進捗報告

平井　遥斗

富山県立大学　情報システム工学科

2023 年 01 月 09 日

共起関係の分析

jaccard 係数

今までちゃんと計算できていると思っていたが，少しおかしいことに気づいたので修正を行った．

間違えていたこと

共起関係の分析は一般的には文章単位で行う．

- Python は，機械学習でよく利用されるプログラム言語であり，他の言語よりも優れている．
- Python は，機械学習だけではなく，Web アプリ開発などでも利用されている．

文章単位で見たときの 2 つの単語の出現回数などを用いる．

しかし，すべての文章をひとくくりに計算していた．

- Python は，機械学習でよく利用されるプログラム言語であり，他の言語よりも優れている．Python は，機械学習だけではなく，Web アプリ開発などでも利用されている．

専門用語の辞書作成

辞書に専門用語を登録

抽出した重要語を Janome の辞書の書式に合わせて csv を作成した．ここで，辞書の大きさを大きくしすぎるとプログラムが正常に動作しないことが分かった．
制限しないと 50 万行から 100 万行の辞書が作成されてしまい．時間がかかる上にたまに正常にプログラムが動作しなくなる．

ホログラム記録	-1	-1	1000	名詞	固有名詞	*	*	*	*	ホログラム	*	*
カード	-1	-1	1000	名詞	固有名詞	*	*	*	*	カード	*	*
前記スクラッチ隠蔽層	-1	-1	1000	名詞	固有名詞	*	*	*	*	前記スクラ	*	*
隠蔽	-1	-1	1000	名詞	固有名詞	*	*	*	*	隠蔽	*	*
顔料	-1	-1	1000	名詞	固有名詞	*	*	*	*	顔料	*	*
屈折率層	-1	-1	1000	名詞	固有名詞	*	*	*	*	屈折率層	*	*
表面	-1	-1	1000	名詞	固有名詞	*	*	*	*	表面	*	*
スクラッチカード	-1	-1	1000	名詞	固有名詞	*	*	*	*	スクラッチ	*	*
組成物	-1	-1	1000	名詞	固有名詞	*	*	*	*	組成物	*	*
重合	-1	-1	1000	名詞	固有名詞	*	*	*	*	重合	*	*
印刷	-1	-1	1000	名詞	固有名詞	*	*	*	*	印刷	*	*
偽造防止性	-1	-1	1000	名詞	固有名詞	*	*	*	*	偽造防止性	*	*
用インキ組成物	-1	-1	1000	名詞	固有名詞	*	*	*	*	用インキ組	*	*
光ラジカル重合開始剤系	-1	-1	1000	名詞	固有名詞	*	*	*	*	光ラジカル	*	*
光カチオン重合開始剤系	-1	-1	1000	名詞	固有名詞	*	*	*	*	光カチオン	*	*
透明基材	-1	-1	1000	名詞	固有名詞	*	*	*	*	透明基材	*	*

図 1: 辞書 (csv)

ネットワークの可視化

2D グラフ

表示する要素が少ないときは 2D の方が見やすいかもしれない，もっと要素が多くなるとごちゃごちゃするので 3D の方が見やすくなる．2D グラフも可視化の一つとして取り入れてもいい．

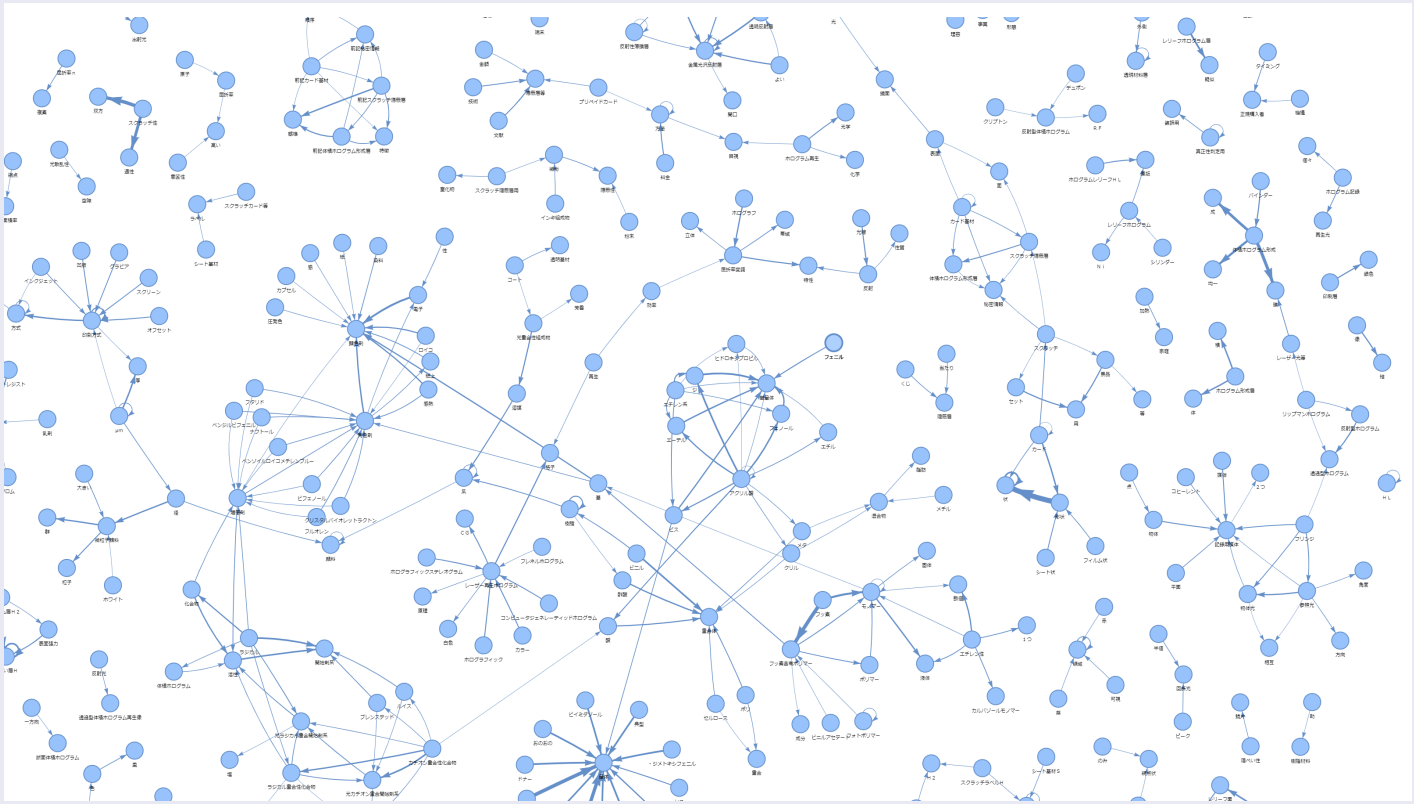

図 2: 2D グラフ

ネットワークの可視化

3D グラフ

自分のデータに適応してみた.
Json ファイルを作成してグラフを作成した

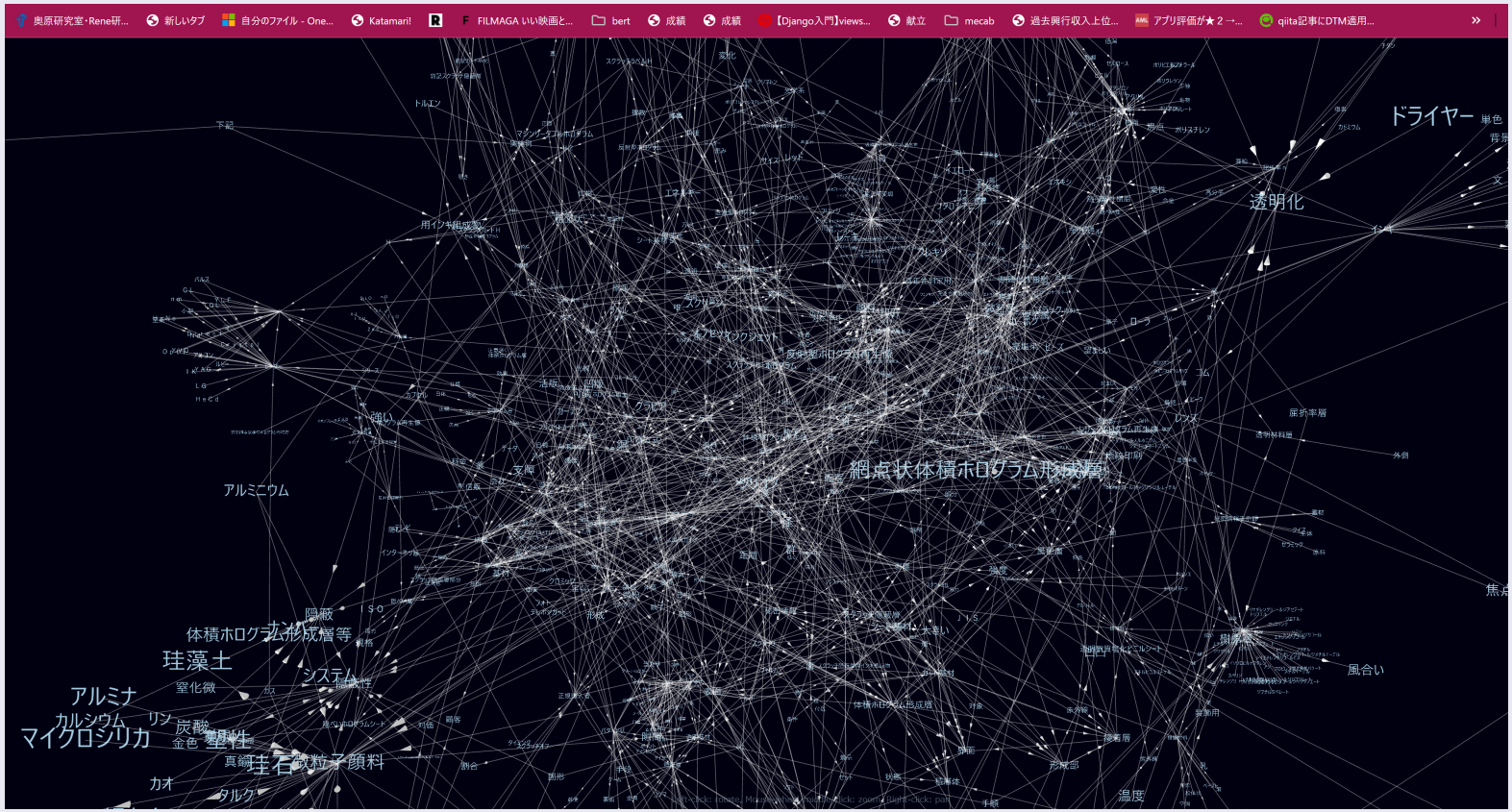

図 3: 3D グラフ

クラスターの解釈

クラスターの解釈

クラスター内の重要語を計算して表示することは可能だが，1 つのクラスターの重要語を計算するのに結構時間がかかるので，10 以上のクラスターを計算しようと思うと，時間がかかりすぎる．

現状

クラスターの中からランダムで 10 個要素を取り出し，その中で重要な単語を 3 つ表示することでクラスターの解釈の表示を行った．

理想

- クラスターの代表的な文章だけを取り出して，重要度を計算する．各クラスターの中心に近い文章ほどそのクラスターの中心的な文章である可能性が高い．
- クラスターからまんべんなく要素を持ってこれる手法があれば全体的な要素を含めることができる．

システム

フロントページ

ここに検索したい単語を入力する.
複数入力したい場合は間にスペースを空ける.

図 4: フロントページ

ロード画面

グラフを作るまで時間がかかるのでロード画面を作成した.
最終的には進捗バーを追加する予定.

図 5: ロード画面

システム

クラス選択画面

クラスを選択できるようにした.
また，それぞれのクラスターがどういうものなのかを表示できるようにした.

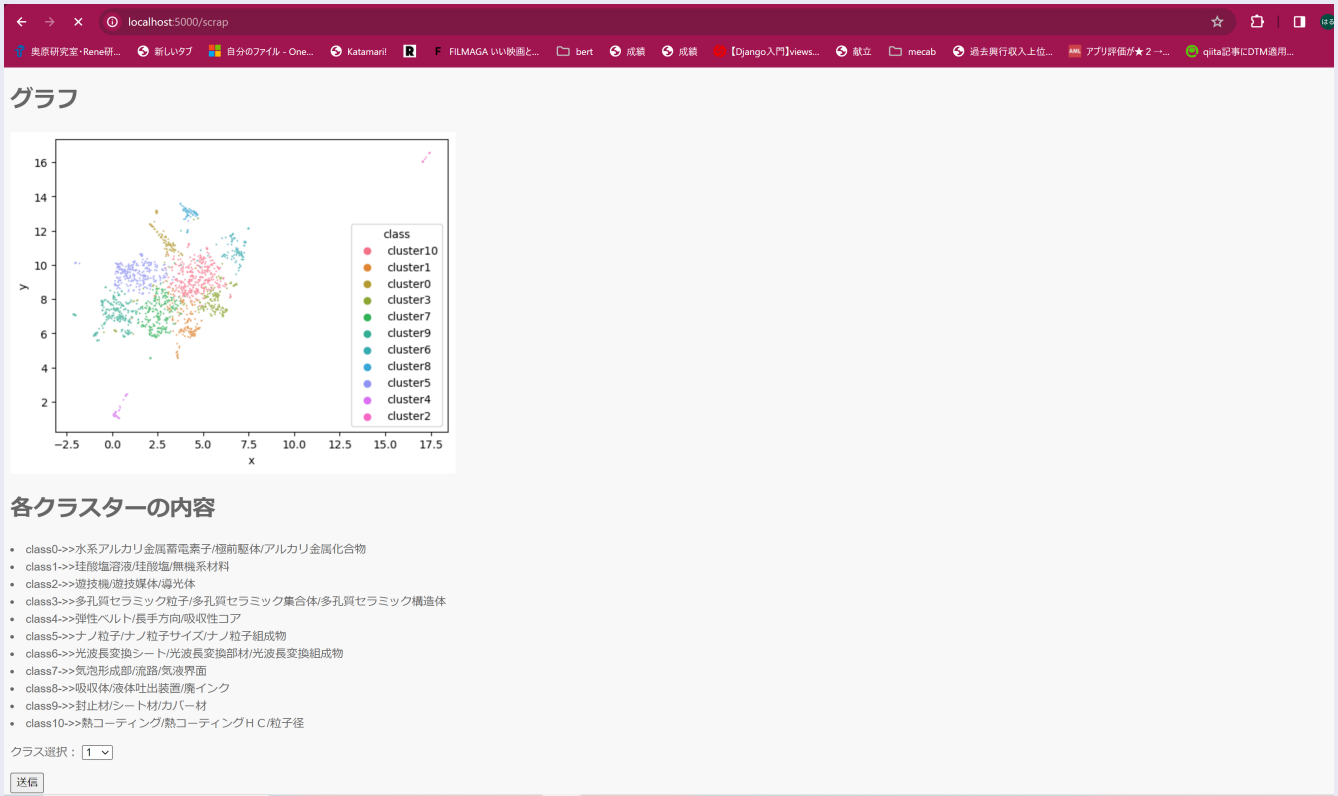

図 6: クラスター表示画面

システム

クラス選択画面

選択したクラスを 2d グラフまたは 3d グラフのどちらでも表示できるようにした.

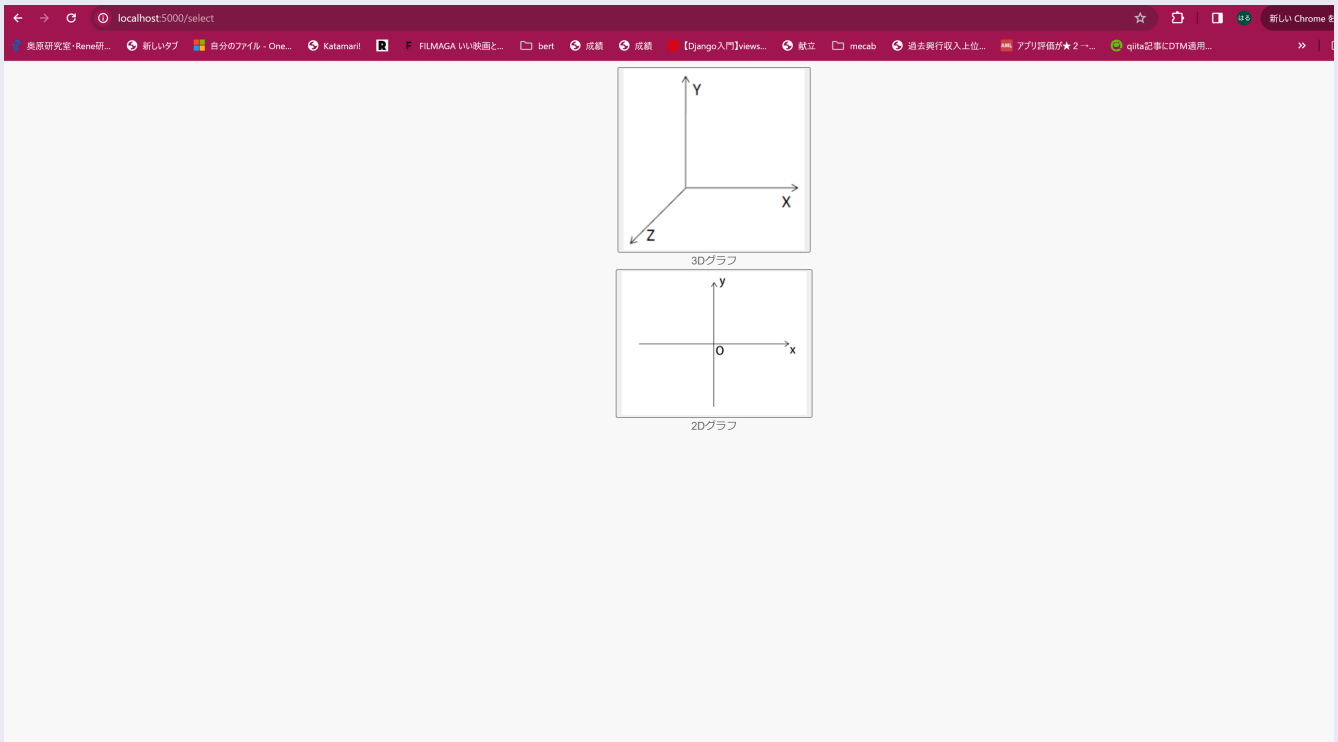

図 7: グラフ表示画面

システムの高速化

分かち書き

現在使用している Janome がとても時間がかかっていることが分かったので，他の手法を試してみた．
おそらく現時点で一番早い Vibrato を試してみた．

結果

動かすことはできたがユーザー辞書ができなかった．
新しすぎてネット上に情報が少なすぎる．

解決方法

Janome を並列処理で動かす．
Janome はユーザー辞書が登録しやすいというメリットがある．

まとめ

まとめ

- まだまだ時間がかかる処理が多いので少しでも早く処理できるように工夫を凝らしたい.
- ユーザーインターフェースを整える.
- 数値実験を行う.