全文テキストデータを利用した近現代日本文学の研究

日比 嘉高

HIBI Yoshitaka

日本研究のための情報源活用法

国立国会図書館ウェビナー　2024

自己紹介

日比 嘉高 HIBI Yoshitaka
名古屋大学大学院人文学研究科 教授

日本の近現代文学・文化論

■『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

このウェビナーについて

ウェビナーのトピック　3つ

1. 近現代日本文学研究者から見た全文テキストデータの状況
2. サービスを使ってみる　次世代デジタルライブラリーと NDL Ngram Viewer
3. 全文テキストデータを利用した文学研究の実践例

1

近現代日本文学研究者から見た全文テキストデータの状況

New!!

検索できる

- 青空文庫
- 国立国会図書館デジタルコレクション
- Google Books
- Maruzen eBook Library (*)
- Kino Den (*)
- 各種新聞のデータベース(*)

(*)は要契約

データ分析できる

- 青空文庫 (**)
- 国立国会図書館次世代デジタルライブラリー(**)
- 国立国会図書館 NDL Lab が提供する
  NDL Ngram Viewer

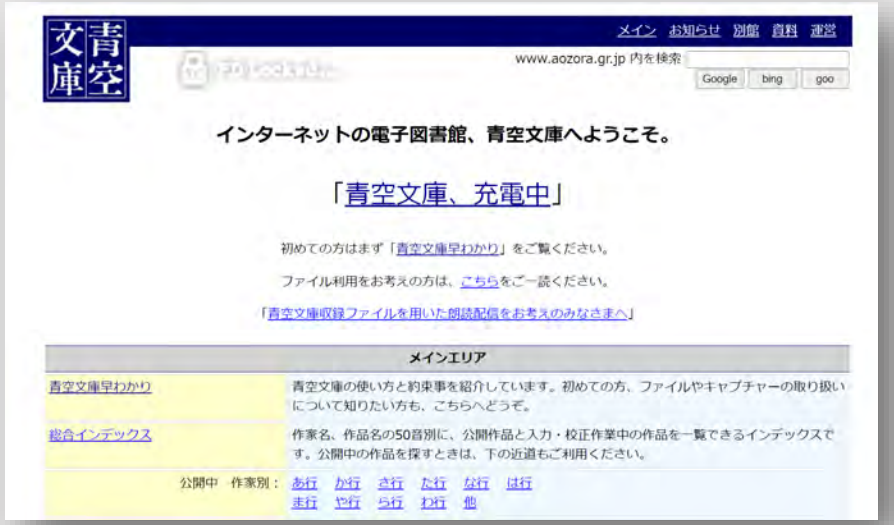
(**)は全文テキストをダウンロードして用いる

代表的な全文検索サービス

文学作品の本文を検索

キーワードを共有する作品の発見

関連キーワードの探索



青空文庫

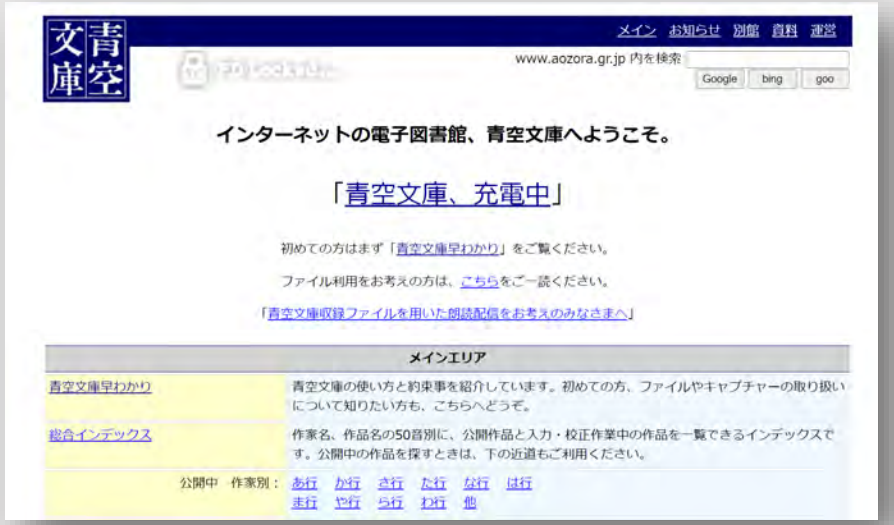
国立国会図書館 デジタルコレクション

全文テキストが利用可能なサイト

より高度なテキスト解析が可能

アイデア次第で広がる可能性

知識なしで使える解析サービスも



青空文庫

国立国会図書館 次世代デジタルライブラリー

青空文庫 「図書カード」のページ

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

ここから
ダウンロード

次世代デジタルライブラリー　個別資料の閲覧画面

2

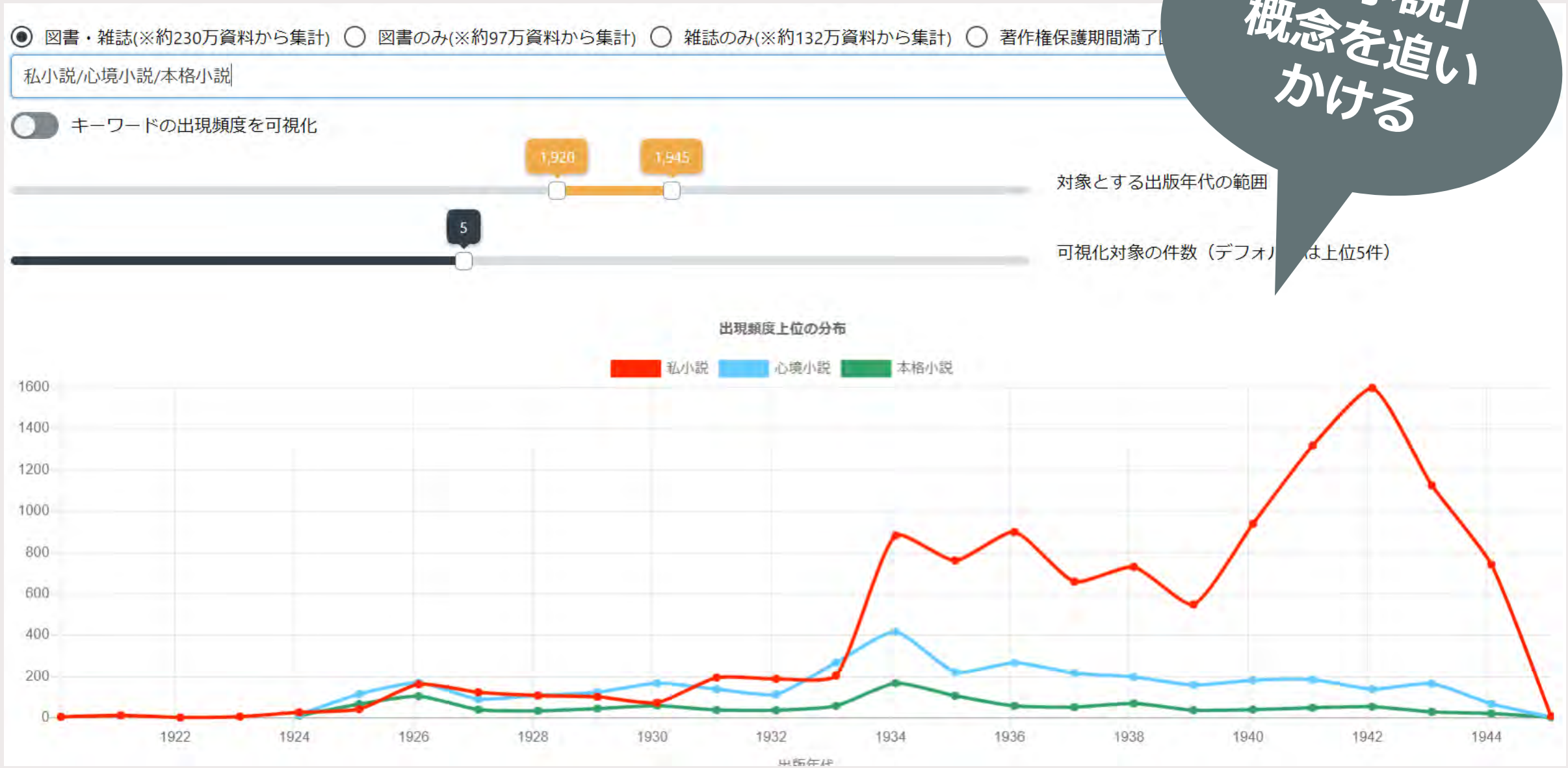
サービスを使ってみる――国立国会図書館の NDL Ngram Viewer

「私小説」概念を追いかける

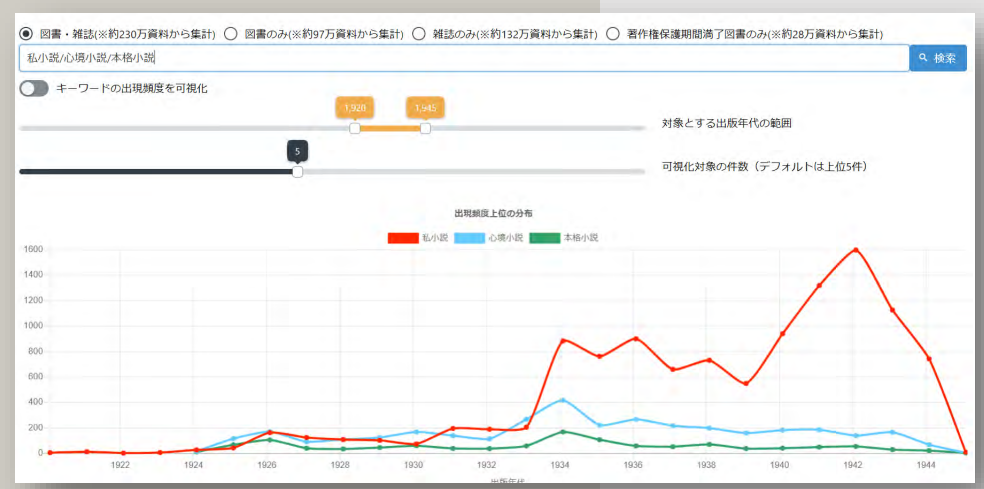
NDL Ngram Viewer

「私小説」「心境小説」「本格小説」の出現年グラフ

「私小説」概念を追いかける



グラフから読み取れること



1. 「私小説」の語の使用頻度の波がわかる。（ただし、資料点数による補正も必要）

2. 波の大きさ＝議論の広がり、その規模の違いがわかる。

3. 語の初出の見当が付く。
「心境小説」の初出は1924年、「本格小説」の初出は1925年あたり。

4. 近接語彙の比較や、交替、衰退の観察ができる。
1925年から増え始めるが、「心境小説」「本格小説」の使用の方が多い。
1930年代後半からは「心境小説」「本格小説」の使用頻度は下がる。とくに「本格小説」はあまり使われない言葉となっていく。

3

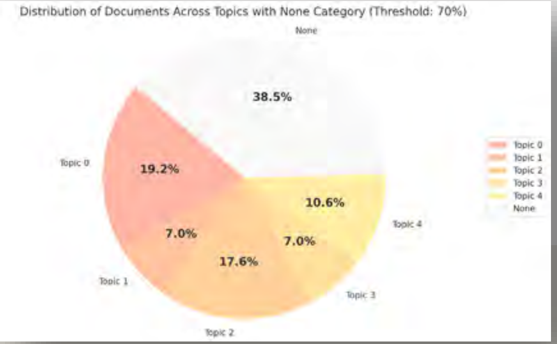
全文テキストデータを利用した文学研究の実践例　～ NDLデータの場合

例❶　トピック分析

- コンピュータを用いた、文学作品のサブジャンルの機械判定
- 未知のサブジャンルの発見
- その交代の観察による新しい文学史叙述の可能性

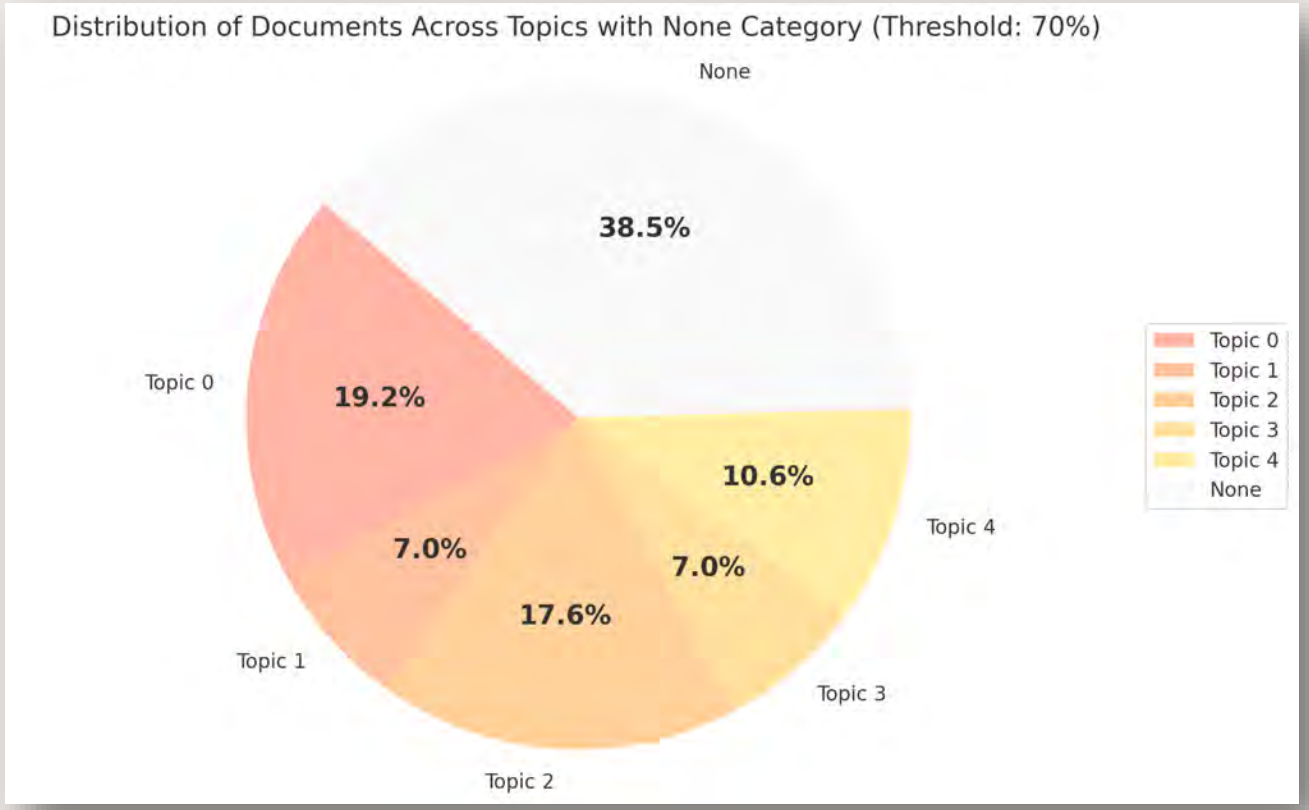
例❷　受容研究

- 文章の一致や類似判定を用いた、受容研究
- 人間には及ばない規模の調査

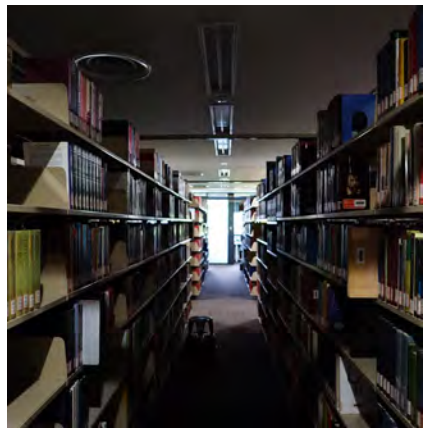


芭蕉全集の句	明治句集の句	類似度
石山の石より白し秋の風	石山の石より白し秋の風	1
道ばたの木槿は馬に喰はれけり	君が家の木槿は馬に喰はれずや	0.438
猪もともに吹かるゝ野分かな	鶏の横に吹かるゝ野分かな	0.357
猪もともに吹かるゝ野分かな	干網に蜻蛉吹かるゝ野分かな	0.357
雲雀なく中の拍子や雉の聲	雲雀なく中の松陰神社かな	0.333
鶯の老を啼くなり茶木畠	鶯の老いを啼くなり花楮	0.333
松杉の尾の上の鐘や秋の暮	鳴り止みし野寺の鐘や秋の暮	0.333
秋の色糠味噌壺もなかりけり	秋の雛笏も冠もなかりけり	0.333
蜻蛉やとりつきかねし草の上	蜻蛉や初汐来る草の上	0.333
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


Distribution of Documents Across Topics with None Category (Threshold: 70%)
None
38.5%
Topic 0
19.2%
10.6%
Topic 4
7.0%
7.0%
17.6%
Topic 1
Topic 3
Topic 2
Topic 0
Topic 1
Topic 2
Topic 3
Topic 4
None

例❶
トピック分析



コンピュータに
小説の全文テキストデータを読ませ
サブジャンルの機械判定をさせてみよう

トピックモデル：
文章のトピックを自動で判定させる手法

研究手順の概略

手順❶

分析するデータの検討

いつごろの、どんなデータを対象とするか。
NDLに公開中の作品のリストがある。

手順❷

必要なデータをダウンロード

必要な作品を、インターネット公開分から、
年とNDC分類で絞り込む。
次世代デジタルライブラリーからAPIでダウンロード。

手順❸

前処理

機械解析をするまでの
データ的な準備作業。

手順❹

機械解析の実行と人による考察

LDAを用いて分析。結果をみて、考察する。

手順❶
分析するデータの検討

リストを入手

- ✓ 国会図書館の以下のページ「国立国会図書館デジタルコレクション書誌情報」に「インターネット公開」された「図書」のリストがある。

  https://www.ndl.go.jp/jp/dlib/standards/opendataset/index.html

- ✓ 上記でダウンロードできる書誌データのxlsxまたはtsvファイルのうち、「著作権保護期間満了」となっている資料は、全文テキスト・データが入手可能。

手順❷
必要なデータをダウンロード

全文テキストの入手

✓ 次世代デジタルライブラリーで、全文テキストをダウンロードするURLを確認
　次の形式：

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

永続的識別子

✓ Pythonを用い、全文テキストのダウンロードURLに連続的にアクセスし、リストアップした作品をダウンロードする。

コードはChat-GPT4が書いてくれます

手順❸❹
前処理
機械解析＋考察

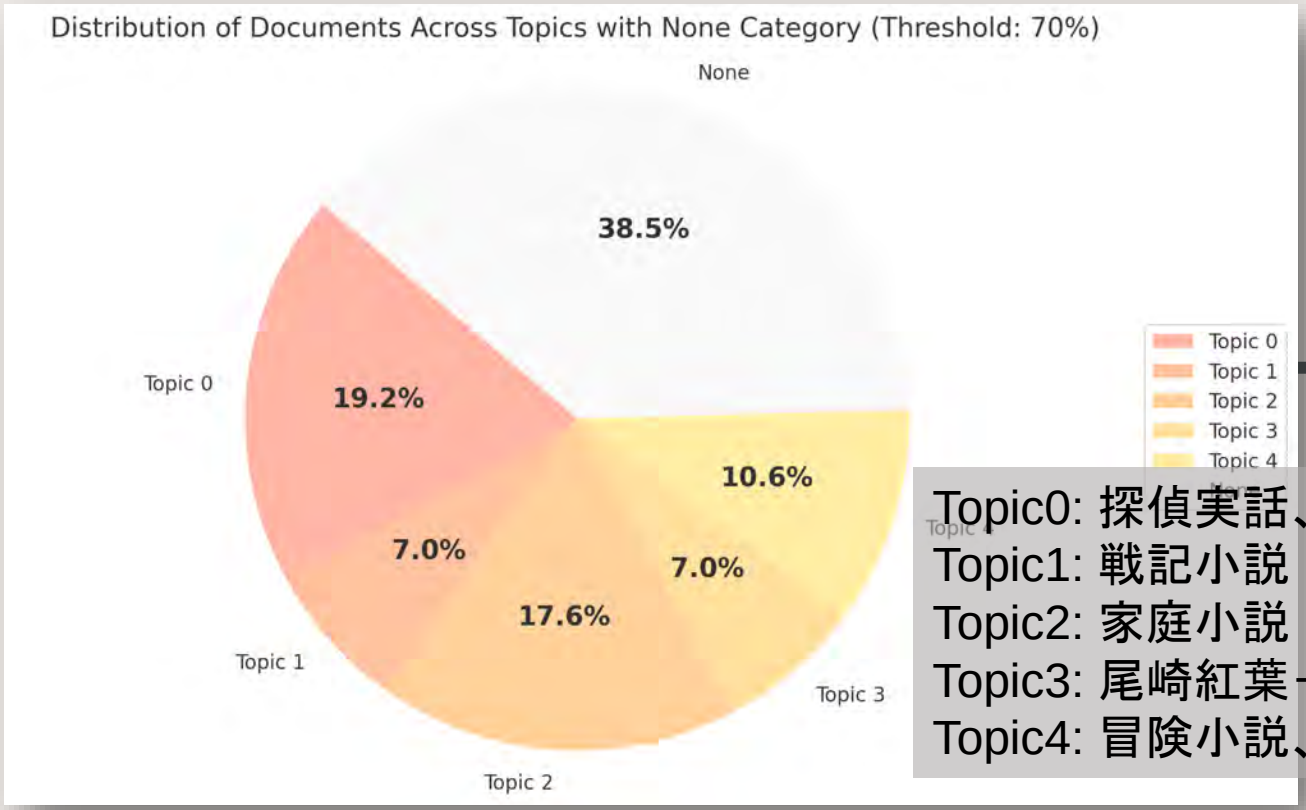
全文テキストに対するトピック分析

コードはChat-GPT4が書いてくれます

- ✓ 前処理＝コンピュータで解析を行う準備作業。
- ✓ ダウンロードで手に入るファイルは、2種類。txt と JSON(JSONL)
- ✓ 本文テキストを、単語ごとに分割し、かつ形態素解析する。
- ✓ 解析は、LDA(Latent Dirichlet Allocation)で行った。
- ✓ トピックモデルと称されるテキスト分析のモデルの１つ。

- ✓ 1901-1905年の期間を分析
- ✓ 各トピックの一致度の高いデータから、5つのサブジャンル名を推定。
- ✓ 全文書に対するサブジャンルの割合も算出。

分析結果の例
トピック分析とサブジャンルの時代的配置



Topic0: 探偵実話、実事小説
Topic1: 戦記小説
Topic2: 家庭小説
Topic3: 尾崎紅葉＋村井弦斎
Topic4: 冒険小説、伝奇小説

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

例❷
受容研究

前処理までの
手順は省略

松尾芭蕉の俳句は
近代俳句にどの程度受容されているか

手順❹
機械解析＋考察

全文テキストに対する受容研究

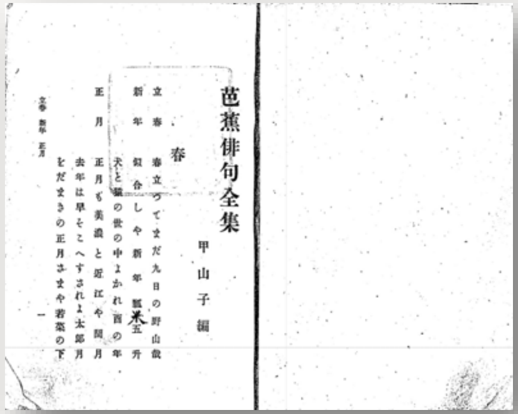
✓ 芭蕉の全発句と
近代俳句の句集2種類とを
N-gramを用いて比較する。

✓ 解析は、2-gramで行う。

2-gramとは、連続する2つの要素(今回の場合は単語)の集合。
それぞれの俳句を2-gramに変換し、
共通する単語の並びの一致率を計算。

✓ 対象とした句集は以下。

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』

- 大塚甲山編、内外出版協会、1903
- 全**1321**句
- 芭蕉句であるか、現代では疑問のある句も含まれるが、明治期の理解の一つとして利用する

B） 『明治句集』

- 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
- 全**27084**句。新聞雑誌から広く採録。

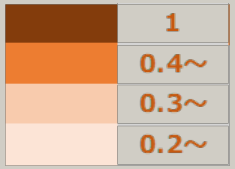
C） 『大正百家選』

- 全1冊。俳諧滝雫吟社編『大正百家選』俳諧滝雫吟社、1918。徳島県を拠点にする吟社。
- 全**12516**句。地元の俳人たちの句を採録。

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

分析結果の例

類似度の計算による受容研究



分析結果の例

類似度の計算による受容研究

a. 初歩的な方法だが、俳句のような文字数が少ないtextを扱うには、有効。

b. 体感として、類似率が 0.3 を越えると、人の読解による対比でも類似性が顕著に見いだせる。0.2では、単に季語が共通しているレベルに留まる。

c. 大量の句から、「類似句」を見つけ出す能力は高い。Digital Humanities的手法の、強み。

ウェビナーのまとめ

1. 解析の対象とできる全文テキストデータが登場しており、研究者の利用を待っている

2. 特別な知識なしで使える解析ツールも登場している

3. 情報科学の知見を援用できれば可能性は、さらに広がる。
共同研究、AIの進化…