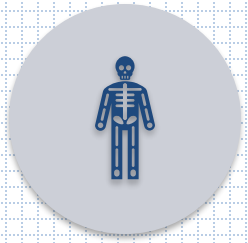
市民による郷土の固有表現抽出と機械学習の連携で加速する古文書読み解き

佐賀大学地域学歴史文化研究センター
講師（研究機関研究員）吉賀夏子
Mail: natsukoy@cc.saga-u.ac.jp
Twitter: summarcat
2021-01-15

自己紹介



ほぼ大分県出身佐賀県在住

県北西部の山中に埋もれて育つ

佐賀大学
農学部卒　農学修士、
理学修士、博士（学術）

15年程度佐賀大学総合情報基盤センター勤務（フロントエンドとデータベース、サイト管理担当）

現在同大学地域学歴史文化研究センターで勤務

文化財や歴史データを扱う情報学を専攻（人文情報学）

学芸員資格あり

市民団体**CODE FOR SAGA**でITを使った社会課題解決の取り組み

研究テーマ

「モノ＋情報＝価値あるモノ」

とくに

身の回りにあるものから価値を創る・見出すには

どうしたらいいのか？

実在する「地域の課題」と「データ」を使った実践的研究例

- **「小城藩日記データベース」**（江戸期・藩の業務記録）の開発（人文系DBの機械可読化）
  - 専門性の高いデータベースの内容を多くの市民と共有し活用するための道筋を創る
- 佐賀県の雨量、川の水位などウェブ上に散財する**「防災関連データ」**の集約とデータプラットフォームの試作・スマートスピーカーを使った情報提供
- （市民活動）佐賀県下を走る**バスデータ**（オープンデータ化済み・公共交通）の応用（アプリ開発）
- （NPO）福岡市の**緊急避難場所**API（オープンデータ化・APIサービス構築済み）を用いたAlexaアプリの開発

これまで入手が難しい・内容把握が困難な「情報」を実在する物事にうまく結びつけて、社会課題の解決に利用する

現在の主な業務

佐賀大学地域学歴史文化研究センターの**「小城藩日記データベース」**の構築と管理、運営

「小城藩日記データベース」とは

https://crch.dl.saga-u.ac.jp/nikki/

- **江戸期に作られた業務日誌のデジタルアーカイブ**
  - **（佐賀県の）小城藩**（佐賀藩の支藩）が記録した**業務日誌**「日記」のタイトル部分を検索できる
  - 小城藩は佐賀藩の支藩
  - 冠婚葬祭、佐賀藩との連絡や幕府、行政、経済、兵役、事件、災害など、当時の幅広い出来事を記録
  - 現在は専門家が「くずし字」から翻訳（翻刻）した**全73984件のタイトル文と原文画像を閲覧**できる

- **オープンデータ**として2018年4月からウェブ公開中
  - **利用許諾：教育・研究利用は実質自由**
    - **原本画像を含む「全データ」を個人でダウンロードして、ホームページ，SNS，分析など気軽に利用することを想定**

佐賀大学附属図書館「日記」原本

そもそも

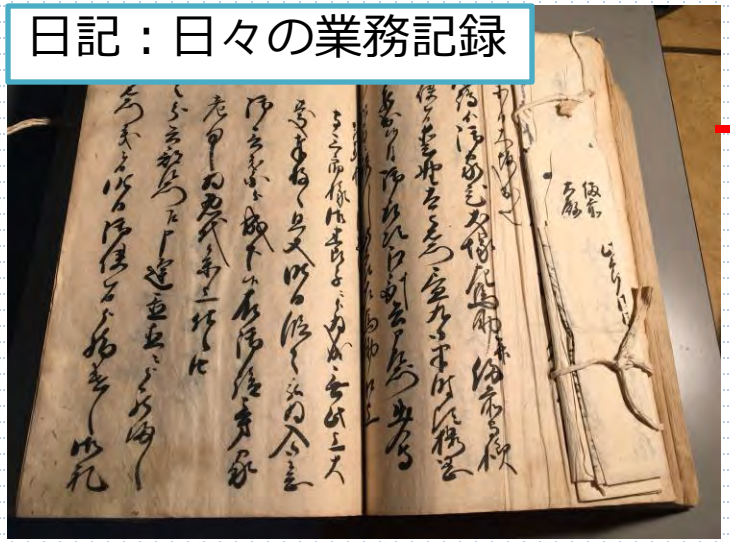
データベースに保存するデータはどう
やってつくったのか？

紙の業務記録

- 日記とその目録
  佐賀大学附属図書館（本庄）に所蔵
- 日記：**84**年分
- 目録：日記の内容を**箇条書き**にしたもの**122**年分
- 箇条書きになった日記の要約を「記事文」と呼ぶ

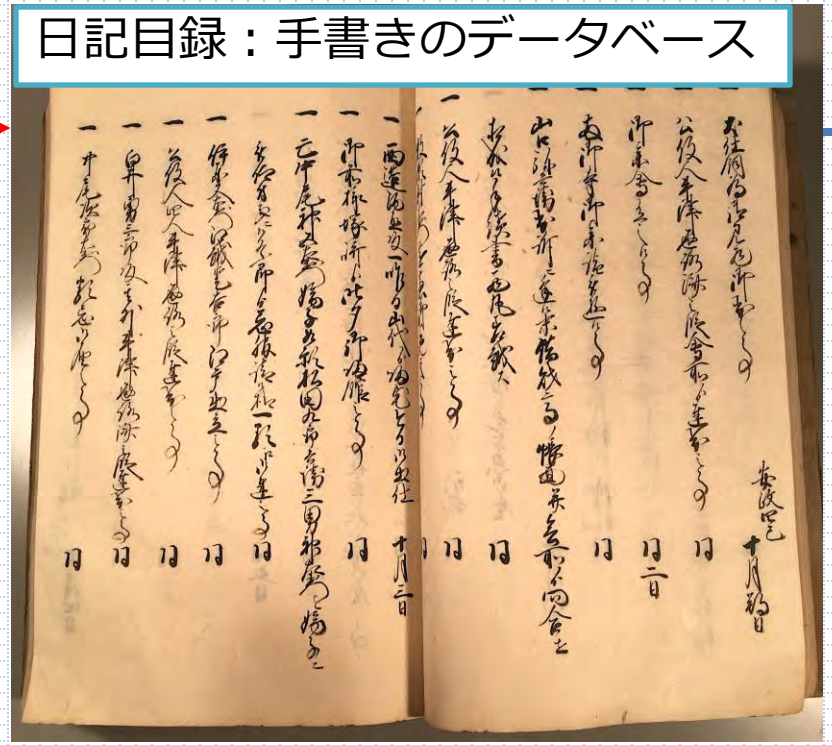
実際のデータベース化

日記：日々の業務記録



「日記」には、様々な事柄が詳細に記載

日記目録：手書きのデータベース



「日記」内容を要約した「日記目録」も、江戸期に並行して作成

数名（最大７名）で、日記目録を翻刻（活字化）

小城藩日記データベースに蓄積

江戸時代に作成

現代に作成

目録化の利点

- 日記の長い内容が、箇条書きに要約されているため、**特定の内容についてすばやく探しやすい**
- **コンピューターにとってもデータの検索や保存がしやすい**
  - もし、「日記」だけが残っていたなら、解読（翻刻）作業やデータ量が**膨大**すぎてデータベース化は困難
  - **江戸時代に**小城藩が目録にしたことが、現代のコンピュータでの検索に**有利に働く**

現在のデータベース

- 現存する「目録」全**73984**件登録完了！
- 当初の想定10万件より目録件数は少なかった
- 目録の記事文に対応する「日記」の記事画像を探す作業中

目録の記事文を読むのは敷居が高い

文のほとんどは漢字でかかれていて、文法も「**候文**」（そうろうぶん）
候文：主に中世から近世にかけて使用された、手紙や文書に用いる文語体

- 元延公御看病御暇御拝領直二御出府可被遊処御不快二付御断御使者江戸へ村川孫四郎被仰付候事
- 野口進之允殿遠慮御免之事
- 祥光山御名代事
- 殿様請役所え以後御出座御参会日二被遊候事
- 大蔵殿死去二付三岳寺へ御代香事
- 右同人呼出之儀医師頭取え誰ソ同道二て罷出候様懸合
- 下川覚介長崎へ被差越候事
- 佐嘉被遊御越候事
- …

記事文からキーワードを抜き出しておけば人もコンピュータも内容を理解しやすい

キーワードや分類を別途用意

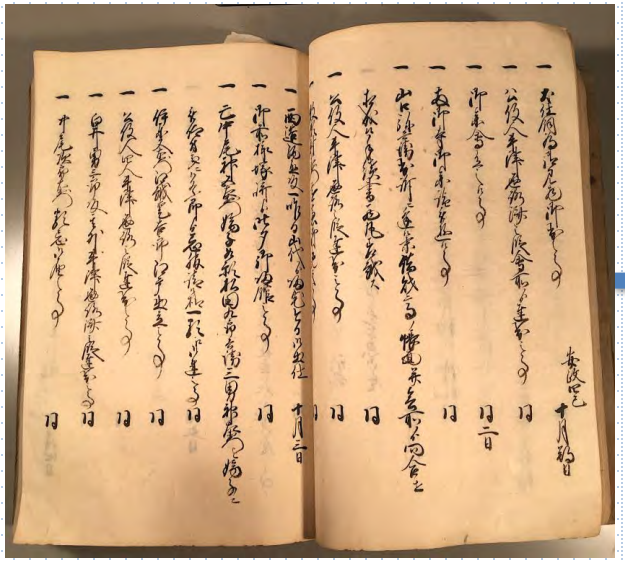
年月日-昇順 表示件数 10 次の 10 件

全て 画像有

#	登録番号	和暦年月日	記事 典拠利用/翻刻	分類	関連する人名
#1	10 SPARQL	寛文1年7月29日 1661 グレゴリオ暦 1661-07-26	右御亡者様へ御名代諫早へ三浦杢之助御香典銀弐枚之事 画像なし	冠婚葬祭 自動タグ 寺社, 冠婚葬祭, 藩主家, 賞罰, 家中 出来事 亡者, 香典, 銀 地名 諫早	諫早直孝後妻, 三浦杢之助 人名 三浦杢之助 役名 名代
#2	2 SPARQL	寛文1年7月13日 1661 グレゴリオ暦	鍋島志摩殿死去之事 画像なし	冠婚葬祭 自動タグ 冠婚葬祭, 家中, 他藩, 藩主家, 医学 出来事 死去	鍋島茂里(1), 鍋島茂里(2) 人名 鍋島志摩

記事文から固有表現（キーワード）を抜き出す手順

日記目録の記事文



くずし字を読み取り
テキストに変換

ほんこく
翻刻

コンピュータで処理しやすい
データに変換

固有表現
抽出

翻刻例）

多賀丸様御上京御供犬塚茂右衛門被仰付候事

人名　トピック（出来事）　役割　人名

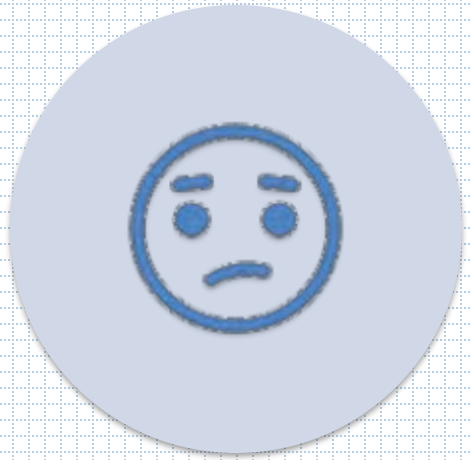
（多賀丸様が上京されるので，お供に犬塚茂右衛門が命じられたとの事．）

地域色（キーワード）の強い固有表現

- 固有表現抽出
  - 資料中でキーワードとなる言葉を抽出して「クラス」に仕分ける
  - **人名・場所・出来事**は地域色が強く収集困難

固有表現クラス名		説明
Person	人名	人名，呼称
Date	日時	日時を表す語
Place	場所	座標で指定可能な地名
Event	出来事	検索キーワードとなり得る語
Role	役職、役割	役職，家族関係
Terms	候文用語	接続詞，定型句
Quantity	数量	数および単位を表す語

ウェブの欠点



多くの人にとって比較的マイナーな

地域固有のデータはなかなか見つからない

対処方法

1. 佐賀県立図書館データベースやWikipediaで集める
2. 地元の郷土資料に詳しい人に手伝ってもらう

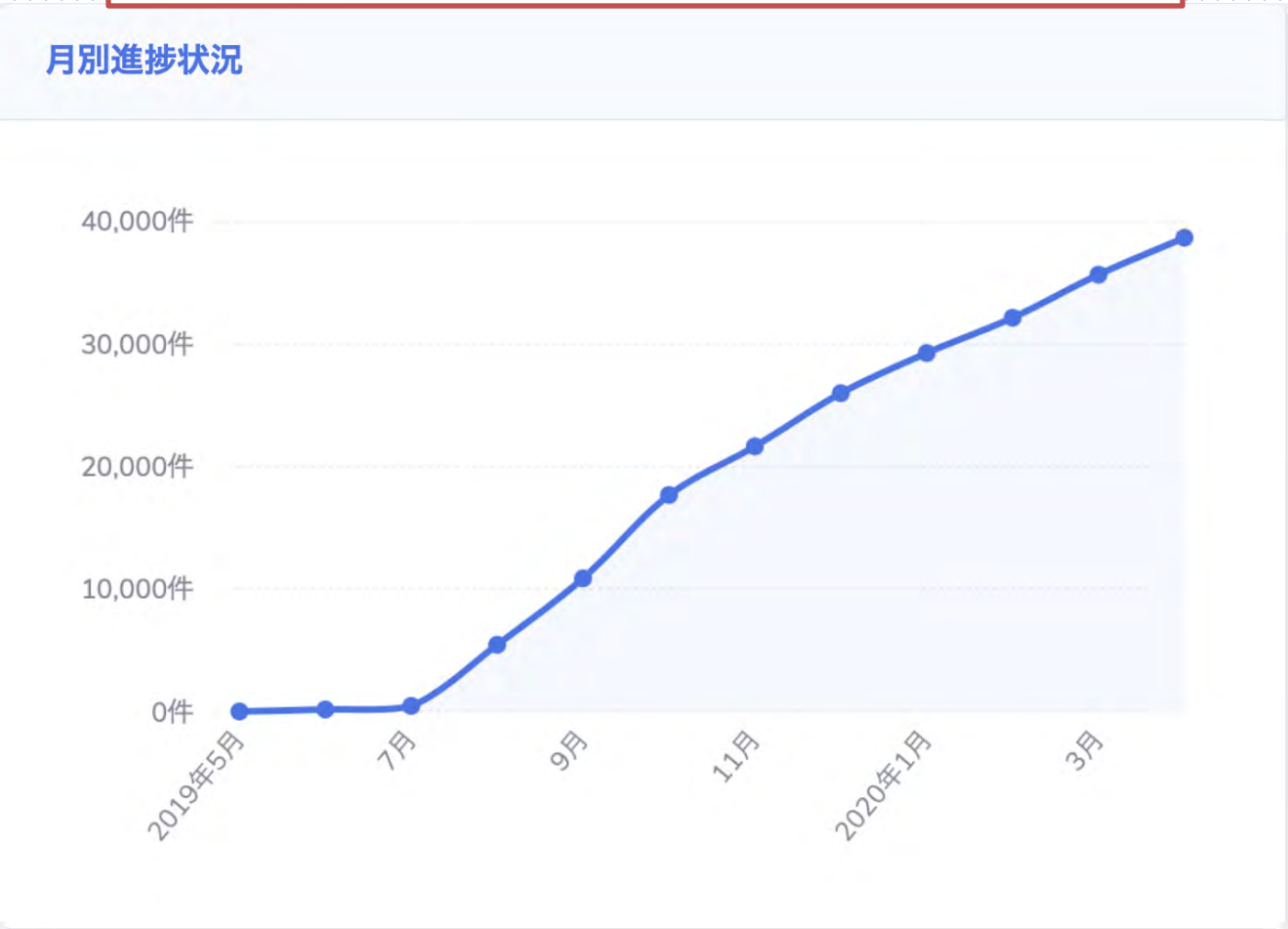
県立図書館DBには人名、地名、寺社名などかなりの情報が入っているが、小城固有の情報は少なめ→

クラウドソーシング

抽出作業は2年を予定していたところ、年度内で終了

- 小城市歴史資料館、佐賀大学で郷土資料を読める人材を依頼
- キーワードの抽出は、専用ホームページで

73984記事のうち、前方40000記事からキーワード抽出
8人で11ヶ月かけて作業
（うち5人有償）



抽出ルールのすり合わせメモの例

固有表現ラベリング凡例（記事＃1～＃100 より抜粋作成）　R1.7.25.

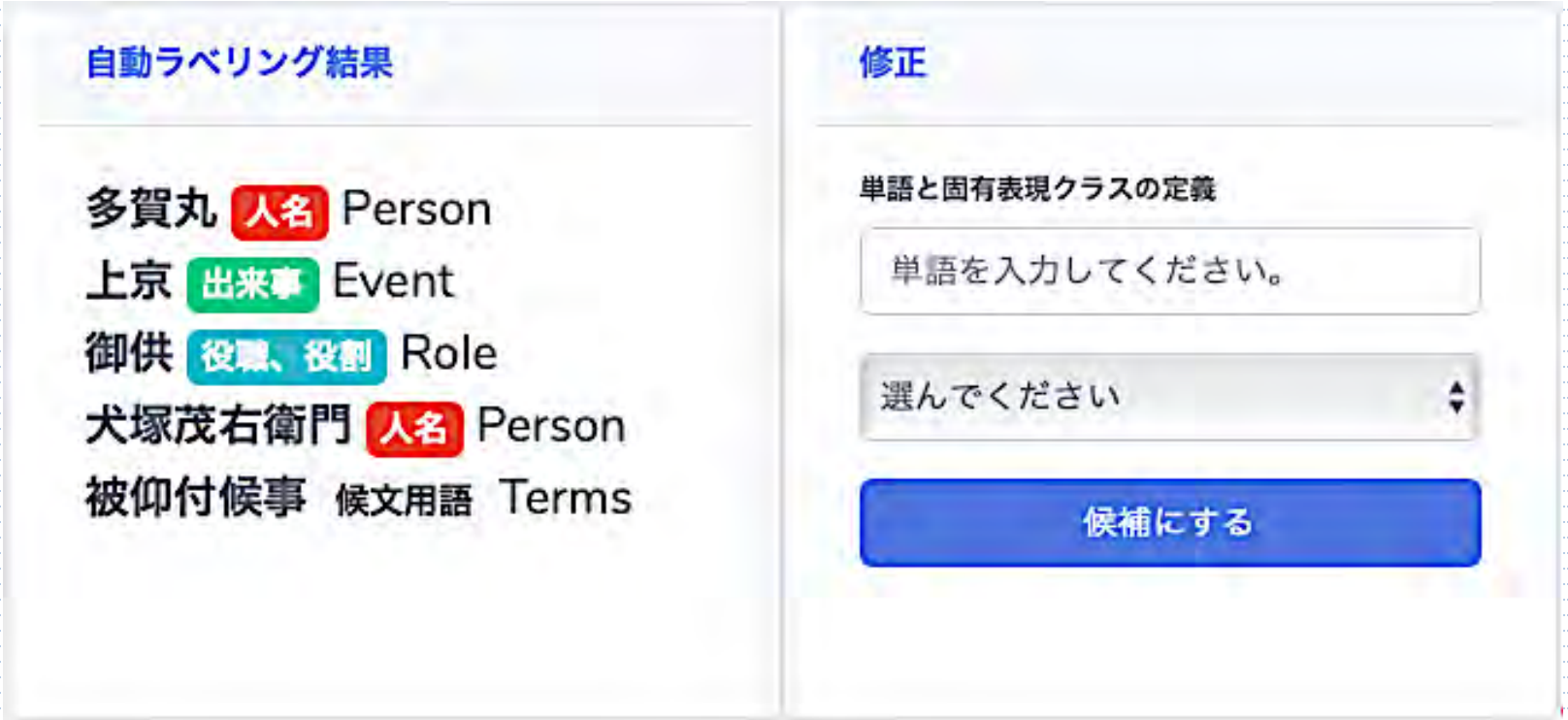
①Person	②Place	③Event	④Role	⑤Terms	⑥Date	⑦Quantity	⑧
人(人名)	場(場所)	出(出来事)	役(役職役割)	候(候文)	日(日時)	数(数量回数単位)	固無(固有表現無し)

＃	記事	固有表現分類
2)	長寿院様御死去之事　諫早石見殿内方直茂公御女 ⇒	人(長寿院,諫早石見),出(死去),役(内方，直茂公御女)
3)	城州様本行寺へ御葬礼之事 ⇒	人(城州),場(本行寺),出(葬礼)
4)	月堂様え御燈爐被相進候御方之事 ⇒	人(月堂),出(燈爐,御方),候(被相進候)　※「～被相進候御方～」「御方」だけ『出』に分類、「被相進候」は『候』へ
5)	月堂様御塔前へ御拝塔之衆之事 ⇒	人(月堂),出(塔前,拝塔)　※「之衆」は分類不要
6)	長寿院様御死骸諫早へ御越事 ⇒	人(長寿院),場(諫早),出(死骸),候(御越事)　※動詞入りの末尾単語は「事」まで入れ『候』へ→御越事，御書事，差迦候事
7)	南祥院様御塔前へ右同断事 ⇒	人(南祥院),出(塔前),候(右同断事)
8)	右御亡者様へ御名代諫早へ三浦杢之助御香典銀弐枚之事 ⇒	人(三浦杢之助),場(諫早),出(亡者,名代，香典，銀),数(弐枚)
9)	左近殿へ御同人御遺物鉄砲壱挺隼一連被遣候事 ⇒	人(左近),出(御同人,遺物,鉄砲,隼),候(被遣候事)，数(壱挺,一連)　※「御同人」の「御」外すと意味が違って来ると判断（目上の方の意）

「御」が目上の意で使われている場合は、「御」は Terms となります。

あらかじめ、抽出とクラス（この図ではラベリング）への仕分け方について、作業グループで検討

抽出作業画面例



あらかじめ固有表現を自動で抽出し、作業者に適宜修正を求める
基本コピー・アンド・ペーストしてボタン押すだけの操作

固有表現リスト

https://winter.ai.is.saga-u.ac.jp/cs/ne-words

10 件/ページを表示

検索 肥州

#	固有表現種類名	単語（固有表現）	単語数
63	人名 Person	肥州様	399
3698	地名 Place	肥州	6
4830	人名 Person	肥州	4
6678	人名 Person	松平肥州	2
#	固有表現種類名	単語（固有表現）	単語数

1 ページ中 1 ページを表示 (全 11,521 件から 4 件抽出)

前 1 次

キーワード（固有表現）分類作業の情報共有

誰でも閲覧可能

集めたキーワードは？

- データベースに集めたキーワードを自動反映させ、多くの記事にキーワードが付いた
- 集めたキーワードは機械学習による記事内容のカテゴリー分け、高度な情報探索に利用できる

可視化ツールで関連キーワードを表示

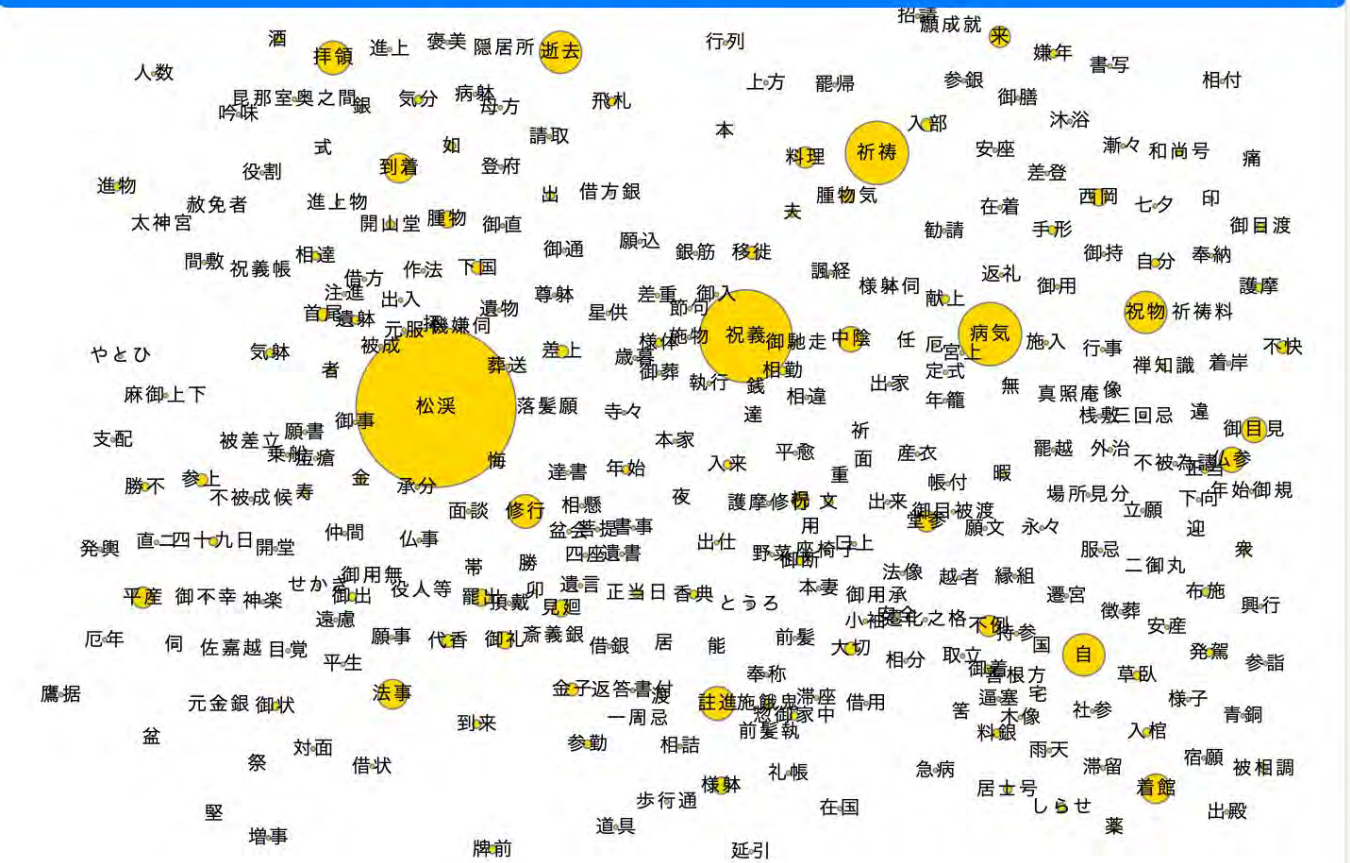
データ転送と描画に時間がかかるので、200件程度までの結果に絞って可視化するのがオススメ

目録記事内容の内訳→

関連語

マウスなどを使って移動、ズームイン・アウトできます。
元に戻すときはリセットボタンを押してください。

リセット



松渓 38　祝義 22　病気 15　祈祷 15　逝去 10　自 10　祝物 10　拝領 8　註進 8
修行 8　到着 7　法事 7　着館 7　中陰 6　仏参 6　御目見 6　料理 5　平産 5

場所

松渓 38　金栗 25　江戸 16　岩蔵寺 14　御願 11　祥光山 9　二丸 7　御家 7
御目見 6　桜岡 5　佐嘉 4　西岡 4　牛頭庵 3　彦島 3　同寺 2　蓮池 2　筑前 2
唐津 2　宗智寺 2　島原 2　江府 2　開山堂 2　鳳来寺 2　御船 2　西丸 2　仏事 1
宅 1　上方 1　太神宮 1　御館下町 1　相懸 1　薬師堂 1　江戸神明宮 1
昆那室奥之間 1　柿久村宝蔵寺 1　御屋敷 1　鈴岩 1　万山狩場 1　真照庵 1
西岡両役所 1　西岡岩蔵寺 1　昆那室 1　住心院 1　称光山 1　黄檗山 1　千栗 1
二御丸 1　長崎奉行 1　天山宮 1　大坂 1　大里 1　御本丸 1　森川 1　松計 1
小城岡町 1　牛津御茶屋 1　新川 1　松尾山 1　松景 1　桟敷 1　祇園社 1　鹿島 1
肥後 1　大堂 1　延岡 1　飫肥 1　京都 1　とゝきかはた 1

↑地図上で描画したいので、**土地の名前に詳しい人**を募集予定

←キーワードから、**さらなる関連キーワード**を見つけられます

様々な形式でデータをダウンロード可能

- 検索結果をそのままお手元にダウンロード
- TSV形式はExcel（スプレッドシート）ですぐに表示できる

興味のある方はぜひ全データをダウンロードして遊んでみてください

https://crch.dl.saga-u.ac.jp/nikki

そして面白いことが見つかったら私に教えてください😉