日本語小説分析システム

BERTopic・c-TF-IDF・spaCyによる テキスト分析技術

システム概要

- **日本語のテキストファイル**を分析し、**テーマ**と**固有表現**を抽出するシステム
- 3つのAI技術を順次実行
  - BERTopic：**トピックモデリング（テーマ抽出）**
    （c-TF-IDF：**トピック**を**代表する単語**を抽出）を含む
  - spaCy ja_core_news_lg：**固有表現抽出（人名・地名等）**
- 利用シーン
  - 小説の内容把握、複数の作品の比較分析、登場人物・地名の出現頻度分析

トピックモデリング

- 大量の文書から潜在的なトピック（テーマ）を自動的に発見する技術

- 各文書や文書内の各文が、どのトピック（テーマ）に属するか、各トピック（テーマ）がどんな単語で構成されるかを分析

- 文書や文書内の文をグループ化

BERTopic：トピックモデリングの応用例

- **カスタマーサポート**
  評価、改善要望等の把握
- **ニュース分析**
  トレンド、テーマの抽出
- **リサーチ**
  研究分野、トピックのリサーチ
- **アンケート分析**
  要望等をテーマ別整理
- SNS分析
  商品、関心事の検知

BERTopic

- BERTopic
  - Grootendorst（2022年）により提案された**トピックモデリング手法**
  - 教師なし学習でトピックを発見
  - トピック数の事前指定は不要
- 処理の流れ
  1. Sentence-BERT埋め込み：**文章を384次元ベクトルに変換**
  2. UMAP（次元削減手法）：**384次元を5次元に圧縮**
  3. HDBSCAN（密度ベースクラスタリング）：**文章をグループ化**
  4. Bag-of-Words（単語の出現回数）：**単語の出現頻度を集計**
  5. **c-TF-IDF の実行：トピック**を**代表する単語**を抽出

c-TF-IDF（Class-based TF-IDF）

- c-TF-IDFとは
  - BERTopicの内部で使用される**重み付けアルゴリズム**
  - 従来のTF-IDFを拡張し、**トピック単位で単語の重要度を計算**
- 計算式c-TF-IDF(x,c) = tf(x,c) × log(1 + A/freq(x))
  - tf(x,c)：**トピックc内での単語xの頻度**（L1正規化済み：合計を1に調整）
  - A：全トピックの**平均単語数**
  - freq(x)：全トピックにおける単語xの**出現頻度**
- 使用モデル
  - paraphrase-multilingual-MiniLM-L12-v2（50+言語対応、384次元）

固有表現抽出（NER）とは？

テキストから**重要な情報を自動抽出**

- **固有表現 = 特定の人・場所・時間などの 固有名詞**

抽出対象

- **人名**：伊藤首相、田中部長
- **組織名**：トヨタ自動車、首相官邸
- **地名**：東京、渋谷駅
- **日付**：15日、2025年10月
- **金額**：1000円、5億ドル

手作業では時間がかかる作業を AIが短時間で処理

固有表現抽出（NER）の応用

- 顧客サポート

「顧客からの投稿」から製品名・日付を自動抽出

- 文書理解支援

契約書から当事者・期日・金額を抽出

- マーケティング

SNSから企業名・製品名のメンションを収集

- リサーチ

論文、研究者名、機関名を整理

固有表現抽出（spaCy）

- **spaCy ja_core_news_lgの概要**
  - Explosion AIが開発した日本語大規模モデル
  - 統計的機械学習ベースのNER（**固有表現抽出**）
  - Universal Dependencies Japanese GSD コーパスで学習
- **モデル性能**
  - **品詞タグ付け**精度：約97-98%
  - **NER（固有表現抽出）**のF1スコア（精度評価指標）：約71.2%
- **抽出可能な固有表現**
  - PERSON（人名）、ORG（組織名）、GPE（地政学的実体：国・都市など）、LOC（場所）、DATE（日付）、TIME（時刻）、MONEY（金額）、PERCENT（パーセンテージ）