テキストから情報を抽出する

テキスト分析(1)

社会システム科学（B-4）

テキスト分析とは？

テキスト分析とは？

- **テキストデータ**を収集・分析して有用な情報や知識を取り出す
  ≒テキストマイニング ← データマイニングのテキスト版

テキストデータとは？

- コンピュータ上の**文字のみ**のデータ
  - → 装飾情報を含まない
- **自然言語**（⇔ 人工言語（特にコンピュータ言語））

テキストデータの特徴

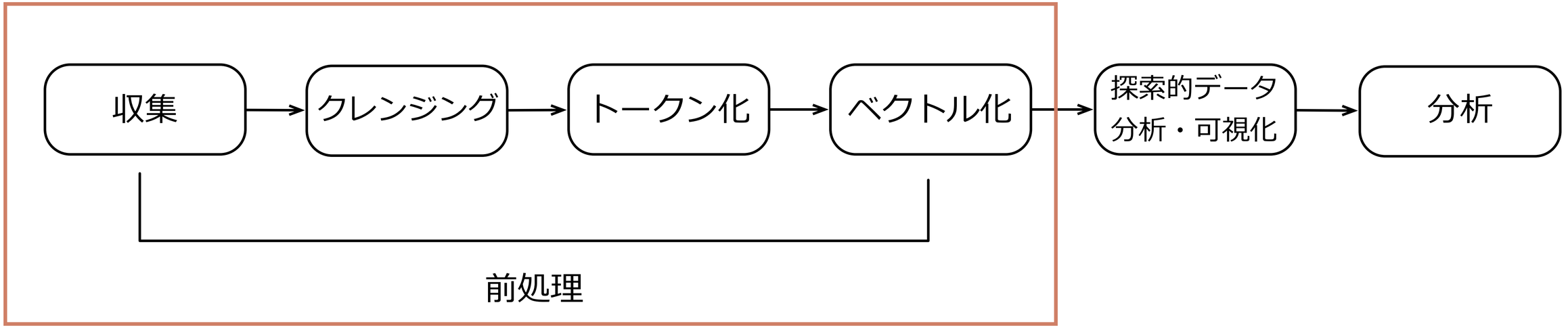
- 離散的（⇔ 連続的）
- 系列データ（sequential）
- 文法によるデータ間の強い制約
- 不定長
- 曖昧

テキスト分析の応用

- 機械翻訳
- パーソナルアシスタント（音声認識・音声合成・会話生成）
- 検索エンジン
- 日本語入力（IME）
- スパムメールフィルタ
- OCR

テキスト分析の手順

テキスト分析の手順



今日はこの辺をします

テキストデータの収集

- コーパスの利用
- 印刷された文書のテキストデータ化（OCR）
- 電子メール
- アンケート
- Webサイト上のテキストデータ → **スクレイピング**
  - ソーシャルメディア（SNS，CGM）

ウェブスクレイピング（Web Scraping）

- Webサイトからデータ（テキストなど）を抽出すること

［狭義］

- 特に**プログラムを使用**して行う（手動ではなく自動）

［注意点］

- Webサイトの負荷 → 不正アクセスと見られる可能性もある。
- API（Application Program Interface）がある場合はそちらを利用する

テキストクレンジング

- クレンジング
  - 不要なデータの除去
  - 破損・欠損データの特定と修復
- テキストクレンジング＝テキストデータ特有の処理
  - 記号・ルビ / URL / HTMLタグ / 編集記号などの除去
  - 表記揺れの処理
- サニタイズ
  - 機密情報や個人情報の除去

トークン化

- テキストデータをトークン（token）という小さな単位に分割
- 英語などのように分かち書きされた言語 → 空白で分割
- 日本語や中国語のように単語の区切りが明示されない言語
  - N-gram
  - 形態素解析

N-gram

・ N文字単位で文書を機械的に分割する方法

［例］

神戸は良い天気です。

・ ユニグラム（N=1）

神/戸/は/良/い/天/気/で/す/。

・ バイグラム（N=2）

神戸/戸は/は良/良い/い天/天気/気で/です/す。

・ トリグラム（N=3）

神戸は/戸は良/は良い/良い天/い天気/天気で/気です/です。

形態素解析

→ 意味のある最小単位のトークン

- 自然言語の文を**形態素**に分割
- 形態素の品詞・読み・原形などを求める

お待ちしております。

文字列	品詞	品詞の種類	活用の種類	活用形	原形	読み
お待ち	名詞	サ変接続			お待ち	オマチ
し	動詞	自立	サ変・スル	連用形	する	シ
て	助詞	接続助詞			て	テ
おり	動詞	非自立	五段・ラ行	連用形	おる	オリ
ます	助動詞		特殊・マス	基本形	ます	マス
。	記号	句点			。	。

MeCab 0.996の結果

トークン化に伴うその他の処理

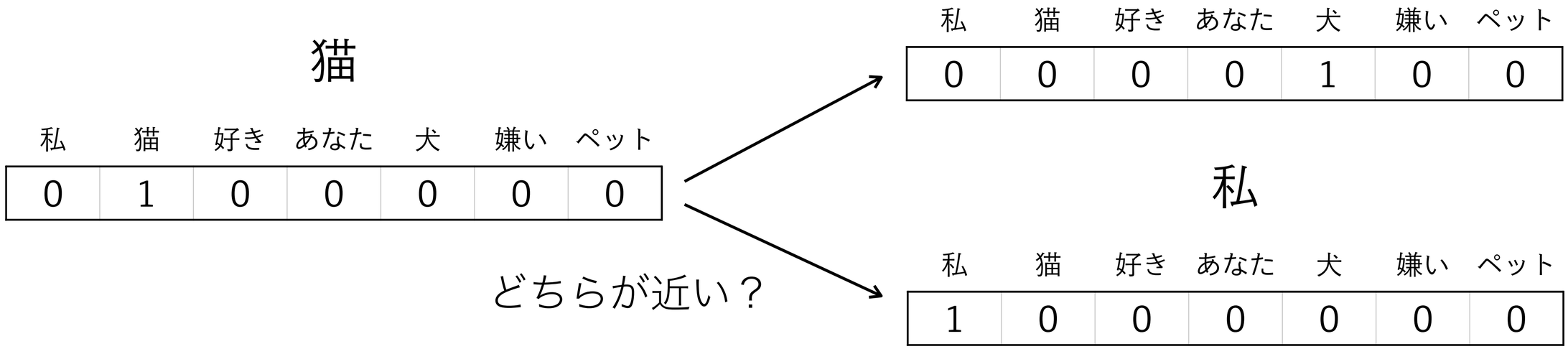
- 照応解析
  - 指示代名詞や省略された語の処理
- 固有表現認識
- 依存構造解析とチャンキング

ベクトル化

- トークン単位のベクトル化
  - O-hot表現
  - Word2Vec
- テキストデータ単位のベクトル化
  - 特徴量ベクトル
  - Bag-of-Words (BoW)
  - TF-IDF
  - Doc2Vec
  - 潜在トピックモデル (LDA)

One-hot表現

- One-hot表現＝1つだけ「1」で残りが「0」でのビット列による表現
- トークンのone-hot表現
  - 各ビットがトークンに対応するビット列による表現
  - 「1」になっているビットで単語を表現する
- 問題点：トークン間の距離（近さ）の表現が困難



Word2Vec

- テキストデータ中での使われ方（コンテキスト）に応じたベクトル
  → 似たような使われ方をするトークンは類似したベクトルで表現

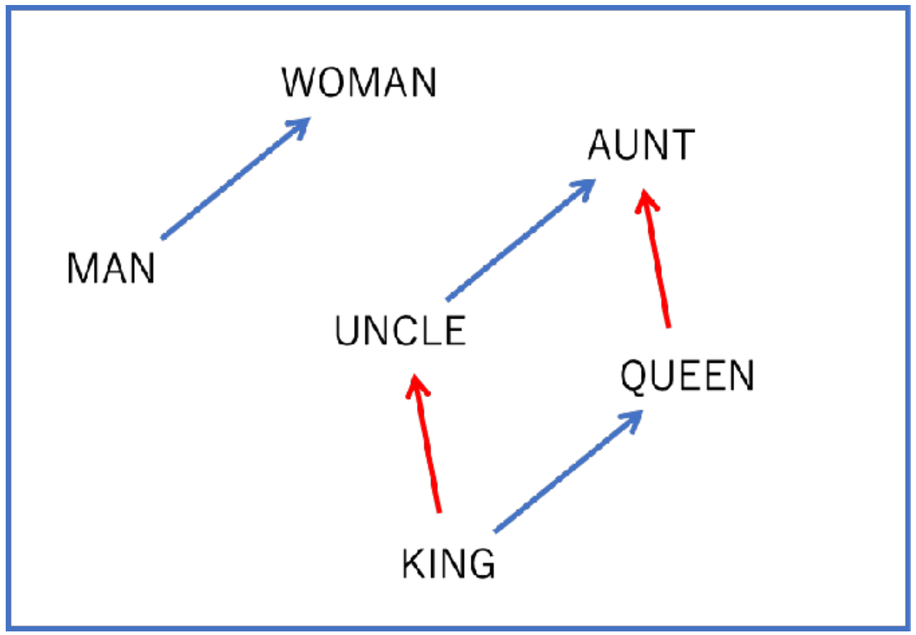
[例]

[方法]

- ニューラルネットワークを使って学習により獲得

Word2Vecの特徴

- トークン間の相対的な関係が保持される。
- トークン間の演算ができる。



KING - QUEEN + UNCLE → AUNT

- どうしたら「彼女」から「奥さん」になれるかを「Word2Vec」に聞いてみた
  https://ainow.ai/2017/10/31/124408

［演習］テキスト分析をやってみる（前半）

［演習］テキスト分析の前半をやってみる（Word2Vecまで）

- ここからは Google Colaboratory で作業します。
- ノートブックの説明を見ながら解説します。
- ノートブックの指示に応じてこちらの資料に戻って参照します。

［補足1］形態素解析パッケージ

Pythonで利用できる形態素解析ツールの種類

- スタンドアロン型
  - C/C++言語などで書かれている（場合が多い）
  - 単独で利用できる
  - 比較的高速
  - OSにインストール＋Pythonからはラッパーを介して利用
- ネイティブライブラリ
  - Pythonで書かれている
  - Pythonでしか利用できない（場合が多い）
  - インストールが簡単（Pythonの標準的な方法でインストールできる）
  - 比較的低速

Pythonで利用できる形態素解析ツール

- スタンドアロン型
  - MeCab
    - よく使われており辞書の種類も多い
    - 比較的高速
  - GiNZA
    - 新しいツールで精度が高い。
    - 係り受け解析もできる。
- ネイティブライブラリ
  - Janome ← 今回はこれを使います

Janomeのインストール

1. Google Colabのコードセルで以下を実行

```
!pip install janome
```

サーバ上でコマンドを実行する

2. 実行されてインストールされる。

```
[10]  1  !pip install janome

      Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
      Collecting janome
        Downloading Janome-0.4.2-py2.py3-none-any.whl (19.7 MB)
           |████████████████████████████████| 19.7 MB 1.3 MB/s
      Installing collected packages: janome
      Successfully installed janome-0.4.2
```