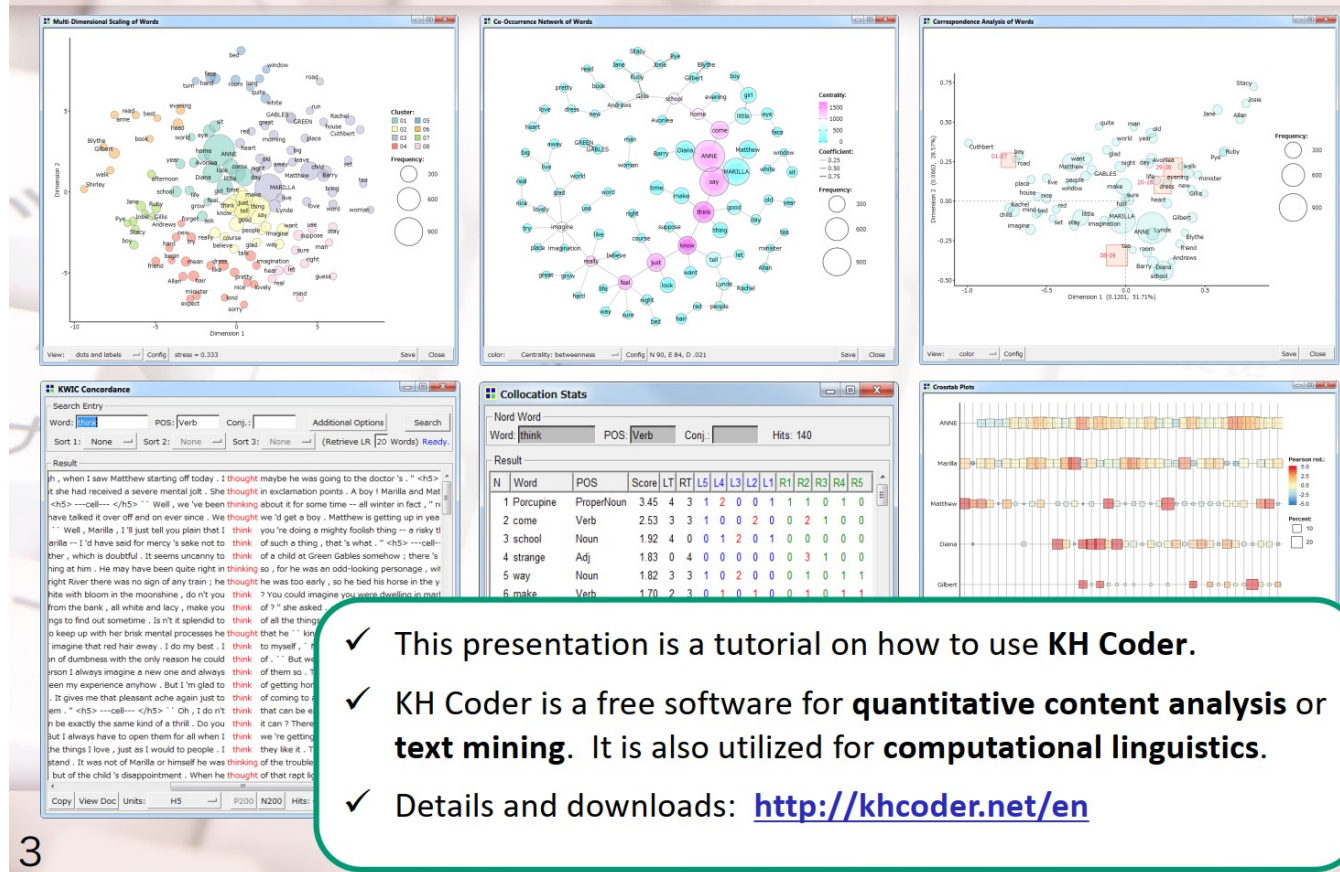
Built-in Japanese text segmentation library and morphological parser (MeCab and Chasen)

Built-in solutions for Japanese encoding problems/garbled texts

Also no code (except for customization of visualization in R Studio)

(Similar to Voyant but developed primarily for Japanese)

Preface



TEXT PREPROCESSING AND DATA CLEANING

K

how do I do the following in excel: if a cell has text with more than 4000 characters to another row; copy all other cells to the new row; automatically



You can use Excel's built-in functions to split the text in a cell with more than 4000 characters to another row. Here's how you can do it:

1. Open your Excel file and select the sheet containing the data.
2. Click on the "Developer" tab in the ribbon, and then click on "Visual Basic editor."
3. In the editor, click on "Insert" and then select "Module" to insert a new module.
4. In the new module, paste the following code:

mathematica

```
Sub SplitLongText()  
    Dim lastRow As Long  
    Dim i As Long  
    Dim j As Long  
    Dim cell As Range  
    Dim cellValue As String  
    Dim splitValue As Variant
```

K

this does split the cells with more than 4000 characters but it does not copy the text in all other cells from the old row to the new split row. could you modify the code?



Apologies for the mistake in my previous response. You are correct that the previous code did not copy the other cells from the old row to the new split row. Here's the updated code that copies all the other cells from the old row to the new split row:

vbnet

Copy code

```
Sub SplitLongText()  
    Dim lastRow As Long  
    Dim i As Long  
    Dim j As Long  
    Dim cell As Range  
    Dim cellValue As String  
    Dim splitValue As Variant
```

K

can you rewrite it to copy columns B to P



Sure, here's the updated code that will copy columns B to P to the new split row:

vbnet

Copy code

```
Sub SplitLongText()  
    Dim lastRow As Long  
    Dim i As Long  
    Dim j As Long  
    Dim cell As Range  
    Dim cellValue As String
```

Functions

Detailed info on problems: [Print](#) [Save as](#)

Close

TEXT ANALYSIS: RESULT INTERPRETATION

Word Frequency List

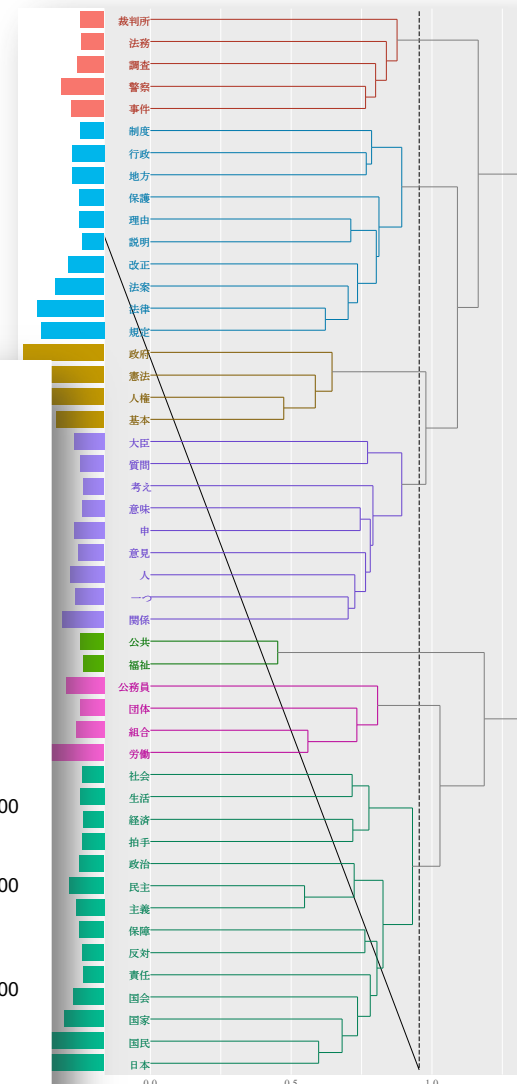
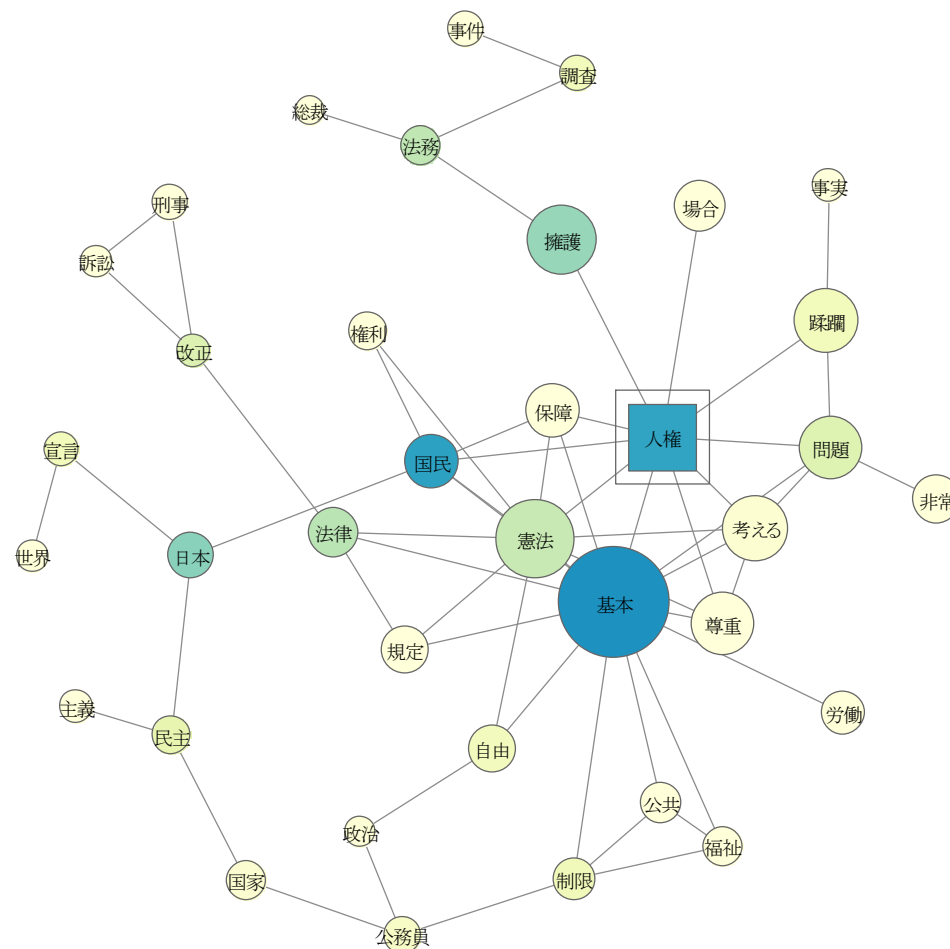
Filter Entry

OR Partial Match Filter

List

#	Word	POS / Conj.	Frequency
1	問題	ナイ形容	7295
2	考える	動詞	6649
3	人権	名詞	5911
4	政府	名詞	5681
5	日本	地名	5538
6	労働	サ変名詞	4979
7	場合	副詞可能	4863
8	法律	名詞	4679
9	憲法	名詞	4528
10	規定	サ変名詞	4432
11	国民	名詞	4089
12	非常	形容動詞	3777
13	法案	名詞	3424
14	基本	名詞	3342
15	警察	名詞	3048
16	持つ	動詞	2965
17	関係	サ変名詞	2940

Bar chart Enable Filter when Search P100 N100



FUTURE POSSIBILITIES

- Next: Using AI to help analyze and interpret graphs/text analysis results?
- Other possibilities: character AI, improved authorship studies, fine-tuned models for specific projects, etc.
- The role of the researcher and teacher of text engagement?
 - As interface of knowledge, along with AI?
 - Not to “replace”, but to do things heretofore not achieved yet

GPT-4



What are a couple of meals I can make with

Based on the items visible in the fridge, here are a couple of meal ideas:

- 1) Yogurt parfait: Layer yogurt, strawberries, and honey or nuts for added texture.
- 2) Carrot and hummus wrap: Spread hummus on a wrap, add shredded carrots and any other desired vegetables.

