アンケートの自由記述を Pythonでテキスト分析してみた

Dockerを使ったコンテナ環境での分析

秋山 浩杜
Hiroto Akiyama

概要

やったこと

- **VSCodeとDockerで環境に依存しにくい分析を実行した**
  - 私物PCが2台ある→ 仕事で一緒に分析することを想定
  - アナリストが使う分析開発環境と、マネージャーが使う確認用環境と想定し、環境に依存しにくい分析をDockerを用いて実行

- **テキスト分析の練習として、自治体に対するアンケートを分析**
  - ワードクラウド・共起ネットワーク・コレスポンデンス分析
  - 前処理や分析のロジックはメインの関心ではないだろうから割愛

今後の展望

- **自動化を組み込めば面白そう**
  - Daily Report: 決まった時間に自動で集計→ 回答割合・今日の特徴・前日比等を要約
  - Weekly Report: 決まった時間に自動集計 + Daily reportを結合してレポート作成

背景

- 「テキストデータの良い分析知らない？」とQ&Aがあった
  → テキストを分析できることは、ビジネス上のニーズがありそう
- **分析環境に依存しない形で、テキストデータを基礎的に集計・可視化できれば、面白そうだぞ…？**
  - 意思決定者は分析者ではない
    → 意思決定者の分析時間・労力を減らして、意思決定に悩む時間を作ろう
  - 個人の分析環境構築は面倒だし、共有がめんどくさい。
    しかも、これからチームで働く上でローカルで分析するのは非現実的
    → 分析環境に依存しない分析を

→ 2つの目標

1. **基礎的なテキストマイニングを実践しよう**
2. **コンテナを使って、ローカルでの分析を卒業しよう**

分析の目標

- **「地方公共団体へのオープンデータへの取組に関するアンケート」を用い、地方公共団体が抱える課題と、2018–2020年調査の年次差を評価する**
- 意義
  - (僕にとって) テキストデータの分析の練習
  - (社会的な意味として) 地方公共団体のオープンデータへの取組に関する課題をうまく解決できていないなら、内閣官房が取組をうまく進められていない根拠の1つとなる
- データソース
  - data.go.jp より、内閣官房「**オープンデータの取組に関する自治体アンケート結果**」を使用 (https://www.data.go.jp/data/dataset/cas_20170628_0004)
  - **2018年調査と2020年調査よりほぼ同じ質問項目を1つ選択し、内容・年次による差を比較**
    - 2018年 (No.31): オープンデータの提供・公開を進めるにあたり、貴団体として必要な情報や現在疑問に思われている点などがございましたら、自由にご記入ください。
    - 2020年 (No.35) : オープンデータの提供・公開・利活用を進めるにあたり、貴団体として必要な情報や支援、現在疑問に思われている点などがございましたらご自由にご記入ください。

仮想的な課題設定

- **設定（勝手な設定だけど、割とありそうな話）**
  - 上司:　秋山くん、アンケートの分析ってしたことある？
  - 秋山:　まあ、ありますよ。大学でもインターンでもあります。
  - 上司:　良かった！じゃあ、**2018年と2020年で2回評価したアンケートがあるんだけど、この2年で課題意識の自由記述に差があったか調べてくれる？もし差があったならどういう違いがあるかを可視化してくれると完璧かな。**
  - 秋山:　え、、、量的な回答じゃないんですか、、、？
  - 上司:　まあ、そんな高度な分析を求めてるわけじゃなくて、単純に集計する程度でいいからさ
  - 秋山:　あ、、、あ、、、いや、、、そういう話じゃなくて、、、
  - 上司:　大丈夫だよ、**先輩のXXさんも確認してくれるから気軽にやってね～**

→ 今回の分析の要件

テキストデータの分析
量的なデータではない！

＋

分析環境への配慮
自分だけが分析するわけではない！

分析練習の2つの目標

- **Dockerを使って、ローカルではない環境で分析を実行する**
  - ローカル環境で仕事が終わるのは、1人で仕事を完結させるときだけなはず
  - サーバー立てるほどの話ではないので、Dockerを使った環境構築を！

2つの環境で同じ結果を得る！

M1 Pro MacBook Pro: メイン環境
仕事をぶん投げられた秋山くん用環境と思えばok

Intel Mac mini: 確認用環境
一緒に仕事をする先輩の環境と想定

- **とりあえず簡単でいいからテキストデータの分析を実行する**
  - サンプルデータとして、「地方公共団体へのオープンデータの取組に関するアンケートを使用」（正直なんでもいい）
  - ワードクラウド + 共起ネットワーク + コレスポンデンス分析くらいでok

アンケートの自由記述を Pythonでテキスト分析してみた

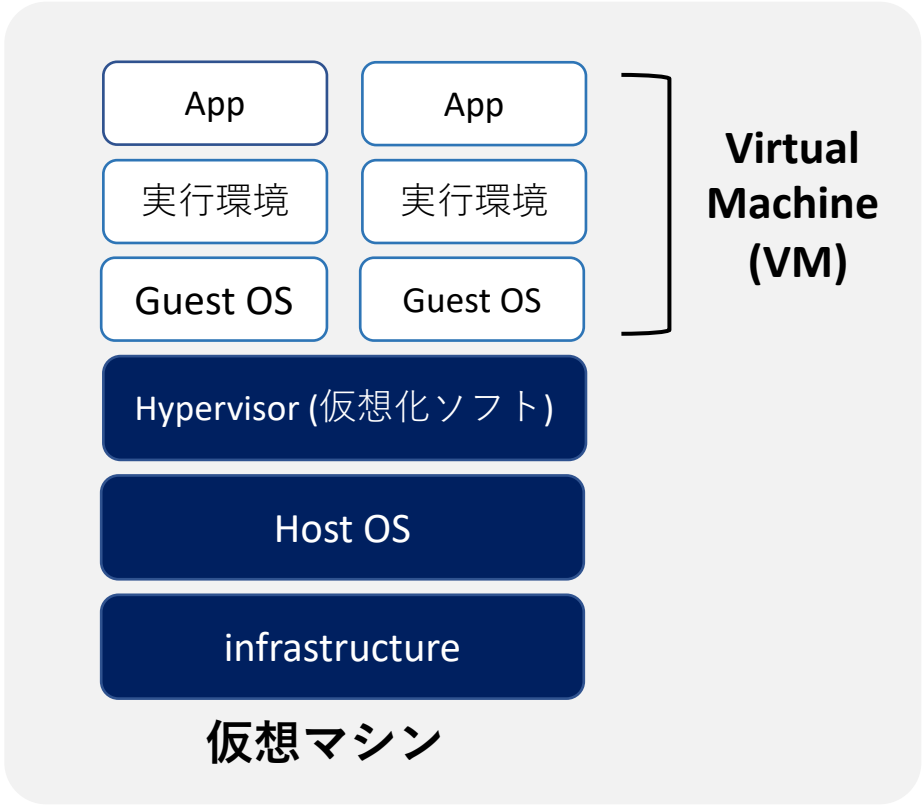
Dockerを使ったコンテナ環境での分析

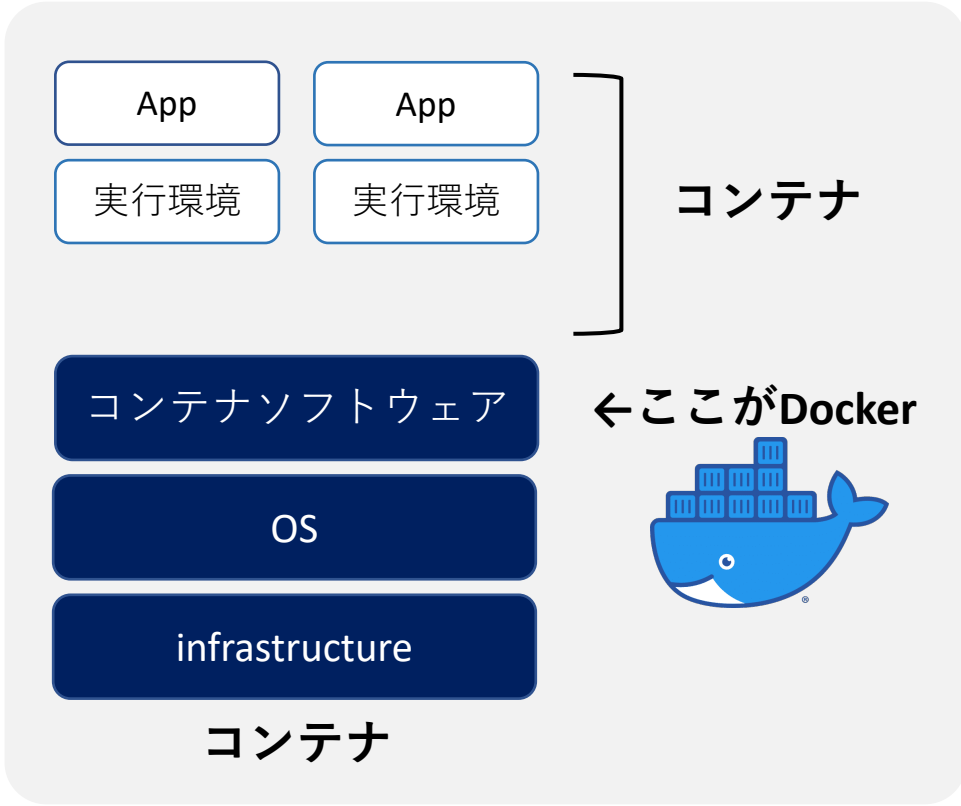
分析環境について

コンテナにより、軽量に分析・開発環境を標準化管理ができる

コンテナ: 1つのOSの上に作る仮想化した実行環境。軽い・早い・簡単！

従来の仮想マシン: Host OSの上にゲストOSを建てて作る仮想環境。安定・安全！





今回の分析環境

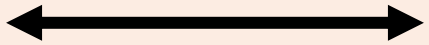
- **メイン分析環境: M1 MacBook Pro (2021)**
  - VSCode + Docker
  - Python (ver 3.10.4) ← Dockerfileに記載しているので、確認環境でも同じになる

- **確認用環境: Intel Mac mini (2018)**
  - Docker DesktopをインストールしVSCodeにRemote containersとDockerの拡張機能を入れたPC

仕事担当者のPC

M1 Pro MacBook Pro: メイン環境
仕事を実行する秋山くんの環境と思えばok

2つの環境で同じ結果を得る！



上司のPC

Intel Mac mini: 確認用環境
一緒に仕事をする先輩の環境と想定

今回の分析環境

- Dropboxでデータファイルの共有→ コンテナで分析環境を標準化

opendata_initiative

- opendata_initiative.code-workspace (VSCodeのプロジェクトファイル)
- .devcontainer (コンテナ作成用のフォルダ)
  - Dockerfile
  - devcontainer.json
- data (データファイルを入れるフォルダ)
- code (コードを入れるフォルダ)
- output (グラフ等の出力物を入れるフォルダ)
- slides (スライドを入れるフォルダ)

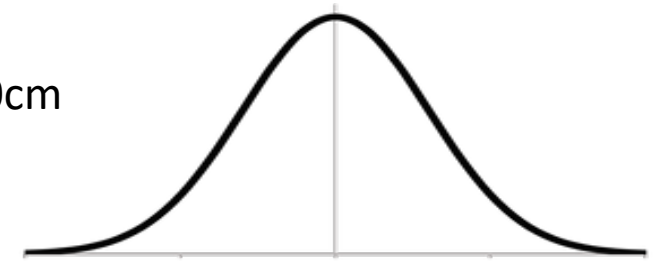
アンケートの自由記述を Pythonでテキスト分析してみた

Dockerを使ったコンテナ環境での分析

分析について

自然言語処理の難しさ: 解析対象が曖昧

- **数値の場合**
  - 身長を分析するとしよう
  - 身長が170cm→ 計測誤差はあるだろうけれど、170cmは170cm
  - 平均・差・比などの指標を解釈しやすい



みんな大好き正規分布

- **言語の場合**
  - 「美しい湖に立つ鳥」という意味は？美しい鳥なのか、美しい湖なのか？
    → 一意に定まらない
  - 「このセッションを受けて楽しかった」と「このセッションを受けてワクワクした」は意味的に同じか？ それとも異なるのか？
    → ネガティブ / ポジティブという点では、ポジティブで共通しているが…
    → 量的にデータを要約することの難しさ

今回の分析でまとめる3つのアウトプット

今回の分析でまとめるもの

1. **ワードクラウド: 単語の頻度を評価**
2. **共起ネットワーク: 単語と単語のつながりを評価**
3. **コレスポンデンス分析: 属性とテキストの内容との関連を評価**

今回の分析でまとめる3つのアウトプット

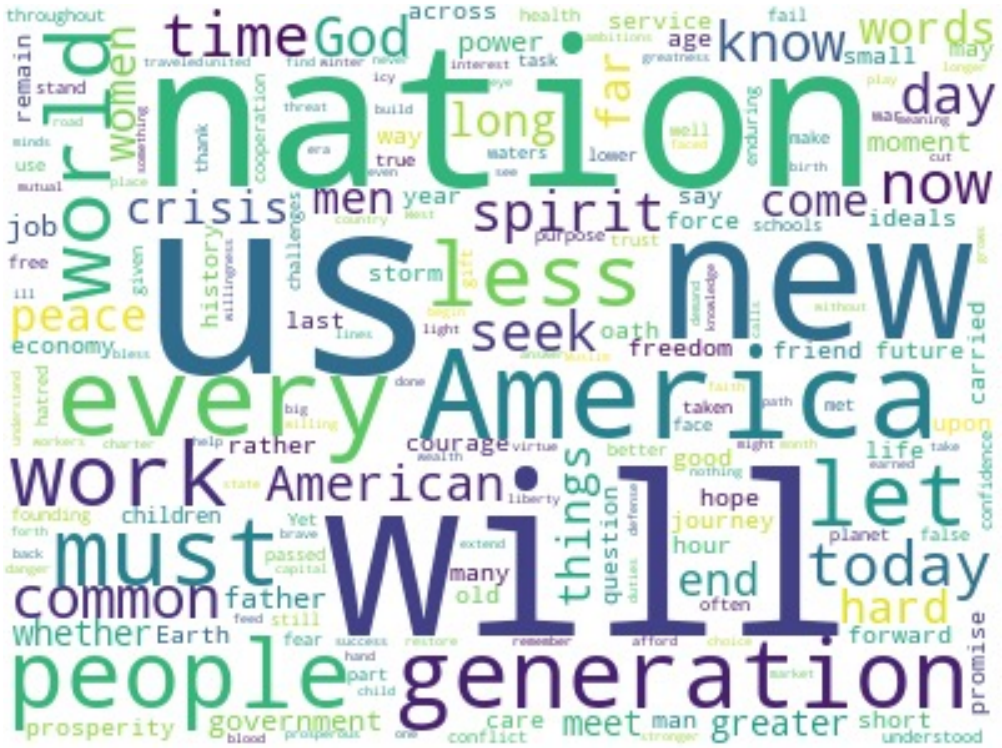
今回の分析でまとめるもの

1. **ワードクラウド: 単語の頻度を評価**
2. **共起ネットワーク: 単語と単語のつながりを評価**
3. **コレスポンデンス分析: 属性とテキストの内容との関連を評価**

1. ワードクラウド

単語の登場頻度を可視化したもの
（見たことある人も多いはず！）

頻度が高いほど単語が大きくなる



Obama元米国大統領の2009年1月のスピーチより秋山作成

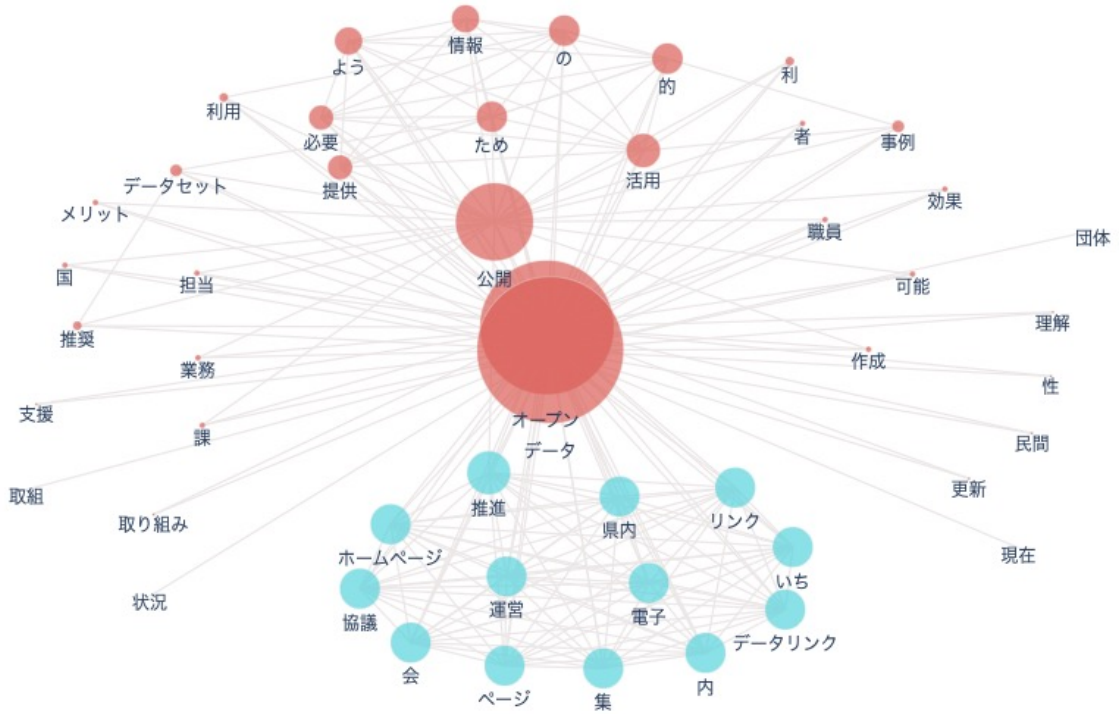
今回の分析でまとめる3つのアウトプット

今回の分析でまとめるもの

1. **ワードクラウド: 単語の頻度を評価**
2. **共起ネットワーク: 単語と単語のつながりを評価**
3. **コレスポンデンス分析: 属性とテキストの内容との関連を評価**

2. 共起ネットワーク

同時に使われやすい単語を調べる
→ コンテクストの可視化



「地方公共団体へのオープンデータへの取組に関するアンケート」より秋山作成

今回の分析でまとめる3つのアウトプット

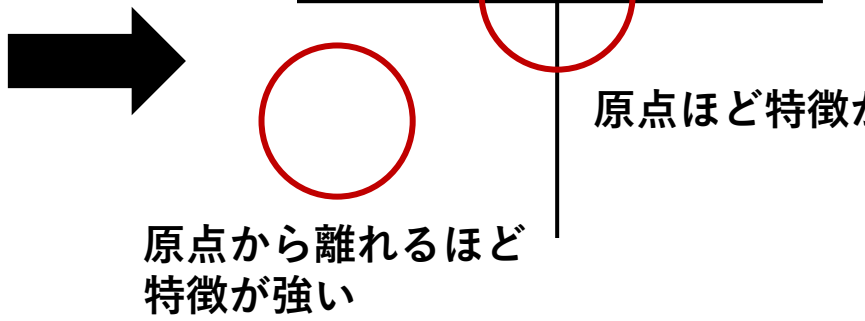
今回の分析でまとめるもの

1. **ワードクラウド: 単語の頻度を評価**
2. **共起ネットワーク: 単語と単語のつながりを評価**
3. **コレスポンデンス分析: 属性とテキストの内容との関連を評価**

3. コレスポンデンス分析

クロス集計表で集計できる行×列で表示されるデータを
散布図で表示し、変数間の関連を可視化する

例) パンの種類と味の感想	アンパン	食パン
甘い		
辛い		



クロス集計表の
横軸・縦軸の変数が
近くにプロットされる
→ 変数間の関連が強い

→ きっと「アンパン」と
「甘い」は近くなる

原点ほど特徴が平均的

原点から離れるほど
特徴が強い

分析結果: 名詞と形容詞のワードクラウドで2018/2020の単語を比較

分析対象: 2018・2020の「地方公共団体へのオープンデータへの取組に関するアンケート」より、必要な情報や支援、疑問に関する自由記述

2018と2020年で登場する単語が変わった様子はない

2018 + 2020 (406コメント)

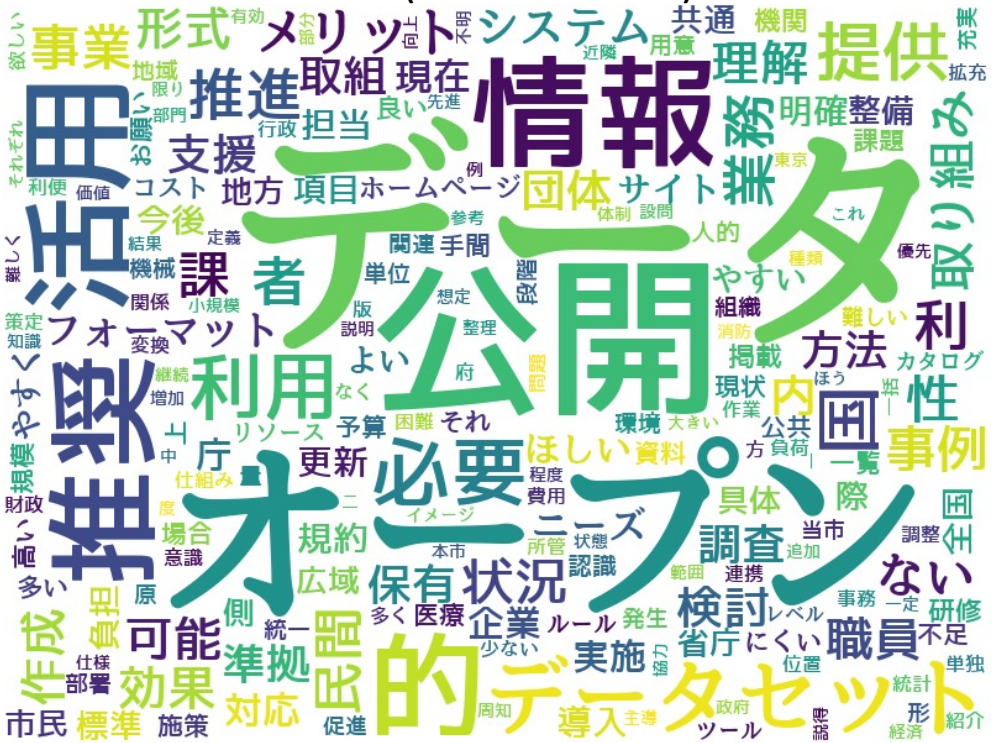
FYI: 前後でワードクラウドが変わった例 [地域再生計画をテキストマイニングして地方創生分析]

分析結果: 名詞と形容詞のワードクラウドで2018/2020の単語を比較

分析対象: 2018・2020の「地方公共団体へのオープンデータへの取組に関するアンケート」より、
必要な情報や支援、疑問に関する自由記述

2018と2020年で登場する単語が変わった様子はない

2018 (177コメント)



2020 (229コメント)

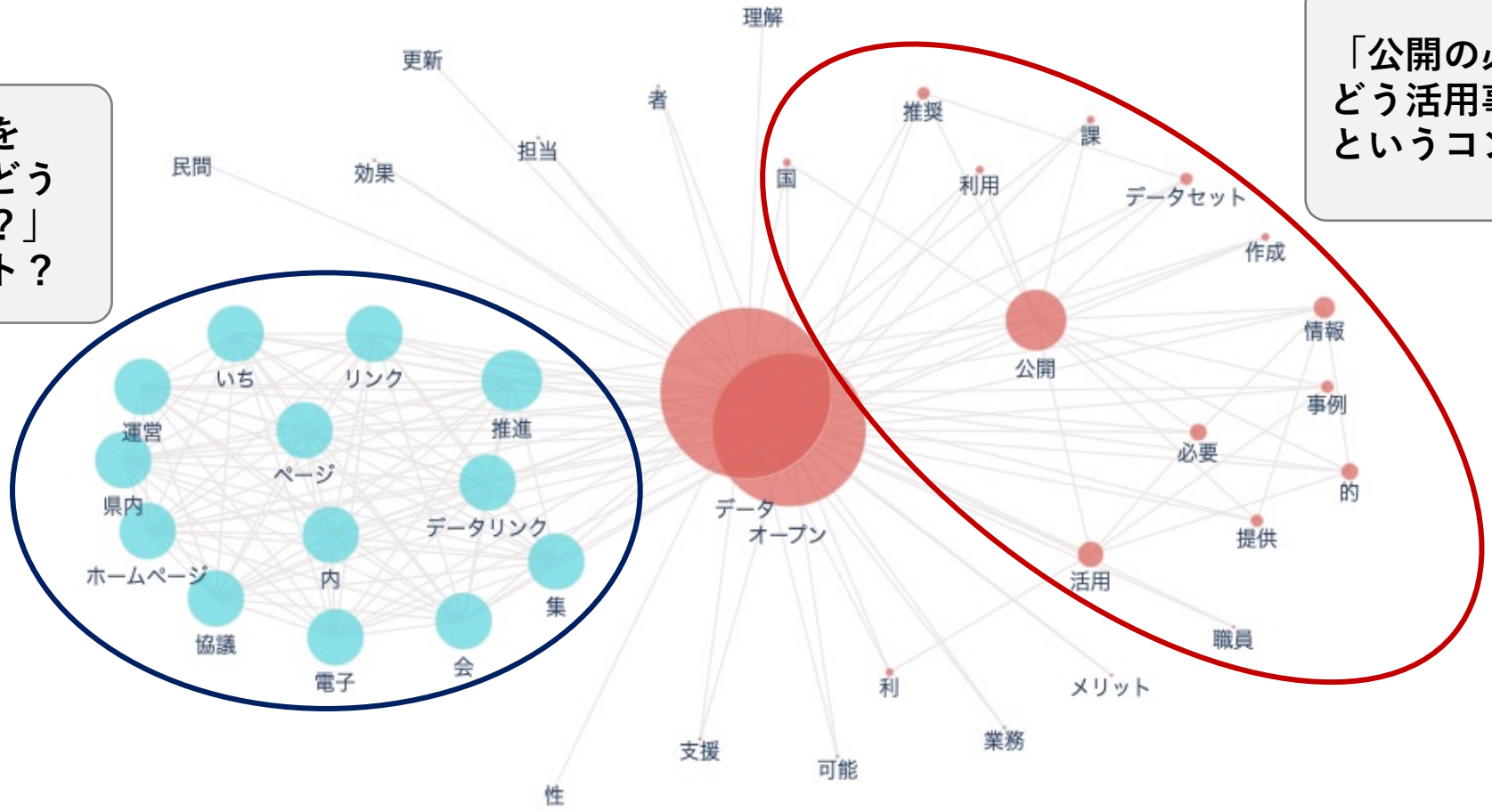
FYI: 前後でワードクラウドが変わった例 地域再生計画をテキストマイニングして地方創生分析

分析結果: 共起ネットワークで文脈を可視化

分析対象: 2018・2020の「地方公共団体へのオープンデータへの取組に関するアンケート」より、
必要な情報や支援、疑問に関する自由記述

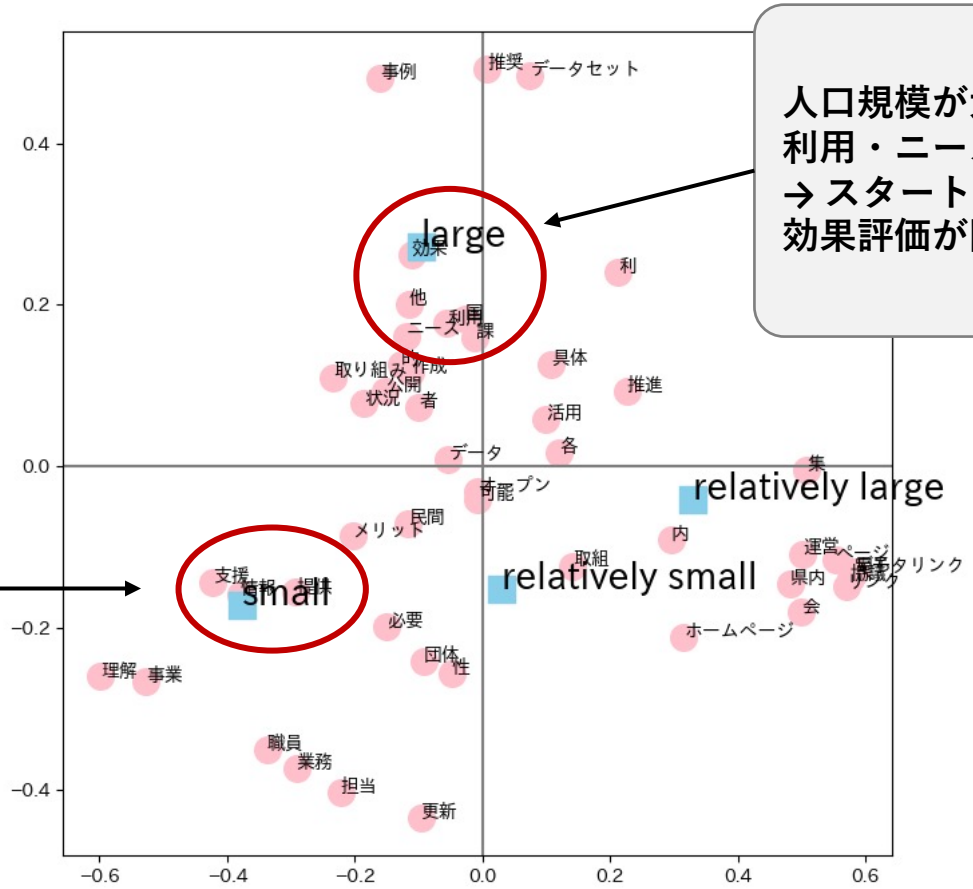
「オープンデータを掲載するページをどう運営していくのか？」というコンテクスト？

「公開の必要性を疑問視 / どう活用事例するの？」というコンテクスト？



分析結果: コレスポンデンス分析で人口とコメント内容の関連評価

市区町村の人口とコメントの内容の間の関連を評価（都道府県のコメントは除外、名詞頻度を評価）
→ わかること:「人口の多さによって課題って変わるのかな？」



人口規模が大きな自治体では
利用・ニーズ・効果等と関連強い
→ スタートしたものの、活用や効果評価が困難？

人口規模が小さな自治体では、
情報・支援・提供との関連が強い
↓
そもそもスタートできていない？

はじめてテキストの分析をやってみた感想

本に書いていることをなぞるのではなく、自分で ”Question設定→ 実践” をしたのは初

- **結果がなんかぼんやりしてない？**
  - 解釈が困難だと感じた。
  - 結局、ドメイン知識を持った上で臨まないと、よくわからない
  - もっと刺さる分析とは？を学ぶ必要がある

- **前処理がかなり難しい**
  - 今回、それほど前処理はしていない→ 分析上大きな問題
  - ストップワードは選んだ
  - かなり雑な解析をしてしまったので、今後は前処理も含めて勉強すべき

- **データハンドリングで少し悩んだ**
  - 目標とする分析に合わせてデータ構造を作るんだよって言ってきたけど、
    自分が初学者になると「言うは易く行うは難し」

課題と今後の勉強方針

- **Pythonよくわからん！！！**
  - 「この行のコードはどういう意味があって、このコードが欠けるとどうまずいか？」がまだよくわかっていない→ もっと書いてもっと作ろう

- **今回は集計レベルの分析しかしていないので、より意思決定に資する分析の勉強を**
  - 正規表現とか単語分散表現とか理解できていない→ 勉強を
  - なにかアクションを促すための分析をできるようになる必要がある
    → テキストを使った分析で意思決定者に刺さる分析とは？？？

- **自動化機能の搭載**
  - 毎日集計されるレポートの場合はどうすべき？
    → 前日比や前週比レポートを自動で出力するプログラムを開発したい

- **Pythonではない言語での開発**
  - 基幹システムに合わせた開発が必要

THANK YOU !

アンケートの自由記述を Pythonでテキスト分析してみた

Dockerを使ったコンテナ環境での分析

秋山 浩杜
Hiroto Akiyama