Research on Modern Japanese Literature Using Full-Text Data

HIBI Yoshitaka

How to use information sources for Japanese studies

NDL Webinar 2024

Self-Introduction

日比 嘉高 HIBI Yoshitaka

Professor, Graduate School of Humanities, Nagoya University

Modern Japanese Literature and Culture

■『プライヴァシーの誕生　モデル小説のトラブル史』新曜社2020

■「科研費採択課題を対象とした研究課題の計量テキスト分析――日本文学の場合」『社会文学』52、2020、pp.89-100

■「人と機械の境を跨ぐ――芭蕉受容のデジタル・ヒューマニティーズ的研究のメモ」『跨境　日本語文学研究』15、2022、pp.12-19

ほか

About This Webinar

3 webinar topics

1. The status of full-text text data from the perspective of researchers of modern Japanese literature
2. Using the services NDL Next Digital Library and NDL Ngram Viewer
3. Practical examples of literary research using full-text text data

1 The status of full-text text data from the perspective of researchers of modern Japanese literature

New!!

Can be searched

- Aozora Bunko
- National Diet Library Digital Collection
- Google Books
- Maruzen eBook Library (*)
- Kino Den (*)
- Databases of various newspapers (*)

Can analyze data

- Aozora Bunko (**)
- National Diet Library Next Generation Digital Library (**)
- NDL Ngram Viewer provided by National Diet Library NDL Lab

(*) requires a subscription

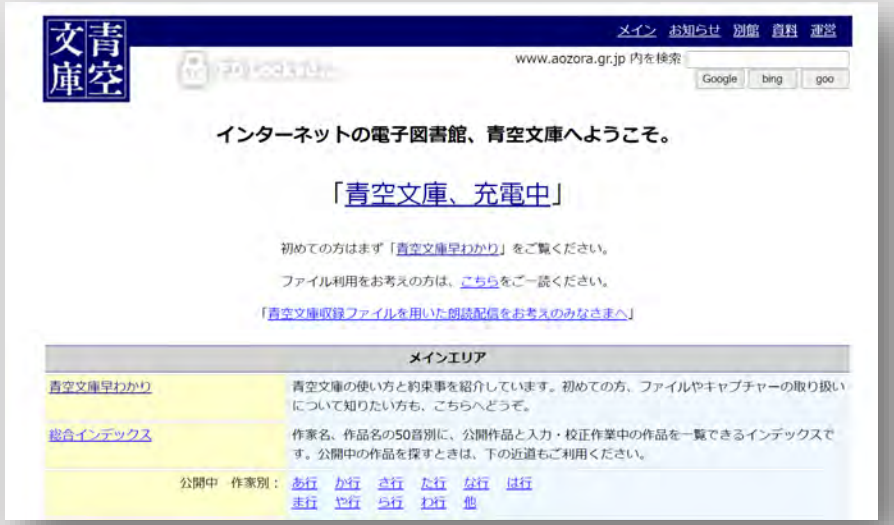
(**) by downloading the full text

Representative full-text search services

Search the text of literary works

Search for related keywords

Discover works that share keywords



Aozora Bunko

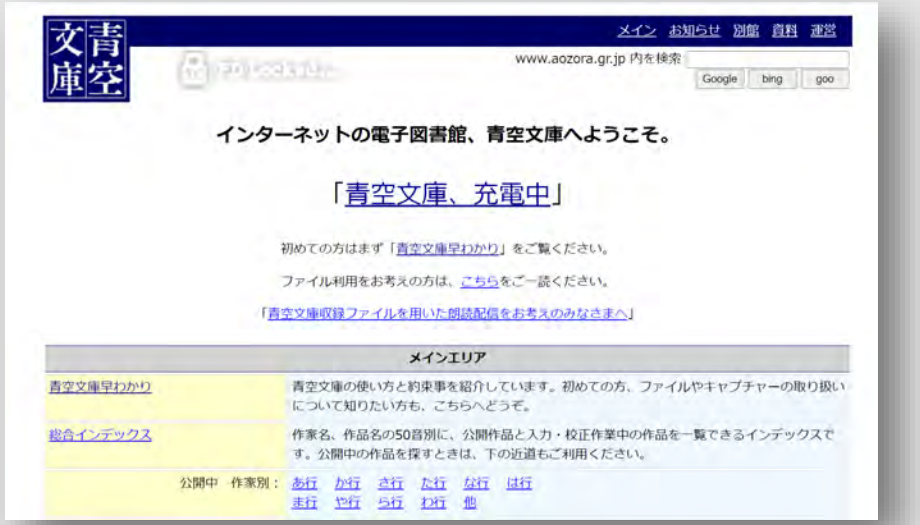
NDL Digital Collections

Sites where full texts are available

More advanced text analysis possible

Possibilities expand depending on your ideas

Analysis services that can be used without any specialized knowledge are also available



Aozora Bunko

NDL Next Digital Library

Aozora Bunko “Book Card” page

ファイルのダウンロード

ファイル種別	圧縮	ファイル名（リンク）	文字集合／符号化方式	サイズ	初登録日	最終更新日
テキストファイル(ルビあり)	zip	1504_ruby_6153.zip	JIS X 0208／ShiftJIS	218160	2001-05-24	2010-11-02
エキスパンドブックファイル	なし	1504.ebk	JIS X 0208／ShiftJIS	654364	2001-05-24	2002-01-30
XHTMLファイル	なし	1504_14585.html	JIS X 0208／ShiftJIS	726627	2004-02-11	2010-11-02

●ファイルのダウンロード方法・解凍方法

Download from here

Next Digital Library Individual document viewing screen

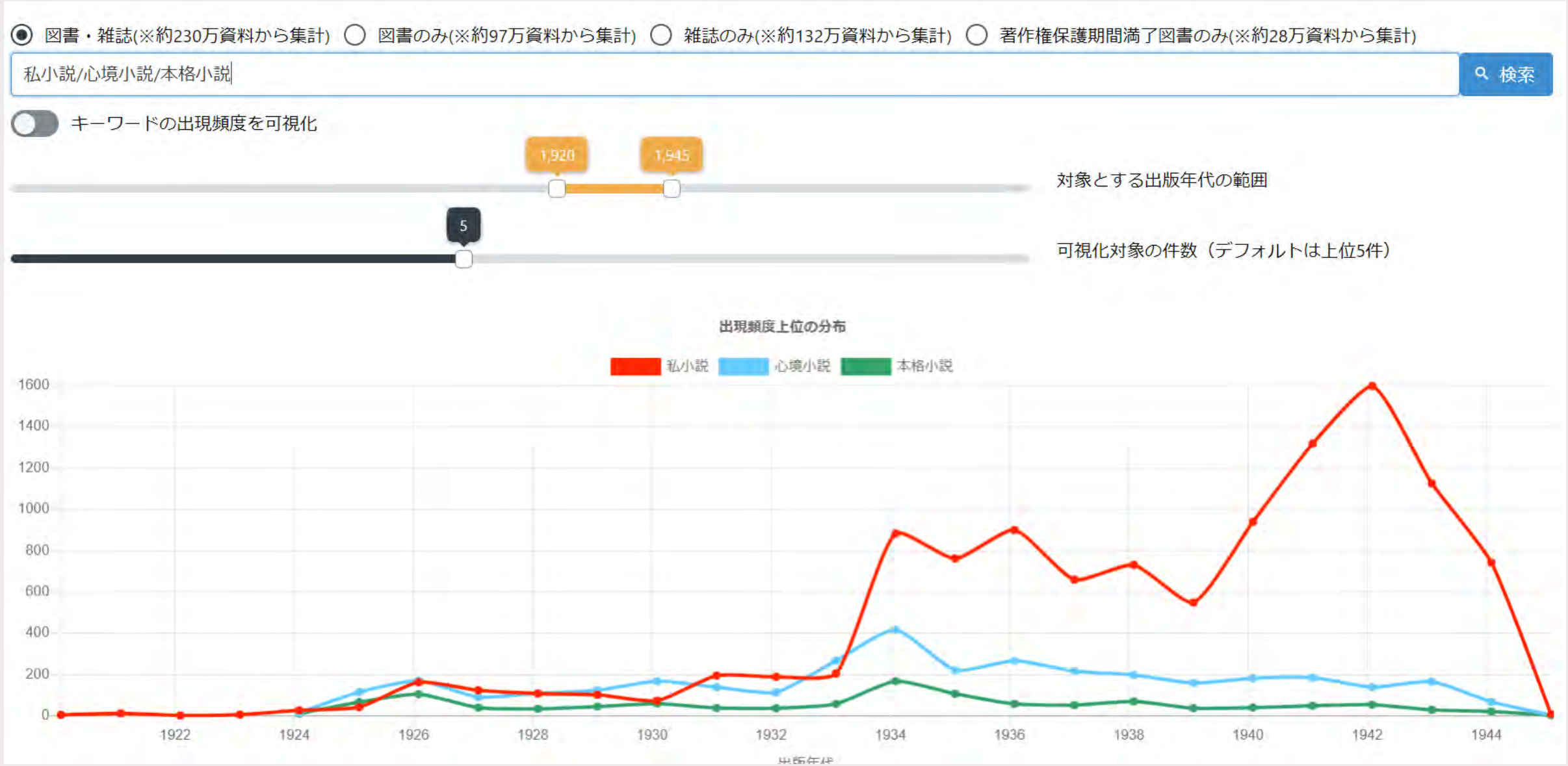
2

Try using the service –
National Diet Library's NDL Ngram Viewer

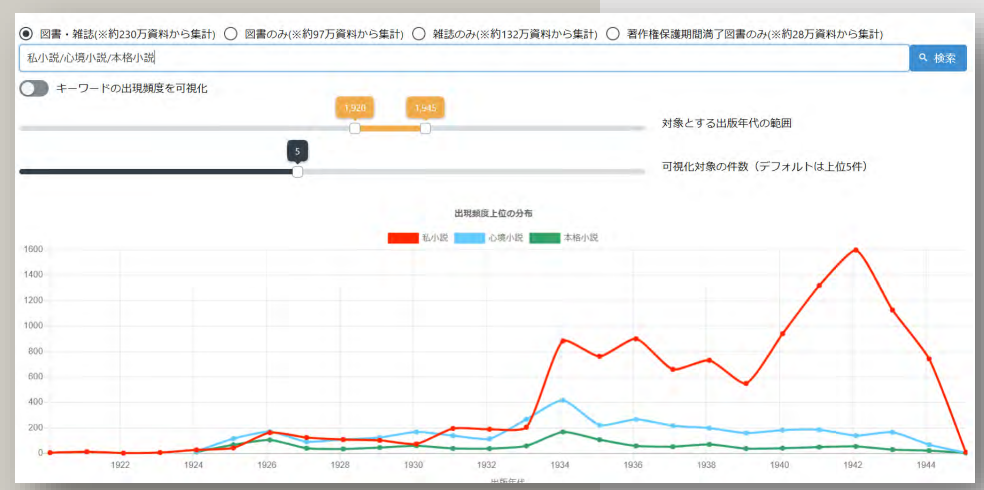
Pursuing the concept of "I-novel"

NDL Ngram Viewer

Graph of the year in which “shishosetsu” “shinkyo-shosetsu,” and “honkaku -shosetsu” appeared



What can be read from the graph



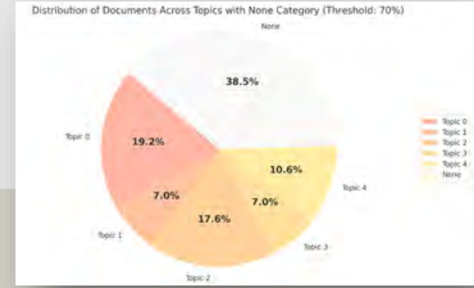
1. The frequency of use of the word “shishosetsu." (However, correction based on the number of materials is also required)
2. The size of the wave = the spread of the discussion, and you can see the difference in scale.
3. Can get an idea of the first appearance of a word. The first `shinkyo shosetsu' were published in 1924, and the first `honkaku shosetsu' were published around 1925.
4. Can compare adjacent vocabulary and observe their replacement and decline. The number started to increase from 1925, but the terms `shinkyo shosetsu' and `honkaku shosetsu' were used more frequently. From the late 1930s onwards, the frequency of use of `shinkyo shosetsu' and `honkaku shosetsu' decreased. In particular, the term `honkaku shosetsu' is becoming less commonly used.

3

Example of literary research using full-text data:
In the case of NDL data

Example ❶ Topic analysis

- Machinery determination of subgenres of literary works using computers
- Discovering unknown subgenres
- Possibility of new literary historical narratives through observation of this alternation

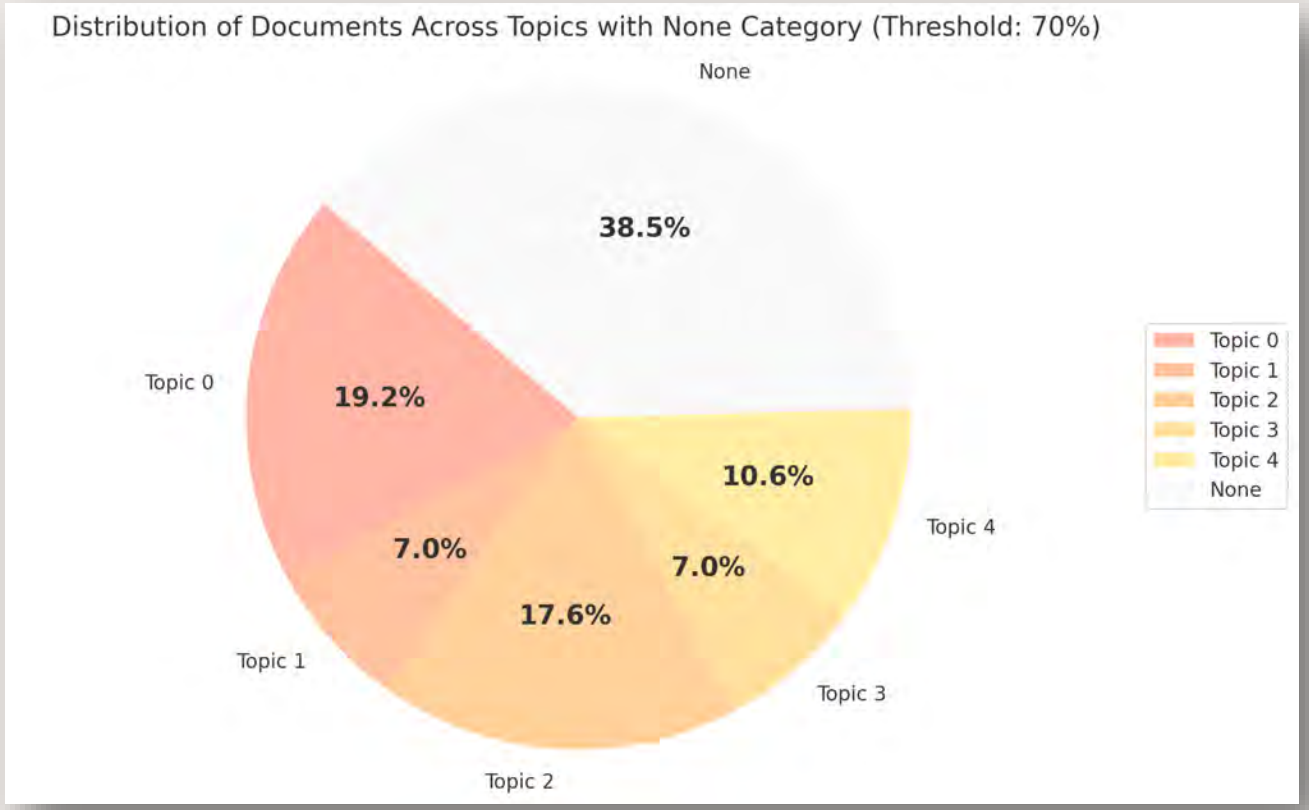


Example ❷ Reception analysis

- Reception research using text matching and similarity judgment
- Investigations on a scale that cannot be compared to humans

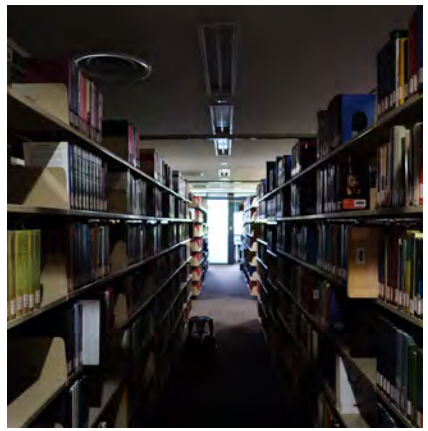
芭蕉全集の句	明治句集の句	類似度
石山の石より白し秋の風	石山の石より白し秋の風	1
道ばたの木槿は馬に喰はれけり	君が家の木槿は馬に喰はれずや	0.438
猪もともに吹かるゝ野分かな	鶏の横に吹かるゝ野分かな	0.357
猪もともに吹かるゝ野分かな	干網に蜻蛉吹かるゝ野分かな	0.357
雲雀なく中の拍子や雉の聲	雲雀なく中の松陰神社かな	0.333
鶯の老を啼くなり茶木畠	鶯の老いを啼くなり花楮	0.333
松杉の尾の上の鐘や秋の暮	鳴り止みし野寺の鐘や秋の暮	0.333
秋の色糖味噌壺もなかりけり	秋の雛笏も冠もなかりけり	0.333
蜻蛉やとりつきかねし草の上	蜻蛉や初汐来る草の上	0.333
このあたり目に見ゆるもの皆涼し	庵の月目に見ゆるもの露涼し	0.313
鶯の笠落したる椿かな	鶯の鳴く時落つる椿かな	0.308
鶯の老を啼くなり茶木畠	鶯の老を啼きけり雲の中	0.308
春の夜や籠人ゆかし堂の隅	春の夜や衣桁に重き戀衣	0.286
春の夜や籠人ゆかし堂の隅	春の夜や芝居打出す川向ふ	0.286
春の夜や籠人ゆかし堂の隅	春の夜や掛物古く桃赤し	0.286
春なれや名もなき山の朝霞	秋風や名もなき山を吹き渡る	0.286
ほとゝぎす啼くや黒戸の濱庇	ほとゝぎす月の桂を塒かも	0.286
岩躑躅染むる泪やほとゝぎす	三軒家蚊帳釣る時やほとゝぎす	0.286
岩躑躅染むる泪やほとゝぎす	鉱毒の貧乏村やほとゝぎす	0.286
岩躑躅染むる泪やほとゝぎす	峡中の暁闇やほとゝぎす	0.286
身にしみて大根からし秋の風	身にしみて市人恋し山住居	0.286
菊の香や奈良には古き佛達	菊の香や傾けつくす大瓶子	0.286
菊の香や奈良には古き佛達	菊の香や我大君は歌聖	0.286
菊の香や奈良には古き佛達	菊の香や一閑張の古机	0.286
菊の香や奈良は幾世の男振	菊の香や傾けつくす大瓶子	0.286
菊の香や奈良は幾世の男振	菊の香や我大君は歌聖	0.286
菊の香や奈良は幾世の男振	菊の香や一閑張の古机	0.286
正月も美濃と近江や閏月	橋一つ美濃と近江や別れ霜	0.273
春の夜や籠人ゆかし堂の隅	春の夜や旅寝淋しき銀屏風	0.267
春の夜や籠人ゆかし堂の隅	春の夜や髣髴として月の富士	0.267

芭蕉全集の句	大正百家選の句	類似度
松島や水を衣裳に夏の月	加茂川や水を枕に夏の月	0.385
雲の峰幾つ崩れて月の山	雲の峰崩れて雄々し不二の山	0.357
梅が香に昔の一字哀れ也	梅が香に編むや大正百	0.286
むく起に隣の花の匂ひかな	日の本に普き花の匂ひかな	0.286
行末は誰肌觸れん紅の花	行末は誰が手折るらん辻の花	0.286
清瀧の波に洗ふや夏の月	誰もかも笑顔揃ふや夏の月	0.286
月代や膝に手を置く宵の中	自ら膝に手を置く牡丹哉	0.286
蘭の香や蝶の翼にたきものす	蘭の香や源語ひもとく文机	0.286
影待や菊の香のする豆腐串	枕頭に菊の香のする夜明哉	0.286
菊の香や奈良には古き佛達	菊の香や貴船祭りの行戻り	0.286
菊の香や奈良には古き佛達	菊の香や国歌を奏す大広間	0.286
菊の香や奈良には古き佛達	菊の香や全土に及ぶ君	0.286
菊の香や奈良には古き佛達	菊の香や旭輝く日本国	0.286
菊の香や奈良には古き佛達	菊の香や得難き物は人	0.286
菊の香や奈良は幾世の男振	菊の香や貴船祭りの行戻り	0.286
菊の香や奈良は幾世の男振	菊の香や国歌を奏す大広間	0.286
菊の香や奈良は幾世の男振	菊の香や全土に及ぶ君	0.286
菊の香や奈良は幾世の男振	菊の香や旭輝く日本国	0.286
菊の香や奈良は幾世の男振	菊の香や得難き物は人	0.286
菊の香にくらがり上る節句哉	菊の香に集る夜会姿かな	0.286
茶の花に人里ちかき山路哉	茶の花に昼の月あり、詩仙堂	0.286
春雨や簑ふきかへす川柳	春雨や又かけかへす蓄音機	0.273
名月や我と筆架の影法師	名月やさて様々の影法師	0.273
春の夜や籠人ゆかし堂の隅	春の夜やよしあし語る浪華節	0.267
春の夜や籠人ゆかし堂の隅	春の夜やからかって居る電話口	0.267
蝶の羽の幾度越ゆる塀の屋根	蝶の羽の重きもそれか別れ霜	0.267
梅が香にのつと日の出る山路哉	梅が香に編むや大正百	0.267
梅が香に昔の一字哀れ也	梅が香に琵琶抱く月の法師	0.267
花の時爺は目出度なりにけり	花の雲若葉の風となりにけり	0.267
山は不二このみちのくに櫻かな	山は不二海は二見や初日の出	0.267



EXAMPLE 1

TOPIC ANALYSIS



Let's have a computer read the full-text data of a novel and make a machinery judgment of the subgenre.

Topic modeling :
A method to automatically determine the topic of texts

Outline of research procedure

Consideration of data to be analyzed
When and what kind of data should be considered; the NDL has a list of publicly available works.

Step❶

Step❷

Obtain a list of data to analyze
Narrow down the list of works you need by year and NDC classification from what is publicly available on the Internet. Download from the Next Generation Digital Library via API.

Pre-processing
Preparation of data for machinery analysis.

Step❸

Step❹

Machinery analysis and human consideration
Analysis using LDA. Results are discussed.

Step❶

Consideration of data to be analyzed

Obtain a list of data to analyze

- ✓ There is a list of books that have been available on the Internet on the following page of the National Diet Library, “国立国会図書館デジタルコレクション書誌情報 National Diet Library Digital Collection Bibliographic Information.” https://www.ndl.go.jp/jp/dlib/standards/opendataset/index.html

- ✓ Among the xlsx or tsv files of bibliographic data that can be downloaded above, full text data is available for materials indicated by “著作権保護期間満了copyright protection period has expired”.

Step❷

Download the necessary data

Obtain the full texts

✓ Check the URL to download the full text in the Next Generation Digital Library in the following format

https://lab.ndl.go.jp/dl/api/book/fulltext/885240

永続的識別子
persistent identifier

✓ Using Python, continuously access the full-text download URL and download the listed works.

Chat-GPT4 will write the code for you.

Step❸❹

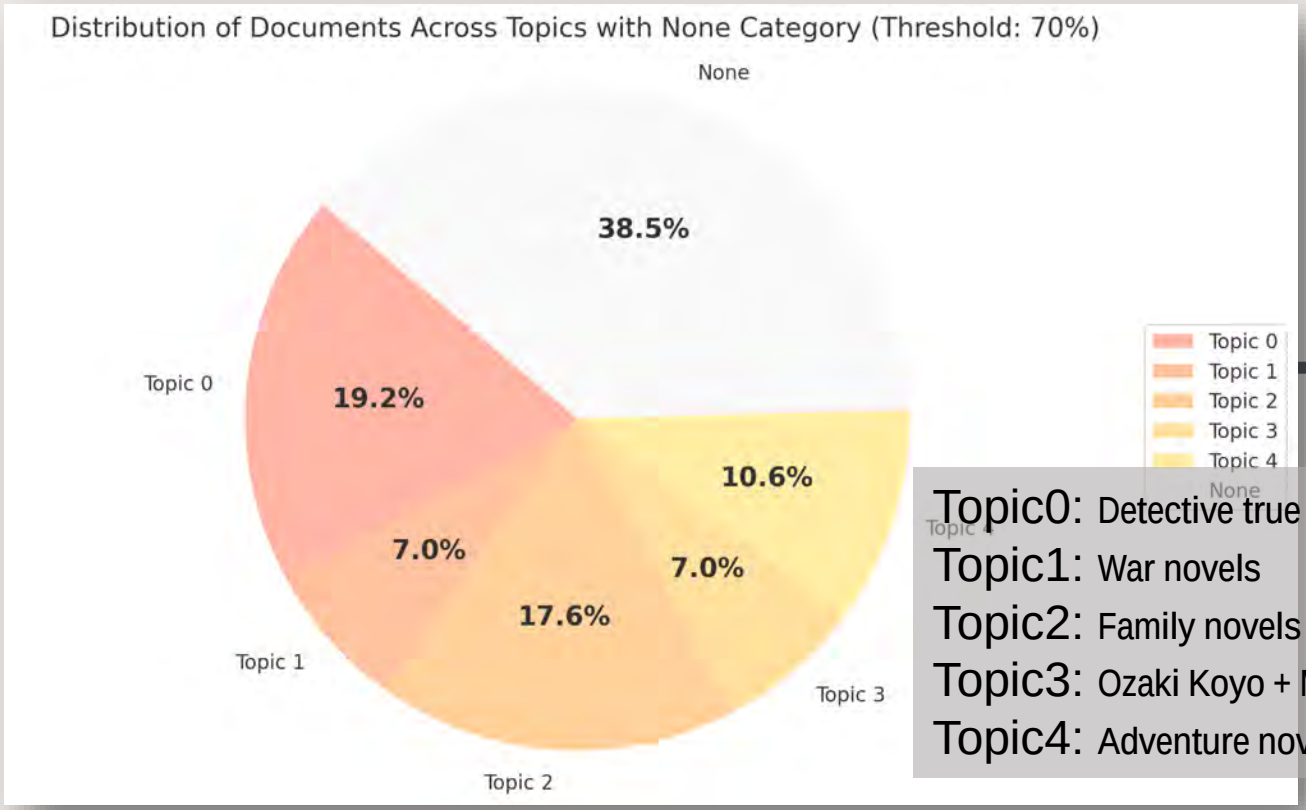
Pre-processing Machinery analysis + discussion

Topic analysis for the full texts

Chat-GPT4 will write the code for you.

- ✓ Preprocessing = Preparing works for performing analysis on a computer.
- ✓ There are two types of files available for download: TXT and JSON (JSONL).
- ✓ Divide the main text into words and perform morphological analysis.
- ✓ The topic analysis was performed using LDA (Latent Dirichlet Allocation).
  - ✓ LDA = One of the text analysis models called the topic model

- ✓ Analyzing the period 1901-1905
- ✓ Five subgenre names are estimated from data with a high degree of agreement for each topic.
- ✓ The ratio of subgenres to all documents is also calculated.



Topic0: Detective true stories, factual novels
Topic1: War novels
Topic2: Family novels
Topic3: Ozaki Koyo + Murai Gensai
Topic4: Adventure novels, legendary novels

Example of analysis results

Topic analysis and chronological arrangement of subgenres

芭蕉全集の句	明治句集の句	類似度
石山の石より白し秋の風	石山の石より白し秋の風	1
道ばたの木槿は馬に喰はれけり	君が家の木槿は馬に喰はれずや	0.438
猪もともに吹かるゝ野分かな	鶏の横に吹かるゝ野分かな	0.357
猪もともに吹かるゝ野分かな	干網に蜻蛉吹かるゝ野分かな	0.357
雲雀なく中の拍子や雉の聲	雲雀なく中の松陰神社かな	0.333
鶯の老を啼くなり茶木畠	鶯の老いを啼くなり花椿	0.333
松杉の尾の上の鐘や秋の暮	鳴り止みし野寺の鐘や秋の暮	0.333
秋の色糖味噌壺もなかりけり	秋の雛笏も冠もなかりけり	0.333
蜻蛉やとりつきかねし草の上	蜻蛉や初汐来る草の上	0.333
このあたり目に見ゆるもの皆涼し	庵の月目に見ゆるもの露涼し	0.313
鶯の笠落したる椿かな	鶯の鳴く時落つる椿かな	0.308
鶯の老を啼くなり茶木畠	鶯の老を啼きけり雲の中	0.308
春の夜や籠人ゆかし堂の隅	春の夜や衣桁に重き戀衣	0.286
春の夜や籠人ゆかし堂の隅	春の夜や芝居打出す川向ふ	0.286
春の夜や籠人ゆかし堂の隅	春の夜や掛物古く桃赤し	0.286
春なれや名もなき山の朝霞	秋風や名もなき山を吹き渡る	0.286
ほとゝぎす啼くや黒戸の濱庇	ほとゝぎす月の桂を塒かも	0.286
岩躑躅染むる泪やほとゝぎす	三軒家蚊帳釣る時やほとゝぎす	0.286
岩躑躅染むる泪やほとゝぎす	鉱毒の貧乏村やほとゝぎす	0.286
岩躑躅染むる泪やほとゝぎす	峡中の暁闇やほとゝぎす	0.286
身にしみて大根からし秋の風	身にしみて市人恋し山住居	0.286
菊の香や奈良には古き佛達	菊の香や傾けつくす大瓶子	0.286
菊の香や奈良には古き佛達	菊の香や我大君は歌聖	0.286
菊の香や奈良には古き佛達	菊の香や一閑張の古机	0.286
菊の香や奈良は幾世の男振	菊の香や傾けつくす大瓶子	0.286
菊の香や奈良は幾世の男振	菊の香や我大君は歌聖	0.286
菊の香や奈良は幾世の男振	菊の香や一閑張の古机	0.286
正月も美濃と近江や閏月	橋一つ美濃と近江や別れ霜	0.273
春の夜や籠人ゆかし堂の隅	春の夜や旅寝淋しき銀屏風	0.267
春の夜や籠人ゆかし堂の隅	春の夜や髣髴として月の富士	0.267

芭蕉全集の句	大正百家選の句	類似度
松島や水を衣裳に夏の月	加茂川や水を枕に夏の月	0.385
雲の峰幾つ崩れて月の山	雲の峰崩れて雄々し不二の山	0.357
梅が香に昔の一字哀れ也	梅が香に編むや大正百	0.286
むく起に隣の花の句ひかな	日の本に普き花の句ひかな	0.286
行末は誰肌觸れん紅の花	行末は誰が手折るらん辻の花	0.286
清瀧の波に洗ふや夏の月	誰もかも笑顔揃ふや夏の月	0.286
月代や膝に手を置く宵の中	自ら膝に手を置く牡丹哉	0.286
蘭の香や蝶の翼にたきものす	蘭の香や源語ひもとく文机	0.286
影待や菊の香のする豆腐串	枕頭に菊の香のする夜明哉	0.286
菊の香や奈良には古き佛達	菊の香や貴船祭りの行戻り	0.286
菊の香や奈良には古き佛達	菊の香や国歌を奏す大広間	0.286
菊の香や奈良には古き佛達	菊の香や全土に及ぶ君	0.286
菊の香や奈良には古き佛達	菊の香や旭輝く日本国	0.286
菊の香や奈良には古き佛達	菊の香や得難き物は人	0.286
菊の香や奈良は幾世の男振	菊の香や貴船祭りの行戻り	0.286
菊の香や奈良は幾世の男振	菊の香や国歌を奏す大広間	0.286
菊の香や奈良は幾世の男振	菊の香や全土に及ぶ君	0.286
菊の香や奈良は幾世の男振	菊の香や旭輝く日本国	0.286
菊の香や奈良は幾世の男振	菊の香や得難き物は人	0.286
菊の香にくらがり上る節句哉	菊の香に集る夜会姿かな	0.286
茶の花に人里ちかき山路哉	茶の花に昼の月あり、詩仙堂	0.286
春雨や簔ふきかへす川柳	春雨や又かけかへす蓄音機	0.273
名月や我と筆架の影法師	名月やさて様々の影法師	0.273
春の夜や籠人ゆかし堂の隅	春の夜やよしあし語る浪華節	0.267
春の夜や籠人ゆかし堂の隅	春の夜やからかって居る電話口	0.267
蝶の羽の幾度越ゆる塀の屋根	蝶の羽の重きもそれか別れ霜	0.267
梅が香にのつと日の出る山路哉	梅が香に編むや大正百	0.267
梅が香に昔の一字哀れも	梅が香に琵琶抱く月の法師	0.267
花の時爺は目出度なりにけり	花の雲若葉の風となりにけり	0.267
山は不二このみちのくに櫻かな	山は不二海は二見や初日の出	0.267

EXAMPLE❷ RECEPTION ANALYSIS

Skip the steps up to pre-processing

To what extent
have Matsuo Basho's haiku
been accepted among modern haiku?

Step❹

Machinery analysis + discussion

Reception analysis for the full texts

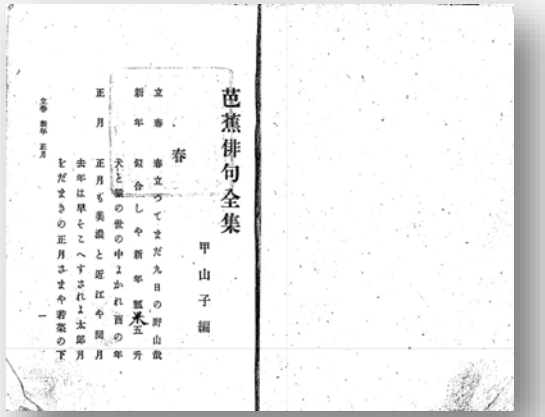
✓ To compare all of Basho's haiku with two collections of modern haiku using N-grams.

✓ Analysis is performed using 2-gram.

2-gram is a set of two consecutive elements (words in this case). Convert each haiku into a 2-gram and calculate the match rate for common word sequences.

✓ The target haiku collections are below.

A）大塚甲山編『芭蕉俳句全集』内外出版協会、1903
B）伊達秋航編『明治句集』新年、春、夏、秋： 春天居書房、1909
C）俳諧滝雫吟社編『大正百家選』俳諧滝雫吟社、1918



A) 『芭蕉俳句全集』 *Basho Haiku Zenshu*

- 大塚甲山編、内外出版協会、1903
- Total haiku **1321**
- Although some of the haiku are questionable now as to whether they are Basho poems, they are used as a way of understanding the Meiji period.

B) 『明治句集』 *Meiji Kushu*

- 全4冊。新年の巻、春の巻、夏の巻、秋の巻。伊達秋航編、春天居書房、1909
- Total haiku **27084**
- Widely collected from newspapers and magazines.

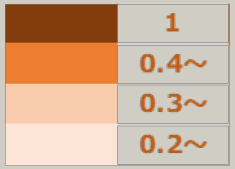
C) 『大正百家選』 *Taisho Hyakkasen*

- 全1冊。俳諧滝雫吟社編『大正百家選』俳諧滝雫吟社、1918。Ginsha is based in Tokushima Prefecture.
- Total haiku **12516**
- Includes poems by local haiku poets.

芭蕉全集の句	明治句集の句	類似度
石山の石より白し秋の風	石山の石より白し秋の風	1
道ばたの木槿は馬に喰はれけり	君が家の木槿は馬に喰はれずや	0.438
猪もともに吹かるゝ野分かな	鶏の横に吹かるゝ野分かな	0.357
猪もともに吹かるゝ野分かな	干網に蜻蛉吹かるゝ野分かな	0.357
雲雀なく中の拍子や雉の聲	雲雀なく中の松陰神社かな	0.333
鷲の老を啼くなり茶木畠	鷲の老いを啼くなり花椿	0.333
松杉の尾の上の鐘や秋の暮	鳴り止みし野寺の鐘や秋の暮	0.333
秋の色糖味噌壺もなかりけり	秋の雛笏も冠もなかりけり	0.333
蜻蛉やとりつきかねし草の上	蜻蛉や初汐来る草の上	0.333
このあたり目に見ゆるもの皆涼し	庵の月目に見ゆるもの露涼し	0.313
鷲の笠落したる椿かな	鷲の鳴く時落つる椿かな	0.308
鷲の老を啼くなり茶木畠	鷲の老を啼きけり雲の中	0.308
春の夜や籠人ゆかし堂の隅	春の夜や衣桁に重き戀衣	0.286
春の夜や籠人ゆかし堂の隅	春の夜や芝居打出す川向ふ	0.286

Example of analysis results

Reception analysis by calculating similarity



a. Although it is a rudimentary method, it is effective for handling texts with a small number of characters, such as haiku.

b. When the similarity rate exceeds 0.3, the similarities can be clearly seen even when compared by human reading comprehension. In 0.2, it simply remains at a level where seasonal words (季語 kigo) are common.

c. It has a high ability to find "similar phrases" from a large number of phrases. A strength of digital humanities methods.

Example of analysis results

Reception analysis by calculating similarity

Summary

1. Full-text data that can be analyzed are now available and are waiting for researchers to use them.
2. Analysis tools that can be used without special knowledge have also appeared.
3. If a researcher could learn and use knowledge of information science, the possibilities for research will expand even further. Collaboration, evolution of AI...