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 12 月 1 日

はじめに

やったこと

- 専門用語や複合語の抽出を行った.
- スクレピング部分の並列化を行った.
- 以前の 3D グラフがあまり良くなかったので前よりも良い 3D グラフを作成できるようにした.
- システムとして使えるように flask を用いてアプリケーションを作成した.（途中まで）

termextract を用いた専門用語の抽出

termextract でできること.

専門用語を抽出して重要度を計算する.

→ 文章から専門用語を単体で取り出し重要度を計算することができる.

```
('体積 ホログラム 形成 層', 432415.3578615052),
('カード 基 材', 211965.10482224714),
('スクラッチ 隠蔽 層', 189256.60822842753),
('ホログラム', 128682.51629494972),
('体積 ホログラム', 111768.93966606639),
('ホログラム 層', 66150.24079657289),
```

図 1: termextract

termextract でできないこと.

専門用語を用いた分かち書き.

→ 「私 は 今日 学校 に 行き ます 。」のような分かち書きを行う必要があるがここまではできない.

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.

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 2: 辞書 (csv)

専門用語の抽出

次元圧縮とクラスタリング

辞書を用いて分かち書きを行い，単語の組み合わせを出力した.

```
[(('カード', '断面'), 81840),
 (('カード', '厚'), 81840),
 (('カード', 'ホログラム'), 80290),
 (('厚', '断面'), 69696),
 (('ホログラム', '断面'), 68376),
 (('ホログラム', '厚'), 68376),
 (('カード', '上記'), 58900),
 (('カード', '表面'), 57040),
 (('上記', '断面'), 50160),
 (('上記', '厚'), 50160),
 (('ホログラム', '上記'), 49210),
 (('断面', '表面'), 48576),
 (('厚', '表面'), 48576),
```

図 3: 抽出前

```
[(('カード基材', '体積ホログラム形成層'), 507144),
 (('スクラッチ隠蔽層', '体積ホログラム形成層'), 442816),
 (('カード基材', 'スクラッチ隠蔽層'), 401376),
 (('スクラッチ', '体積ホログラム形成層'), 338096),
 (('体積ホログラム形成層', '断面'), 335104),
 (('体積ホログラム形成層', '光'), 321640),
 (('カード基材', 'スクラッチ'), 306456),
 (('カード基材', '断面'), 303744),
 (('体積ホログラム形成層', '秘密情報'), 296208),
 (('カード', '体積ホログラム形成層'), 294712),
 (('カード基材', '光'), 291540),
 (('体積ホログラム形成層', '体積ホログラム形成層'), 279378),
 (('ホログラム', '体積ホログラム形成層'), 276760),
 (('カード基材', '秘密情報'), 268488),
```

図 4: 抽出後

クラスターの数がこれで正しいのかわからないので，クラスター分析を行った.

スクレイピングの並列化

並列化

今まではスクレピングにだいたい 1 時間半から 2 時間位かかっていたため並列化を行った.
並列にする数を大きくすれば大きくするほど時間が短くなると思ったが，10 スレッドよりも 5 スレッドの方が速く，スレッドの数を大きくしすぎると逆に時間がかかるということが分かった.
結果 20 分程度まで縮めることができた.

3D グラフ

シルエット分析

3D アニメーションなどを作成できる javascript のライブラリー「Three.js」を使って 3D グラフを作成することができた.
json ファイルを作れば 3D グラフにできる.

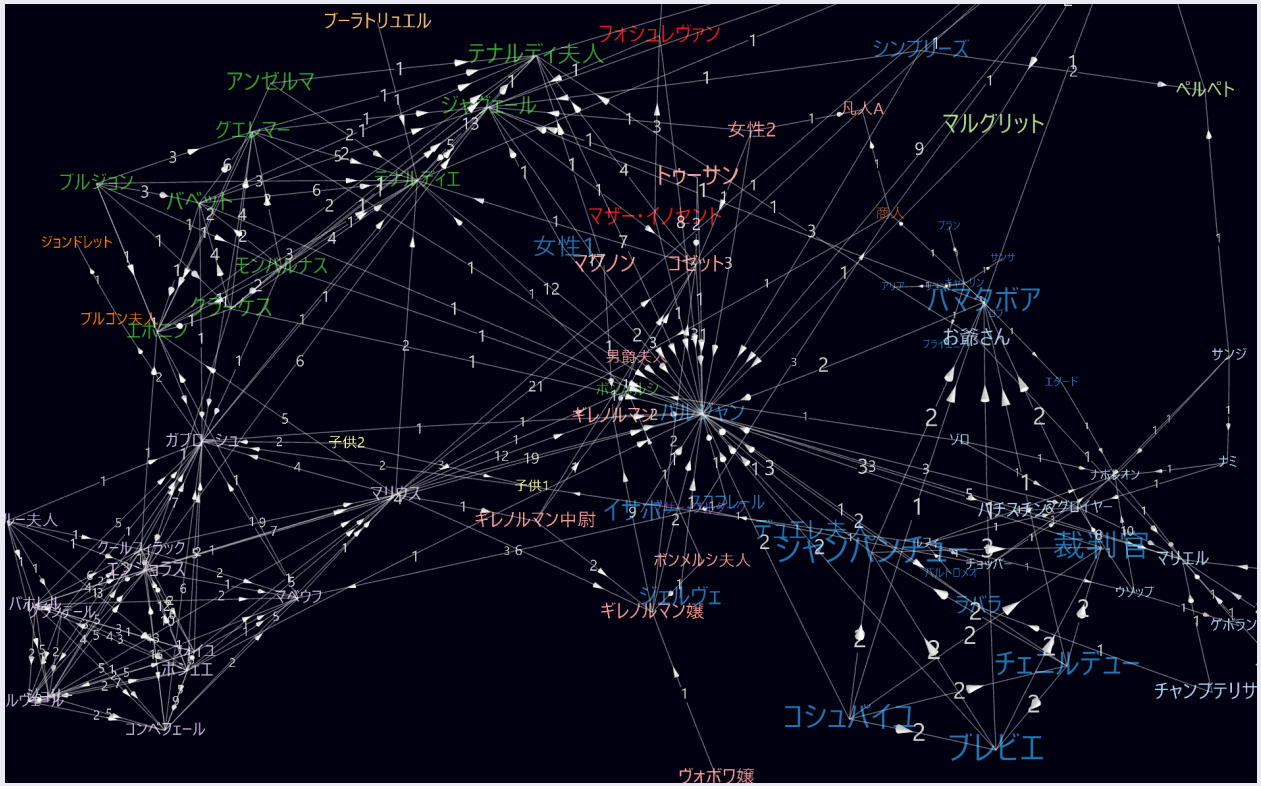

図 5: 3D グラフ

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 6: 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 7: ロード画面

共起語ネットワーク

クラスタリングをしたベクトルを表示することができた.

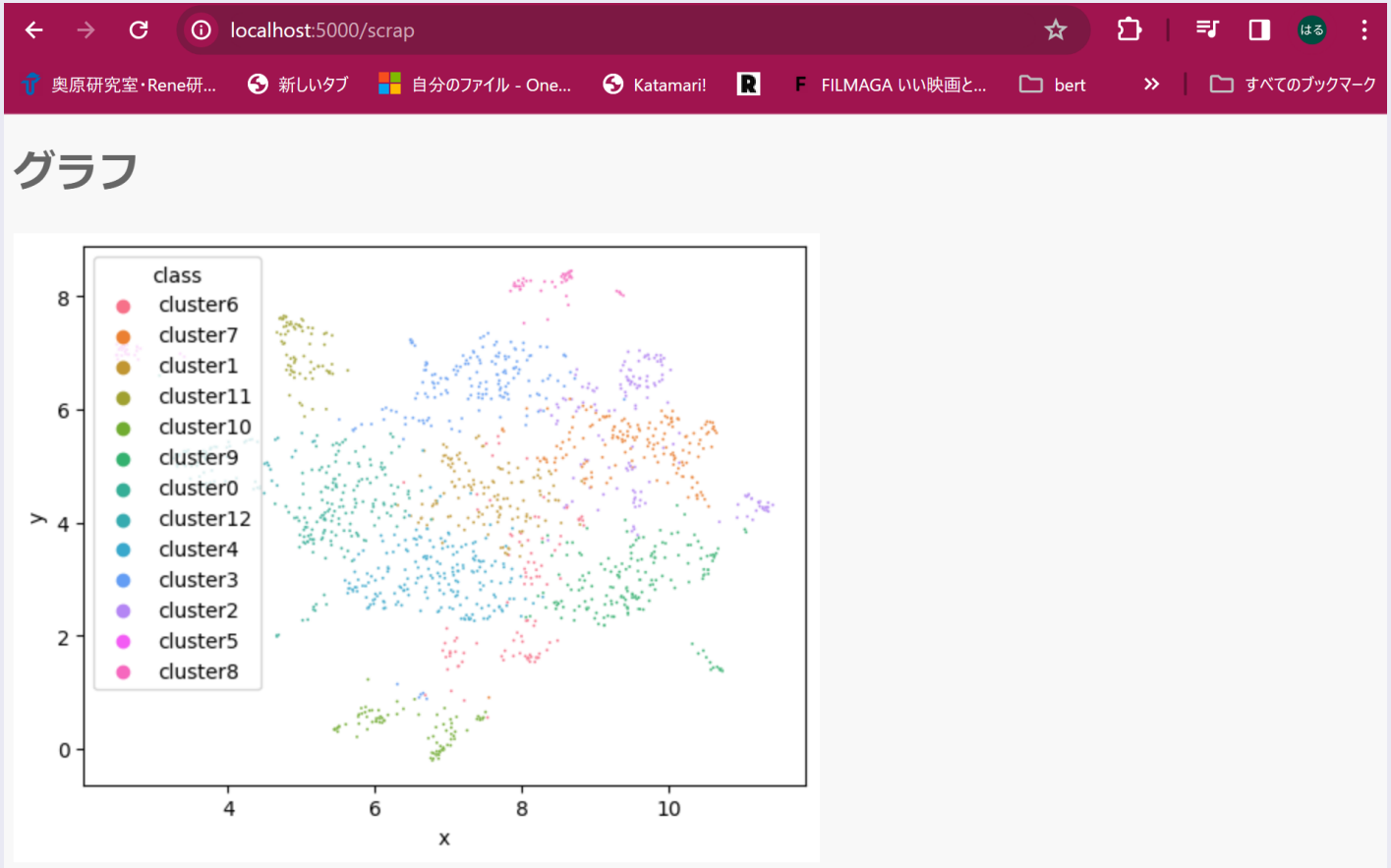

図 8: グラフ

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.
- グラフから先の部分をアプリケーションに組み込む.
- ユーザーインターフェースを整える.
- k-means よりも外れ値に強い k-medoid というクラスタリング手法があると最近知ったので実装したい.