
人文情報学 Digital Humanities

九州大学システム情報科学研究所 藤田 郁

2025年8月20 – 22日



人文情報学 スケジュール 第1日目

1. 人文情報学とはどのような分野なのか？

1. 「情報学」とはどのような分野なのか？
2. 従来の「人文学」研究はどのようなものか
3. 人文学と情報学を掛け合わせることの意義
4. 人文情報学で行われている研究やその成果1 公開されているデータベースの紹介

2. データベースとアノテーション

3. テクスト分析演習 1 (オンラインツール)
4. テクスト分析演習 2 (前処理)
5. テクスト分析演習 3 (前処理)

人文情報学 スケジュール 第1日目

1. 人文情報学とはどのような分野なのか？
- 2. データベースとアノテーション**
 1. 1-4で紹介したデータベースの検索機能などを実際に使用してみる。
 2. アノテーション, メタデータについて
- 3. テクスト分析演習 1 (オンラインツール)**
 1. 既存のオンラインツールを使ったテキスト分析
 2. テクスト分析でよく使われる用語や指標 (述べ語数, 異なり語数, ドキュメントの長さ, 語彙密度, 平均文長, 読み易さの指標, 特徴語, n-gram, 共起等)
4. テクスト分析演習 2 (前処理)
5. テクスト分析演習 3 (前処理)

人文情報学 スケジュール 第1日目

1. 人文情報学とはどのような分野なのか？
2. データベースとアノテーション
3. テクスト分析演習 1 (オンラインツール)
- 4. テクスト分析演習 2 (前処理)**
 1. 分析対象を電子データ化し, 保存する。 (英語)
 2. 分析対象を電子データ化し, 保存する。 (日本語)
 3. 正規表現
- 5. テクスト分析演習 3 (前処理)**
 1. 分かち書き・形態素解析 (日本語)
 2. ステミング (レマ化) (英語, 日本語)

人文情報学 スケジュール 第2日目

1. テクスト分析演習 4 (前処理)

1. 品詞タグとはなにか
2. 品詞タグ付け (英語・日本語)

2. テクスト分析演習 5 (頻度)

1. 既存の大規模コーパスで頻度を調べる
2. 素頻度と相対頻度

3. テクスト分析演習 6 (頻度)

4. テクスト分析演習 7 (頻度)

5. テクスト分析演習 8 (視覚化)

人文情報学 スケジュール 第2日目

1. テクスト分析演習 4 (前処理)
2. テクスト分析演習 5 (頻度)
- 3. テクスト分析演習 6 (頻度)**
 1. ジャンル (レジスター) 間の比較 : マン・ホイットニーのU検定, 対数尤度比, MI-Score
 2. TF-IDF
- 4. テクスト分析演習 7 (頻度)**
 1. 頻度表を作成 (英語・日本語)
 2. 作成した頻度表を考察する (記号の扱い, 機能語と内容語, ストップワード)
- 5. テクスト分析演習 8 (視覚化)**
 1. ワードクラウド
 2. ヒートマップ
 3. 共起ネットワーク

人文情報学 スケジュール 第3日目

1. テクスト分析演習 9 (機械学習)

1. 階層的クラスタリング

2. テクスト分析演習 10 (機械学習)

1. コレスポネンス分析

3. テクスト分析演習 11 (機械学習)

1. トピックモデル

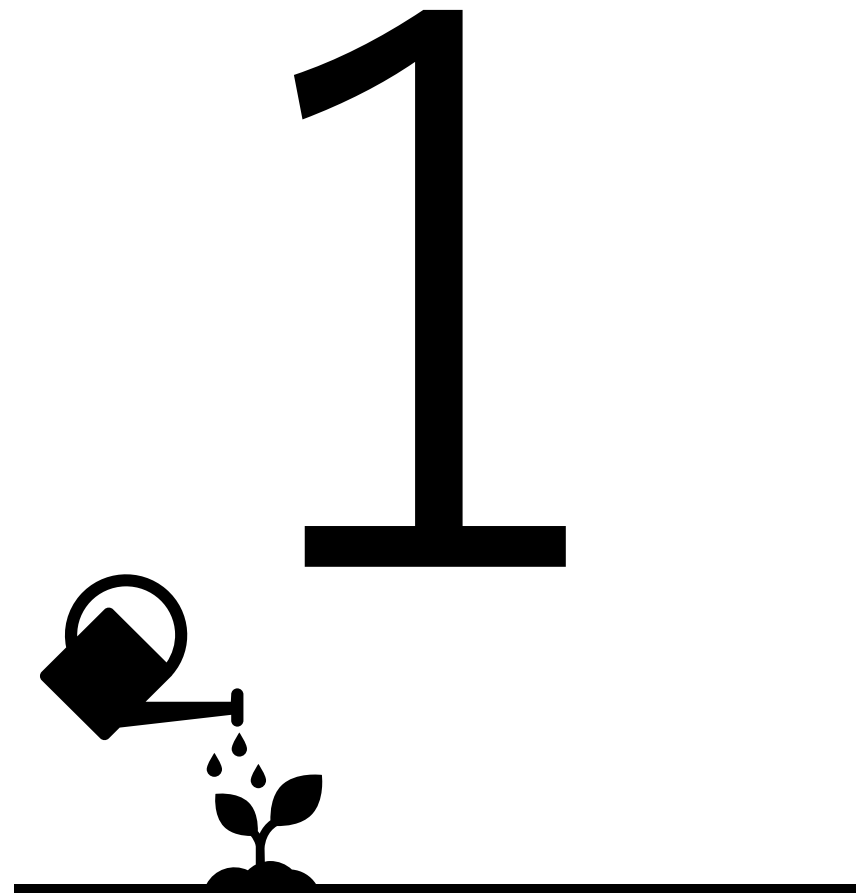
~~4. テクスト分析演習 12 (生成AI/LLM)~~

- ~~1. LLMを使ったテキスト生成/分析~~

5. まとめ

1. 人文情報学と従来の人文学の融合

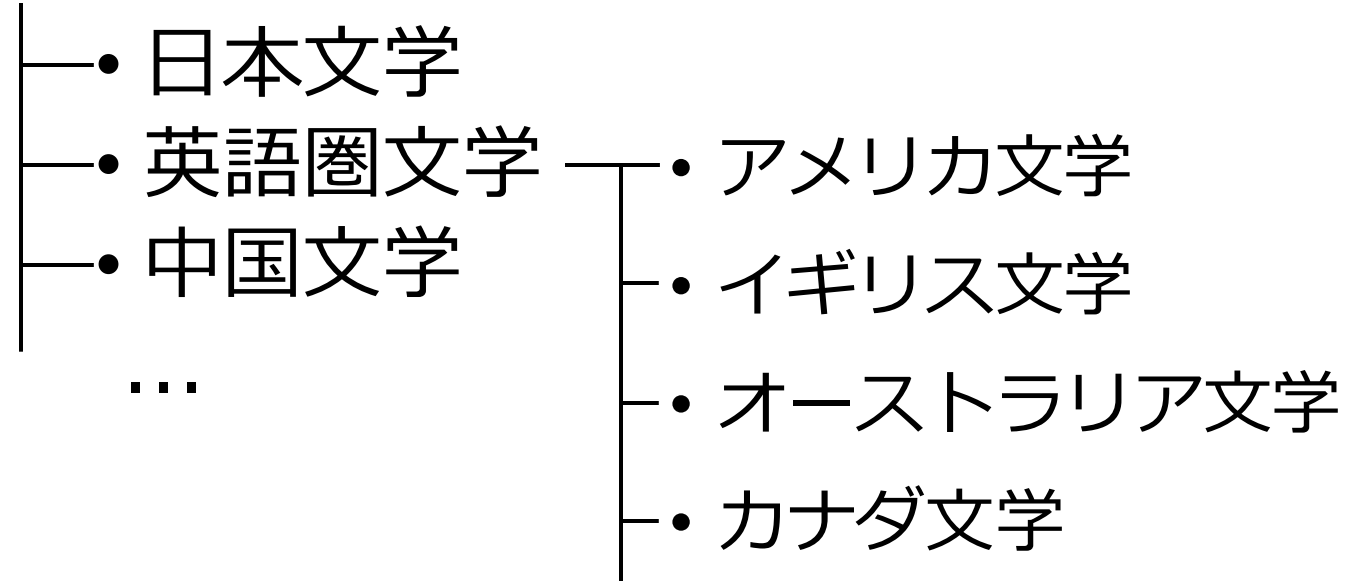
人文情報学とは
どのような分野
なのか？





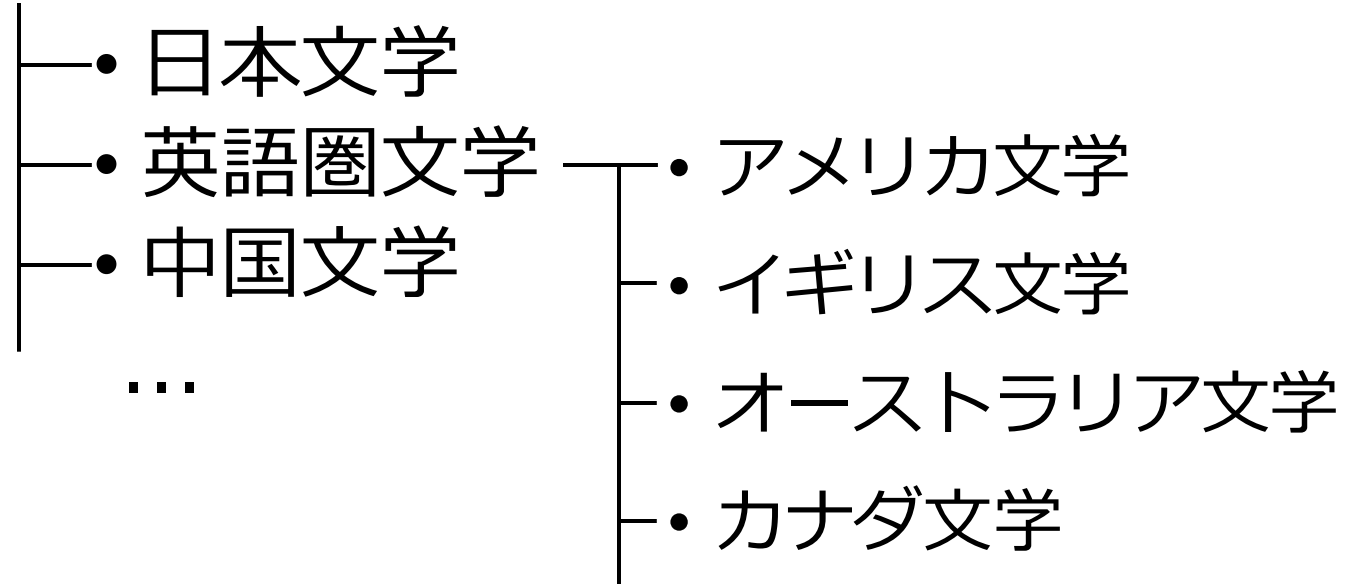
1. 人文情報学とはどのような分野なのか？

- 人文学と情報学を掛け合わせたのが「人文情報学」
- Digital **Humanities**
- 情報学の手法や知見を人文学研究に応用
- 人文学：歴史・**文学**・哲学・宗教・芸術等





- 人文学と情報学を掛け合わせたのが「人文情報学」
- Digital **Humanities**
- 情報学の手法や知見を人文学研究に応用
- 人文学：歴史・**文学**・哲学・宗教・芸術等



1. 人文情報学とはどのような分野なのか？

1. 「情報学」とはどのような分野なのか？

- 「情報に関するあらゆる学問領域をカバーする学問」
- 「情報の発生, 獲得, 表現, 蓄積, 流通, 検索活用な等情報処理の全ての家庭における理論から応用, 人間と社会との関わりまで追求する, ロジック, コンピューティング, 概念形成, システム, あるいはコンテンツを含む応用など, 幅広い情報に関する学問領域」
- 「情報学は, 既存の計算機科学や情報工学などの理工学から, 人文社会, 生命科学の全分野にまたがる複合領域を構成」
 - 数学, 情報や知識を処理するための論理, アルゴリズム, 量子計算, ソフトウェア, 情報メディア, 認知科学, 言語学, 情報社会学, 情報セキュリティ

(小野, 2002: pp. 4-6)

1. 人文情報学とはどのような分野なのか？

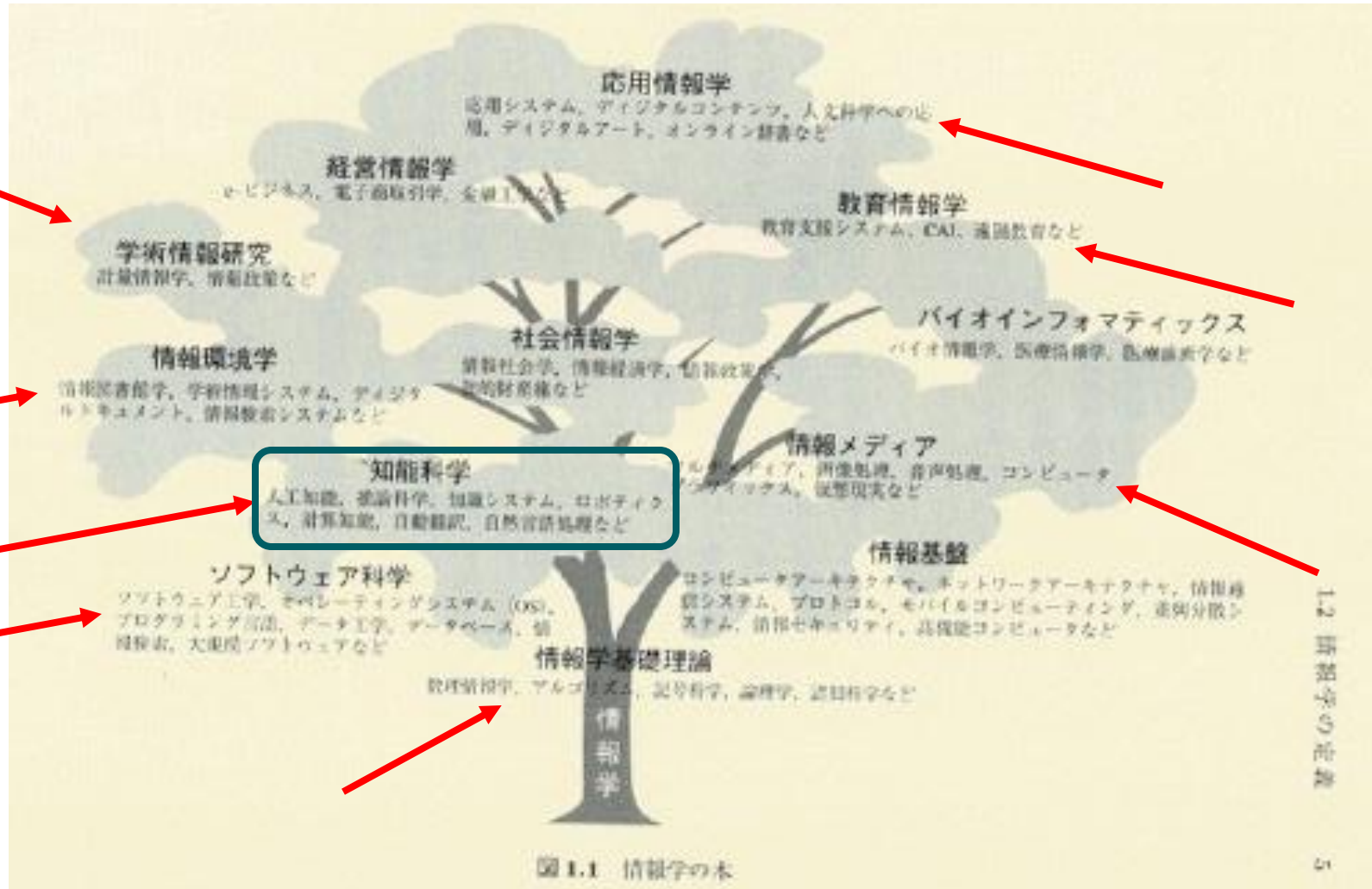
1. 「情報学」とはどのような分野なのか？



(小野, 2002: pp. 4-6)

1. 人文情報学とはどのような分野なのか？

1. 「情報学」とはどのような分野なのか？



(小野, 2002: pp. 4-6)

1. 人文情報学とはどのような分野なのか？

1. 「情報学」とはどのような分野なのか？

- 知能科学：**人工知能**，推論科学，知識システム，ロボティクス，計算知能，**自動翻訳**，**自然言語処理**
- 人工知能 (AI)：人間の知能を模倣し，人工的に作られた知能・システム・メカニズム (cf. 自然知能) ただし，その定義は一様ではない。
→ 「知能」の定義が明確ではない。
- 自動翻訳：異なる言語のテキストや音声を翻訳するプロセスに人間が介入しないもの (e.g., Google翻訳, DeepL) ※**機械学習**，**深層学習**
- 自然言語処理：人間が日常的に使う言葉（自然に生まれた言語）をコンピュータで処理したり，分析したりする

1. 人文情報学とはどのような分野なのか？

1. 「情報学」とはどのような分野なのか？

- 知能科学：**人工知能**，推論科学，知識システム，ロボティクス，計算知能，**自動翻訳**，**自然言語処理**

人工知能 (AI: Artificial Intelligence)

生成AI (Generative AI)

機械学習 (Machine Learning)

深層学習 (Deep Learning)

自然言語処理 (NLP: Natural Language Processing)

コンピュータに大量のデータを入力することでコンピュータがデータ内のパターンや規則性を学習（分析）し、新たなデータの予測や分類を行う。

教師あり学習：正解データを与える

教師なし学習：正解データ不要

人工知能 (AI: Artificial Intelligence)

生成AI (Generative AI)

機械学習 (Machine Learning)

深層学習 (Deep Learning)

自然言語処理 (NLP: Natural Language Processing)

機械学習の一つ

人間の脳の神経回路を模倣(ニューラルネット・ニューラルネットワーク)し、大量のデータから特徴を自動的に学習する

1. 人文情報学とはどのような分野なのか？

1. 「情報学」とはどのような分野なのか？

- 自然言語処理：人間が日常的に使う言葉（自然に生まれた言語）をコンピュータで処理したり，**分析**したりする

言語そのもの（文字列や音声）は分析できない

- 私 + あなた = ???
- I + play + the + piano = ???

文字や音声（あるいは画像）を数値に置き換える

1. 人文情報学とはどのような分野なのか？

- どのような調理器具を使うか
- ガスなのか, IHなのか, 炭火なのか
- 材料をどのように調達するか
- どのように調理するのか
- どの順番で材料を入れるのか
- どの順番で材料を入れるのがベストか
- そもそもこの料理におけるベストとは



- 白菜の生産者は誰なのか
- 白菜の産地はどこで, いつ頃植えられ, いつ収穫されたのか
- にんじんの糖度はどれくらいか
 - 福岡産のにんじんと新潟産のにんじんは何が違うのか
 - 豆腐は販売されるまでにどのような加工をしているのか

1. 人文情報学とはどのような分野なのか？

情報

- どのような調理器具を使うか
- ガスなのか, IHなのか, 炭火なのか
- 材料をどのように調達するか
- どのように調理するのか
- どの順番で材料を入れるのか
- どの順番で材料を入れるのがベストか
- そもそもこの料理におけるベストとは

人文

- 白菜の生産者は誰なのか
- 白菜の産地はどこで, いつ頃植えられ, いつ収穫されたのか
- にんじんの糖度はどれくらいか
 - 福岡産のにんじんと新潟産のにんじんは何が違うのか
 - 豆腐は販売されるまでにどのような加工をしているのか



人文情報学

1. 人文情報学とはどのような分野なのか？

2. 従来の「人文学」研究はどのようなものか

- 研究者（ら）の知識，経験をフル活用した質的な研究
 - データの量的な傾向ではなく，意味・文脈・主観的な判断を重視
 - 比喩，隠喩，（暗示された）主題など、定量化しづらい要素に注目
- 個別作品の精読（close reading）
 - 文学作品や歴史文書，美術作品などを時間をかけて丹念に読み解く
 - 文体，語彙，修辞技法，主題などに注目
 - 解釈や批評を通じて作品の意味や価値を探る

1. 人文情報学とはどのような分野なのか？

2. 従来の「人文学」研究はどのようなものか

- 歴史的・哲学的・文化的文脈の重視（精読に含まれる場合も有）
 - 作品が生まれた時代や社会，思想背景を深く掘り下げる
 - 作者の伝記や政治・宗教的影響なども考慮に入れる
- 手作業による資料収集・分析
 - 古文書，写本，原典資料などを図書館やアーカイブで実際に調査
 - 注釈作成，資料比較，翻訳なども手作業で行う
- 解釈の多様性と主観性
 - 明確な「正解」があるわけではなく，解釈に幅がある
 - 研究者の視点や理論が研究の立脚点

1. 人文情報学とはどのような分野なのか？

3. 人文学と情報学を掛け合わせることの意義

- 手作業による資料収集・分析

- 古文書，写本，原典資料などを図書館やアーカイブで実際に調査
- 注釈作成，資料比較，翻訳なども手作業で行う

1-1. 『万葉集』は，奈良時代に編纂された，現存する日本最古の歌集であると言われている。この『万葉集』の研究をするため，原典にあたる（読む・調査する）必要があると考えるが，原典資料にあたらうとする際に考えられる困難や課題を思いつくだけ挙げてください。

1. 人文情報学とはどのような分野なのか？

3. 人文学と情報学を掛け合わせることの意義

- 解釈の多様性と主観性
 - 明確な「正解」があるわけではなく、解釈に幅がある
 - 研究者の視点や理論が研究の立脚点

1-2. 「学術研究」において、明確な「正解」がなく、また解釈に幅があることの利点と弱点は何だと思えますか。

夏目漱石の『吾輩は猫である』を考えます。本作品の語り（ナレーション）は「吾輩」であり「猫」ですが、Aさんはこの「吾輩」と「猫」は同一人（動）物ではないと主張しています。Bさんは、「吾輩」も「猫」も同一人物であり、猫であると言います。Cさんは「吾輩」も「猫」も同一人物だが、実は猫ではなく「幽霊」なのだと主張しています。明確な正解はなく、また作者も亡くなっている以上、解釈は自由ですよね。

1. 人文情報学とはどのような分野なのか？

3. 人文学と情報学を掛け合わせることの意義

コンピュータは文学作品を「読めない」という批判

[再掲]

- 研究者（ら）の知識，経験をフル活用した質的な研究
 - データの量的な傾向ではなく，意味・文脈・主観的な判断を重視
 - 比喩，隠喩，（暗示された）主題など、定量化しづらい要素に注目
- 個別作品の精読（close reading）
 - 文学作品や歴史文書，美術作品などを時間をかけて丹念に読み解く
 - 文体，語彙，修辞技法，主題などに注目
 - 解釈や批評を通じて作品の意味や価値を探る

1. 人文情報学とはどのような分野なのか？

3. 人文学と情報学を掛け合わせることの意義

コンピュータは人文学に何をしてくれる？

- 客観的，科学的視点を取り入れることができる
 - 「〇〇だと思う」に（科学的）な根拠をプラスできる
 - 説得力が増す
- 古文書や美術作品など，経年劣化等に逆らえない物質を別の形で保存（保管）することができる

1. 人文情報学とはどのような分野なのか？

3. 人文学と情報学を掛け合わせることの意義

「科学」：

- **実証性 (Verifiability)**: ある仮説が（実験などで）実証できるか
- **客観性 (Objectivity)**: 得られた結果が客観的に認められるか
- **再現性 (Replicability/Reproducibility)**:
（実験など）同一条件下であれば誰が行なっても同じ結果になるか

1. 人文情報学とはどのような分野なのか？

3. 人文学と情報学を掛け合わせることの意義

デジタル人文学（デジタルヒューマニティーズ・人文情報学）：

- **蓄積系**：データベースの作成，アーカイブ等
 - 例：コーパス作成等（本講義のデータベースとアノテーションで少し触れます）
- **解析系**：史資料を分析・解析し，新たな知見を得る
 - 例：画像解析，テキスト分析（本講義のメイン！）等
- **可視化系**：（解析系で得られた）知見や研究成果を可視化
 - 例：地図やVR(バーチャルリアリティー；Virtual Reality;仮想現実)含

1. 人文情報学とはどのような分野なのか？

4. 人文情報学で行われている研究やその成果：

公開されているデータベースの紹介

コーパス(corpus/corpora):

自然言語の文書や音声を集めてデータベース化したもの

- Corpus of Contemporary American English (COCA; 1 billion):

<https://www.english-corpora.org/coca/>

- Corpus of Historical American English (COHA; 475 million):

<https://www.english-corpora.org/coha/>

- British National Corpus (BNC; 100 million):

<https://www.english-corpora.org/bnc/>

- 現代日本語書き言葉均衡コーパス (<https://clrd.ninjal.ac.jp/bccwj/>)

1. 人文情報学とはどのような分野なのか？

4. 人文情報学で行われている研究やその成果：

公開されているデータベースの紹介

日本古典籍くずし字データセット：

<https://codh.rois.ac.jp/char-shape/>

人文学オープンデータ共同利用センター：

<https://codh.rois.ac.jp/index.html.ja>

Helsinki Corpus:

<https://varieng.helsinki.fi/CoRD/>

1. 人文情報学とはどのような分野なのか？

4. 人文情報学で行われている研究やその成果：

公開されているデータベースの紹介

Gale Primary Sources:

<https://www.gale.com/jp/primary-sources>

Project Gutenberg:

<https://www.gutenberg.org/>

青空文庫:

<https://www.aozora.gr.jp/>

1. 人文情報学とはどのような分野なのか？

4. 人文情報学で行われている研究やその成果：

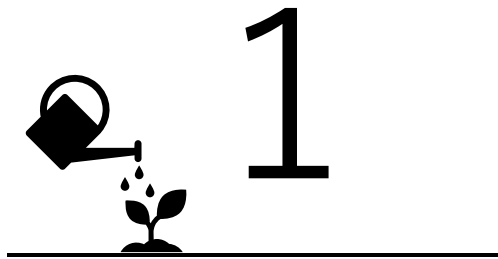
公開されているデータベースの紹介

平安京オーバーレイマップ：

<https://www.arc.ritsumeai.ac.jp/archive01/theater/html/heian/>

1-3. 平安京オーバーレイマップを見て、気付いたことを挙げてください。数は問いません。

人文情報学とは
どのような分野
なのか？

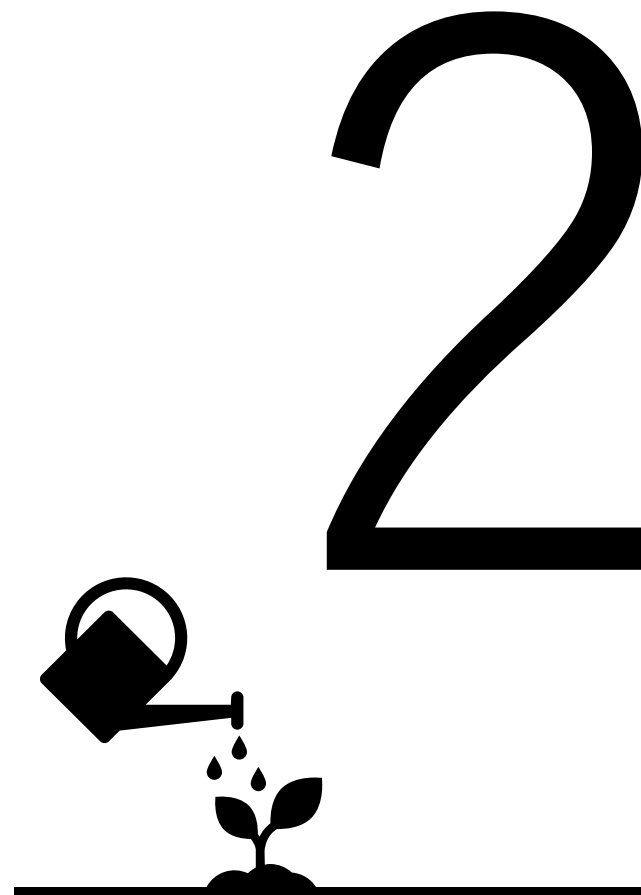


引用文献・参考文献

《引用文献》

- 小野欽司. (2002). 『情報学研究の将来像』. In: 小野欽司, 上野春樹, 根岸正光, 坂内正夫, 安達淳. (2002). 『情報学とは何か』. 東京: 丸善. pp. 1-14.

データベースと アノテーション



2. データベースとアノテーション

2. アノテーション, メタデータについて

データベース：研究対象となるテキストや画像, 音声, 映像などの資料を体系的に保存・管理する仕組み

- 膨大なデジタル資料を検索・抽出しやすくする
- 特定の構造（書簡, 詩, 語彙など）に基づいた整理が可能
- 複数の研究者間での共有や再利用を容易にする

1-4. で挙げたコーパスもデータベースの一つ

2. データベースとアノテーション

2. アノテーション, メタデータについて

アノテーション: テキストや画像などの資料に対して意味的・構造的な情報を付加する作業またはその情報

- 文法, 語彙, 文体, 主題, 人物関係などの情報を付加
- 自然言語処理や機械学習に活用する
- 解釈を明示的にデータ化する

例えば…

- 文ごとの話者情報の付与 (戯曲や書簡), 品詞タグの付与
- 古典語の語彙に現代語訳を添える
- 「これは比喩表現である」といった修辞ラベルを付与

2. アノテーション, メタデータについて

メタデータ：史資料そのものではなく、史資料に関する情報（データについてのデータ）

- 資料の検索性・識別性を高める
- 出典, 作者, 制作年, ジャンル, 言語などの属性を明示
- デジタル保存や引用時の正確性を担保

例えば…

- 書誌情報（著者, タイトル, 出版年, 出版者など）
- デジタル化日, ファイル形式, 文字コードなどの技術的情報
- キーワード, 分類コード

2. データベースとアノテーション

2. アノテーション, メタデータについて

TEI : Text Encoding Initiative

- 文献資料を電子的に構造化・記述するための国際的な規格（ガイドライン）
- テクストに意味や構造を与えるためのマークアップ言語（XML : Extensible Markup Language）

主な用途

- 文献の構造化, 意味付け, デジタルアーカイブの構築

使用対象

- 文学作品, 歴史的文書, 辞書, 書簡, 戯曲, 碑文 など

目的 : テクストの構造や意味、文脈情報を明示的に記述することにより、

- コンピュータによる処理（検索・分析・変換）を可能にする
- デジタルアーカイブの標準化と永続性を確保する
- 異なる研究者・プロジェクト間での再利用・共有を容易にする

2. アノテーション, メタデータについて

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>...</teiHeader>    <!-- メタデータ -->
  <text>
    <body>...</body>          <!-- 本文 -->
  </text>
</TEI>
```

```
.....
<sp who="#hamlet">
  <speaker>Hamlet</speaker>
  <p>To be, or not to be, that is the question:</p>
</sp>
```

```
.....
<l>Shall I compare thee to a summer's day?</l>
<l>Thou art more lovely and more temperate.</l>
```

```
.....
<w lemma="run" pos="VB">ran</w>
```

2. データベースとアノテーション

1. 1-4.で紹介したデータベースの検索機能などを実際に使用してみる

人文学オープンデータ共同利用センター：

<https://codh.rois.ac.jp/index.html.ja>

2-1.人文学オープンデータ共同利用センターのウェブサイト
にアクセスし、「データセット一覧」から興味のあるデータ
セットにアクセスし、検索等を実行してどのようなデータが
収蔵されていて、どんなことがわかるのか、簡単に説明して
ください。
+感想も

データベースと アノテーション



引用文献・参考文献

《参考文献等》

- TEI (Text Encoding Initiative) : <https://tei-c.org>

テキスト分析演習

1

(オンラインツール)

3



1. テキスト分析でよく使われる用語や指標 1

- 述べ語数(tokens) : あるテキストに含まれる語の総数 (繰り返しを含む) ; (句読点等の記号は特別な断りがない限り含まない)
e.g., I saw a cat and he saw a dog. 述べ語数 = 9 (語)
- 異なり語数(types) : テキストに登場する異なる語の数 (ユニークな語の数)
e.g., I saw a cat and he saw a dog. 異なり語数 = 7 (タイプ)

1. テクスト分析でよく使われる用語や指標 1

- TTR (Type Token Ratio):

$$\text{TTR} = \frac{(\text{types})}{(\text{tokens})}$$

語彙の多様性 (lexical diversity) を測る指標;

TTRの値が高い→ 語彙が多様, 同じ語の繰り返しが少ない

TTRの値が低い→ 限られた語が繰り返し使われている

e.g., I saw a cat and he saw a dog. (9 tokens, 7 types)

$$7/9 = 0.778$$

I witnessed a cat and he saw a dog. (9 tokens, 8 types)

$$8/9 = 0.889$$

1. テクスト分析でよく使われる用語や指標 1

$$STTR = \frac{1}{N} \sum_{i=1}^N \frac{types_i}{tokens_i}$$

- STTR(Standardized type-token ratio) : テクストを一定の語数ごとの区間 (たとえば100語) に分割し,それぞれの区間でTTRを算出して平均値を取る方法
 - 各区間のTTRを平均することで, テクスト長の影響を抑えることができる
 - テクストの比較可能性が高まる
 - e.g., 500語のテキストを100語ごとに分割 (5区間)
 - 各区間のTTRを求めて平均 → STTR
- 長さの異なるテキスト間でも比較しやすい
- 言語発達研究, 児童の作文分析, スタイル分析などでよく使用される

1. テキスト分析でよく使われる用語や指標 1

- n -gram : テキスト中の連続した n 個の語 (または文字) の並び
 - $n = 1$: unigram (ユニグラム) (単語1個)
 - $n = 2$: bigram (バイグラム) (単語2個)
 - $n = 3$: trigram (トライグラム) (単語3個)
 - $n \geq 4$: 4-gram、5-gram... (一般に n -gram と総称)

I saw a cat and he saw a dog.

- 2 gram: I – saw, saw – a, a – cat, cat – and, and – he ...
- 3 gram: I – saw – a, saw – a – cat, a – cat – and ...

1. テキスト分析でよく使われる用語や指標 1

- n -gram : テキスト中の連続した n 個の語 (または文字) の並び
 - 言語モデル : 次に来る語を予測 (例 : "I want to" → "go")
 - 頻度分析 : よく使われるフレーズや表現の抽出
 - 文体分析 : 作家やジャンルごとの語順傾向の分析
 - 音声認識・機械翻訳 : 単語や音の連鎖の予測に活用
 - キーワード抽出 : 固定句 (例 : "due to", "on the other hand" など) の検出

1. テキスト分析でよく使われる用語や指標 1

- 共起(co-occurrence) : 2つ以上の語が, ある基準で定められた範囲内で同時に出現すること
 - 語単位 : 語と語が, 前後 n 語以内に現れる
 - 文単位 : 同じ一文内に出現
 - 段落単位 : 同じ段落内に出現
 - 文書単位 : 同じ文書内に出現
- 文学作品で「愛」という語と共起する語 → 「涙」「別れ」「手紙」など
→ 感情的トピックの分析
- 歴史的文書で「王」という語と共起する語 → 「命令」「反乱」「裁き」 → 政治的文脈の分析

1. テキスト分析でよく使われる用語や指標 1

- コロケーション (collocation) :ある語が他の語と頻繁に連続して現れること ; ある語が他の語と頻繁に連続して現れ, 自然で流暢な言語使用を形成する語の組み合わせ
 - 「大きい雨」 ?? ; 「激しい雨」 ◎
- 言語学習 : ネイティブらしい表現習得
- 自然言語処理 : コロケーション辞書, 自動要約, 機械翻訳の精度向上
- コーパス言語学 : 頻出コロケーションの抽出による語法研究や文体分析

2. 既存のオンラインツールを使ったテキスト分析

Voyant Tools: <https://voyant-tools.org/?lang=ja>

- Stéfan Sinclair
- Geoffrey Rockwell

A screenshot of the Voyant Tools web interface. At the top, it says "テキストを追加する" (Add text). Below that is a large text input area with the instruction: "一つのURL、もしくは複数のURLを行わずつ入力するか、あるいはテキスト全文を貼りつけてください" (Enter one URL, or multiple URLs without a line break, or paste the full text). At the bottom of the input area are two buttons: "開く" (Open) and "アップロードする" (Upload). To the right of these is a blue button with a checkmark and the text "結果を表示する" (Show results).

Voyantツールは、デジタルテキストの読解と分析のためのウェブ上の環境です。
小風尚樹(Naoki Kokaze), 佐藤正尚(Masanao Sato), 杉浦清人(Kiyoto Sugiura), 鈴木親彦(Chikahiko Suzuki), 王一凡(Yifan Wang), and 永崎研宣(Kiyonori Nagasaki)

2. テキスト分析でよく使われる用語や指標 2 (Voyant tool内の語)

- ドキュメントの長さ(Document Length) : 1ファイル (作文) あたりの語数
- 語彙密度(Vocabulary density) : 語の種類豊富さ ; この値が高ければ高いほど, 使われている語の種類が多い (内容の複雑さは増す)
- Average Words Per Sentence : 平均文長 ; 一つの文が何語で構成されているか, の平均値 ; 値が高ければ高いほど, 一文が長い傾向にある ; 長い文は読みづらい, 理解しづらい文傾向にある
- Readability Index : 読みやすさの指標 ; この値が高ければ高いほど, 読み易い, 理解し易い
- 特徴語 : 全体と比較して, 特定の文書 (ファイル, 作文) に特徴的に現れる (用いられている) 語

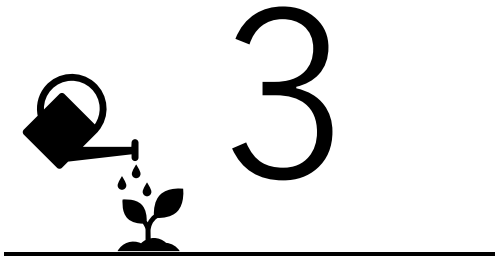
3. テクスト分析演習 1 (オンラインツール)

2. 既存のオンラインツールを使ったテキスト分析

3. Voyant tool で “Austen’s Novels” の作品コーパスを選択し、2つの作品についてどのような特徴があるのか、Voyantツール内で提示された情報をもとに説明してください。

テキスト分析演習

1
(オンラインツール)



引用文献・参考文献

《参考文献等》

- Voyant Tools: <https://voyant-tools.org>

テキスト分析演習

2

(前処理)

4



1. 分析対象を電子データ化し, 保存する。 (英語)
2. 分析対象を電子データ化し, 保存する。 (日本語)
3. 正規表現

テキストエディタの Sublime Text をダウンロードします :

<https://www.sublimetext.com/3>

4. テキスト分析演習 2 (前処理)

1. 分析対象を電子データ化し, 保存する。(英語)
 2. 分析対象を電子データ化し, 保存する。(日本語)
- PDFで読み込まれたデータをプレーンテキストファイル(.txt)に変換し, 保存します。
 - PDFファイルのままではテキストの分析は難しい
 - MS Wordではダメなのか? : MS Wordなどはテキスト等の情報を単に保存するだけではなく, 見易さや様式の設定のしやすさなど, 見えない情報が多く存在
 - プレーンテキストファイル: 文字列をデータとしてシンプルに保存
 - コンピュータが扱いやすい状態
 - PDF ファイルの中身と Sublime Text の中身を比較してみる
 - PDF ファイルの通りになっていますか…?

3. 正規表現 (Regular Expression, regex)

- 正規表現 (regex) とは, 文字列のパターンを表現・検索・抽出するための記述方法
- プログラミングやテキスト分析, 検索置換処理などで利用される

3. 正規表現 (Regular Expression, regex)

.	任意の1文字	$a.b \rightarrow acb, a1b$ にマッチ
*	直前の文字の0回以上の繰り返し	$ab^* \rightarrow a, ab, abb, abbb$ にマッチ
+	直前の文字の1回以上の繰り返し	$ab^+ \rightarrow ab, abb, abbb$
?	直前の文字の0回または1回	$ab? \rightarrow a, ab$
[]	指定した文字のいずれか	$[abc] \rightarrow a, b, c$
[^]	指定した文字以外	$[^abc] \rightarrow a, b, c$ 以外の文字
()	グループ化	$(ab)^+ \rightarrow ab, abab, ababab$
^	行の先頭	$^The \rightarrow$ 行頭が The の行
\$	行の末尾	$end\$ \rightarrow$ 行末が end の行
	または	$ab cd \rightarrow ab, cd$ にマッチ

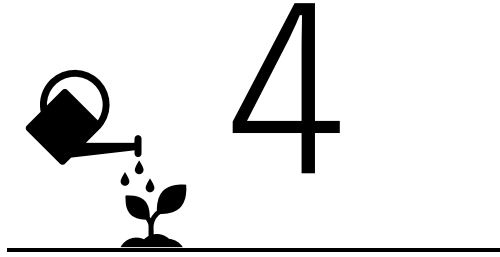
<code>\d</code>	数字1文字 ([0-9])
<code>\D</code>	数字以外の1文字
<code>\w</code>	英数字・アンダースコア1文字 ([a-zA-Z0-9_])
<code>\W</code>	英数字・アンダースコア以外
<code>\s</code>	空白文字 (スペース、タブ、改行など)
<code>\S</code>	空白文字以外
<code>\n</code>	改行
<code>\t</code>	タブ
<code>\b</code>	語の境界
<code>{n}</code>	直前の文字がn回繰り返す
<code>{n,m}</code>	直前の文字がn回以上m回以下
<code>\1 or \$1</code>	()でグループ化した中身を参照 (数字は左から数えて何個目のグループかを指す。\ <code>\or</code> \$はエディタによって違いあり)

3. 正規表現 (Regular Expression, regex)

- 4-1. 英語のPDFファイルとプレーンテキストファイルに貼り付けたデータはどのようなところが異なっていましたか
- 4-2. 日本語のPDFファイルとプレーンテキストファイルに貼り付けたデータはどのようなところが異なっていましたか
- 4-3. 修正して保存した.txtファイルをアップロードしてください
- 4-4. .txtを赤枠と青枠の正規表現で検索をし, 何か違いはありましたか。

テキスト分析演習

2
(前処理)



引用文献・参考文献

《参考文献等》

- Sublime Text: <https://www.sublimetext.com/3>

他にもおすすめのテキストエディタ

- miエディタ (mac OS): <https://www.mimikaki.net/>
- サクラエディタ (Windows): <https://sakura-editor.github.io/>
- VS Code (Windows, mac OS): <https://code.visualstudio.com/>
- BBText (mac OS): <https://www.barebones.com/products/bbedit/>

テキスト分析演習

3

(前処理)

5



5. テキスト分析演習 2 (前処理)

1. 分かち書き・形態素解析 (日本語)

- 文章中の語を単位ごとに区切って表記すること
- 日本語の文章は通常、語と語の間に空白を入れずに書かれるため、機械処理やテキスト分析を行う際に「どこで語が切れるか」を明示的に分ける作業が必要

e.g., 私は大学に行きます。

私 は 大 学 に 行 き ま す 。

- 語とは意味を持つ文字列の最小単位
- 英語はスペースを入れて記述するため、語の単位の区切りが明確

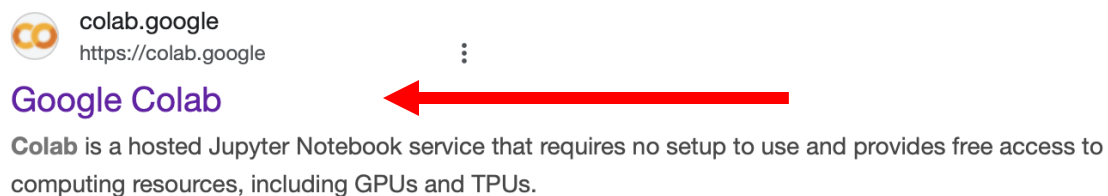
1. 分かち書き・形態素解析 (日本語)

- 形態素解析：文章 (テキスト) を最小の意味を持つ単位である「形態素」に分割し (分かち書き), それぞれの品詞や活用形などの情報を付与する処理
 - 形態素解析ツールを用いることで, 分かち書きが可能
 - MeCab; JUMAN++; SudachiPy; GiNZA / spaCy
-
- 形態素：これ以上小さく分けられない, 意味を持つ最小の単位
単語の構成要素の中で「意味・文法的機能」を担う単位
e.g., 私 は 大学 に 行きます 。
名詞 助詞 名詞 助詞 動詞 助動詞
He go-es to the university. / It is ir-regular. /
I final-ly pass-ed the exam.

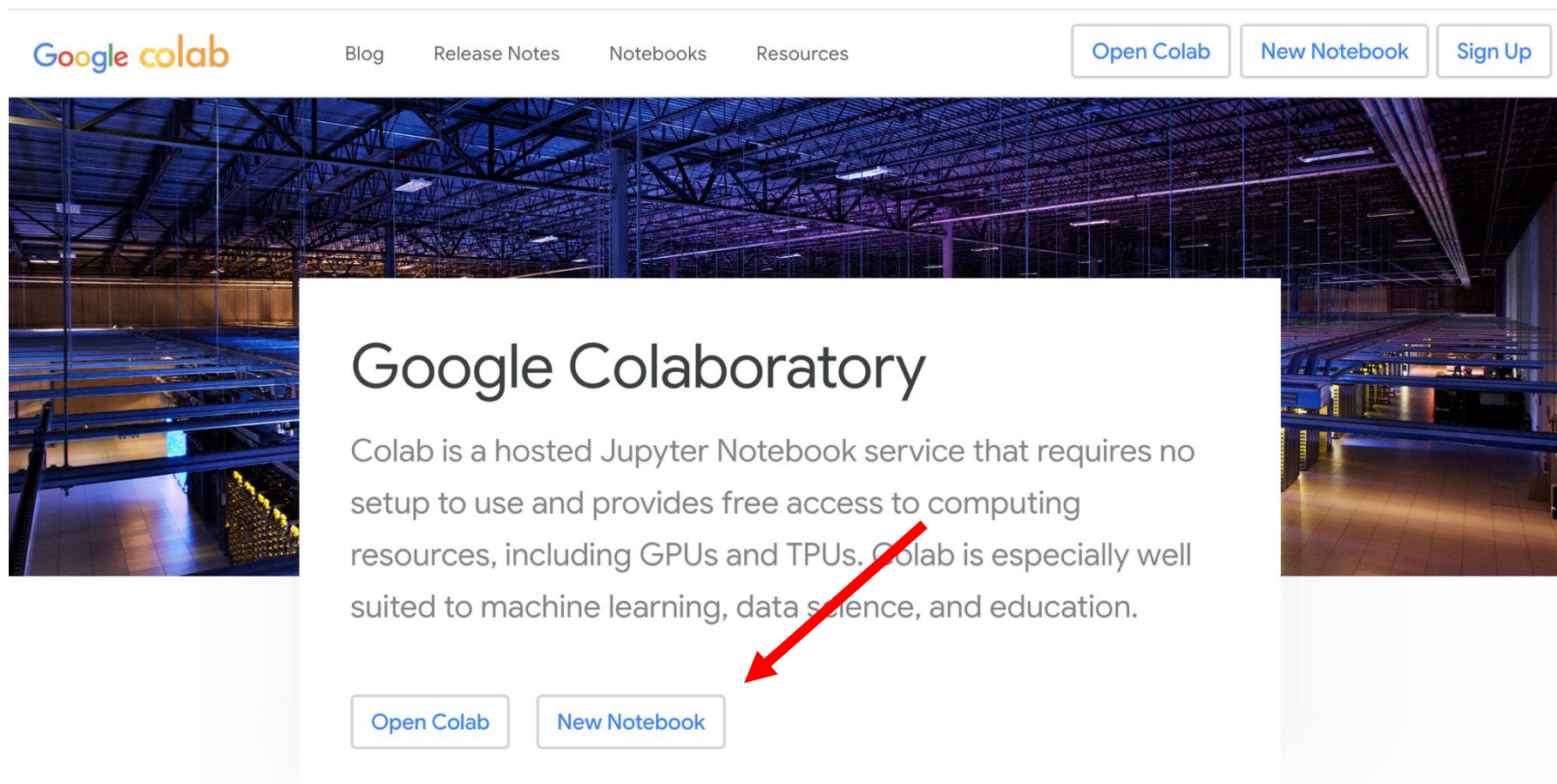
5. テキスト分析演習 2 (前処理)

Google Colab(oratory)を使って, python
を実行できる環境にしましょう!

- 必要なもの : Google のアカウントとパスワード
- ブラウザを開いて, 検索ウィンドウに「Google Colab」と入力



Google Colab(oratory)を使って, python
を実行できる環境にしましょう!



Google colab

Blog Release Notes Notebooks Resources

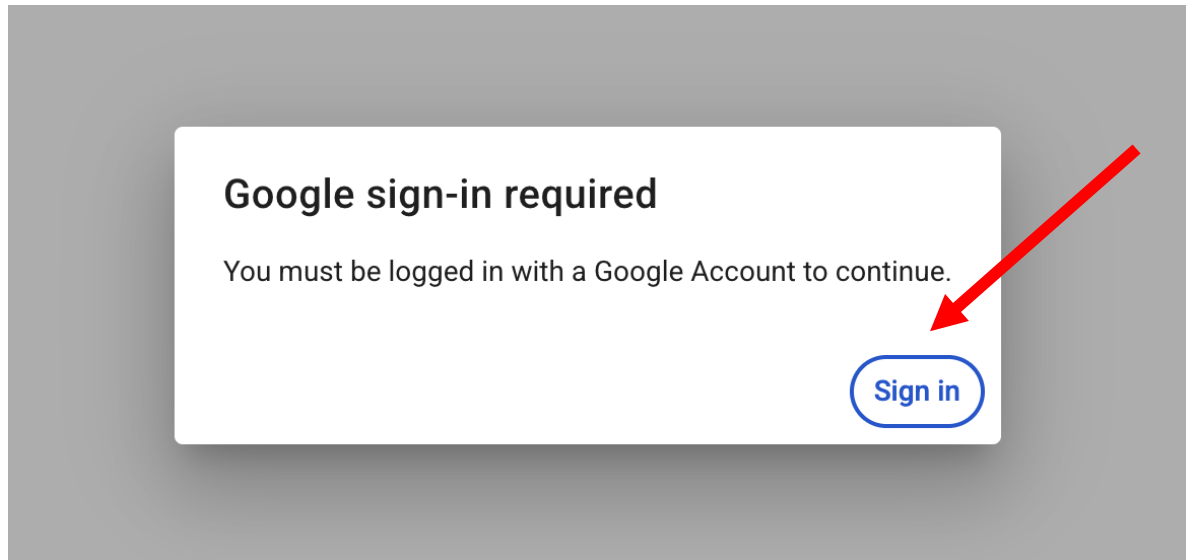
Open Colab New Notebook Sign Up

Google Colaboratory

Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education.

Open Colab New Notebook

Google Colab(oratory)を使って, python
を実行できる環境にしましょう！



Googleのアカウント（メールアドレス）を入力し, 次へ進むとパスワード入力画面が表示されます

プログラミング用語（ちょっとだけ）

分からない用語が出てきたらIT用語辞典e-Words等を活用
(<https://e-words.jp>)

- 変数(へんすう) : 変数 (variable) とは,コンピュータプログラムのソースコードなどで,データを一時的に記憶しておくための領域に固有の名前を付けたもの。(IT用語辞典e-Words <https://e-words.jp>)

```
[ ] text = "My name is Iku Fujita"
```

```
[ ] list = ["My", "name", "is", "Iku", "Fujita"]
```

プログラミング用語（ちょっとだけ）

- 引数(ひきすう)：引数（argument）とは、プログラム中で関数やメソッド、サブルーチンなどを呼び出すときに渡す値のこと。渡された側はその値に従って処理を行い、結果を返す。オペレーティングシステム（OS）の操作などで利用者がコマンドを実行する際に指定するパラメータ（コマンドライン引数）などを指すこともある。（IT用語辞典e-Words <https://e-words.jp>）

プログラミング用語（ちょっとだけ）

- 関数（かんすう）：関数（function）とは、コンピュータプログラム上で定義されるサブルーチンの一種で、数学の関数のように与えられた値（引数）を元に何らかの計算や処理を行い、結果を呼び出し元に返すもののこと。
（IT用語辞典e-Words <https://e-words.jp>）

プログラミング用語（ちょっとだけ）

パス：経路のこと。コンピュータ等上の“住所”と思ってください。

- 絶対パス：絶対パス（absolute path）とは、ファイルなどの所在を書き表すパス（path）の表記法の一つで、階層構造の頂点（最上位階層）からの位置関係を記述する方式。（IT用語辞典e-Words <https://e-words.jp>）
- 相対パス：相対パス（relative path）とは、ファイルなどの所在を書き表すパス（path）の表記法の一つで、現在位置からの相対的な位置関係を記述する方式。起点となる位置から目的の位置までの道筋にある要素を順に並べて記述する。（IT用語辞典e-Words <https://e-words.jp>）

2. ~~ステミング~~(レマ化) (英語, 日本語)

レマ化 (lemmatization) とは :

単語の活用形や変化形を辞書的な基本形 (lemma) に変換する処理

e.g., 英語

running → run

better → good

cats → cat

went → go

am, are, is → be

e.g., 日本語

「走った」 → 「走る」

「美しかった」 → 「美しい」

「食べます」 → 「食べる」

2. レマ化 (英語, 日本語)

レマ化 (lemmatization) をすると何が嬉しいのか :

- 単語の変化形を統一することで語彙のバラつきを抑え, 単語の出現頻度の正確な把握が可能
- “cats” と “cat” を同一として扱うことで次元数(※)を減らす

※表の行列の数と考えてください。行列の数が多くなればなるほど, 計算等が複雑になります。

Google Colab でレマ化してみます ([Lemmatization.ipynb](#)) 。

テキスト分析演習

3
(前処理)



引用文献・参考文献

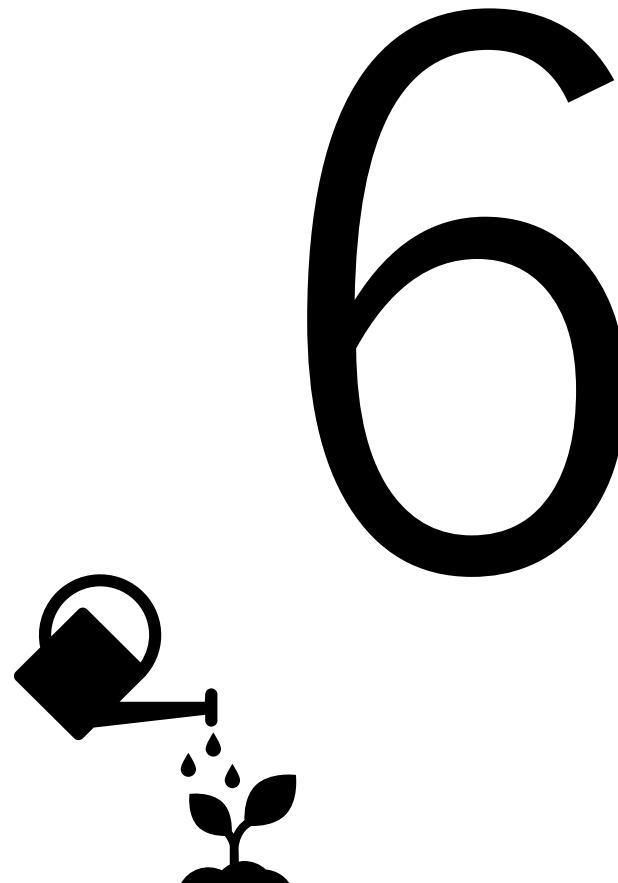
《参考文献等》

- IT用語辞典e-Words <https://e-words.jp>

テキスト分析演習

4

(前処理)



1. 品詞タグとはなにか

- 品詞 (part-of-speech) タグ : PoS/POS tags; POSタグ
- テキスト内の単語に「名詞」「動詞」「形容詞」「副詞」などの品詞情報を付与するラベル
- このラベルを付与するツール : タガー (tagger)
- 英語品詞タグセット (辞書) :
 - Penn Treebank POS Tags
(https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)
 - Universal POS Tags (Universal Dependencies)
(<https://universaldependencies.org/u/pos/>)
 - CLAWS Tagset (<https://ucrel.lancs.ac.uk/claws7tags.html>)
 - Brown Corpus Tagset
(<https://varieng.helsinki.fi/CoRD/corpora/BROWN/tags.html>)

1. 品詞タグとはなにか

- 品詞 (part-of-speech) タグ : PoS/POS tags; POSタグ
- テキスト内の単語に「名詞」「動詞」「形容詞」「副詞」などの品詞情報を付与するラベル
- このラベルを付与するツール : タガー(tagger)→ 日本語では形態素解析器(mecab等) で辞書を用いてタグを付与する
- 日本語品詞タグセット (辞書) :
 - https://zenn.dev/ymmt1089/articles/20220919_mecab_dictionary
 - IPADIC
 - UniDic (<https://clrd.ninjal.ac.jp/unidic/>)

2-1. 品詞タグ付け (英語)

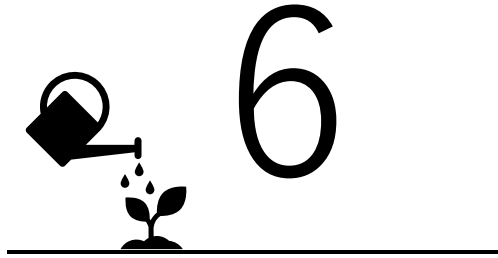
- POSTag_Eng.ipynb

2-2. 品詞タグ付け (日本語)

- POSTag_JP.ipynb

テキスト分析演習

4
(前処理)



引用文献・参考文献

《参考文献等》

テキスト分析演習

5

(頻度)



7

1. 既存の大規模コーパスで頻度を調べる

- 英語 : Corpus of Contemporary American English (<https://www.english-corpora.org/coca/>)
- 日本語 :
現代日本語書き言葉均衡コーパス (<https://clrd.ninjal.ac.jp/bccwj/>) (今日は少納言)

2. 素頻度と相対頻度

- **素頻度** : テキスト内で単語 (ある対象) が出現した回数そのもの
 - e.g., I love apples and I love oranges.
 - I が 2回 出現 → 素頻度 = 2
 - love が 2回 出現 → 素頻度 = 2
 - apples が 1回 出現 → 素頻度 = 1
 - 計算が簡単
 - テキストの中でどの単語がどれだけ使われているかをそのまま把握できる

2. 素頻度と相対頻度

- **相対頻度**：単語の素頻度を，テキスト全体の単語数（総トークン数）で割った値であり，全体に対する割合（頻度の比率）

- 相対頻度 = $\frac{\text{素頻度}}{\text{テキスト内の総語数}}$
- e.g., I love apples and I love oranges.
 - トークン総数 = 8 (I, love, apples, and, I, love, oranges, .)
 - “love” の素頻度 = 2
 - 相対頻度 = $2 / 8 = 0.25$ (25%)
- テキスト量が異なる場合でも 比較可能
- テキスト間比較，時系列比較に便利

2. 素頻度と相対頻度

- 相対頻度 ; **PMW (Per Million Words) 等**

- $PMW = 100万語あたりの単語の頻度 (出現回数)$

- $PMW = \frac{\text{素頻度}}{\text{テキストの総語数}} \times 1,000,000$

- e.g., apple という語が8回出現した文書の総語数は2000語

- 相対頻度 = $8 / 2000 = 0.004$

- $PMW = 8 / 2000 * 1000000 = 4000$

- 1000語あたり $PTW = 8 / 2000 * 1000 = 4$

2. 素頻度と相対頻度

素頻度：

- 同一テキスト内でどの単語が多く使われているかを見るとき
- 頻出語抽出, キーワード候補の把握

相対頻度：

- テキスト間で単語使用頻度を比較するとき (e.g., 作家ごとの文体比較, 時代ごとの語彙比較)
- コーパス分析, 時系列研究

1. 既存の大規模コーパスで頻度を調べる

- 英語 : Corpus of Contemporary American English
(<https://www.english-corpora.org/coca/>)

7. COCA で “maybe” と “perhaps” をそれぞれ検索し、頻度とジャンル、文中の位置に着目して、どのように違いがあるかを述べてください。

テキスト分析演習

5
(頻度)



引用文献・参考文献

《参考文献等》

テキスト分析演習

6

(頻度)

8



1. 頻度表を作成1 (英語・日本語)

- 英語と日本語それぞれ素頻度と相対頻度の頻度表を作成
 - 英語頻度 : `Freq-EN.ipynb`
 - 日本語頻度 : `Freq-JP.ipynb`
- .csv (comma separated values) ファイルにて出力
- 出力されたファイルをダウンロードし, excelを使って頻度表 csvファイルを開く

2. 作成した頻度表を考察する1 (記号の扱い)

8-1. ダウンロードしたcsvファイルを見て, いらないと思うものを挙げてください

3. 頻度表を作成2 (英語)

- 英語の素頻度と相対頻度の頻度表を再度作成
 - “ken’s”や “zig-zag” は一語として扱いたい (英語)
- 上記(アポストロフィ’とハイフン-)以外の記号は不要
- いずれのファイルも頻度の高い→少ない (降順) で表示
- .csv (comma separated values) ファイルで出力

4. 作成した頻度表を考察する2

8-2. 最初に作成した頻度表と, 2回目に作成した頻度表を見比べて, 気付いたこと, 変わったことを (見た目等) 挙げてください。

8-3. 英語の素頻度の表 (2回目) を見て, 気付いたことを挙げてください。 (内容, 語の種類)

4. 機能語と内容語 (英語)

- 内容語(content words): 意味を持つ語 (内容を伝える語)
 - 名詞: book, teacher, freedom
 - 動詞: run, speak, build
 - 形容詞: happy, difficult, red
 - 副詞: quickly, always, very
- 機能語(function words): 文法的, 構造的な役割を持つ語
 - 前置詞: in, on, at, for
 - 冠詞: a, an, the
 - 助動詞: can, will, be, have
 - 代名詞: he, she, it, they
 - 接続詞: and, but, because
 - 限定詞: this, those, some

4. 機能語と内容語(日本語)

- 内容語(content words):意味を持つ語 (内容を伝える語)
 - 名詞 : 本、先生、自由
 - 動詞 : 走る、話す、作る
 - 形容詞 : 嬉しい、難しい、赤い
 - 副詞 : 速く、いつも、とても
- 機能語(function words):文法的, 構造的な役割を持つ語
 - 助詞 : は、が、を、に、で、の、と、から、まで
 - 助動詞 : ~です、~ます、~ない、~た、~られる、~ようだ
 - 接続詞 : しかし、だから、そして
 - 指示語 (一部) : これ、それ、あれ (文法的に指示する語)

4. 機能語と内容語 (英語)

It is a truth universally acknowledged, that a single man in possession of a good fortune must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered as the rightful property of some one or other of their daughters.

Pride and Prejudice, Jane Austen

4. 機能語と内容語 (英語)

truth universally acknowledged, single man
possession good fortune want
wife. However little known feelings views such
man first entering neighbourhood,
truth so well fixed minds surrounding
families, considered rightful property
one other daughters.

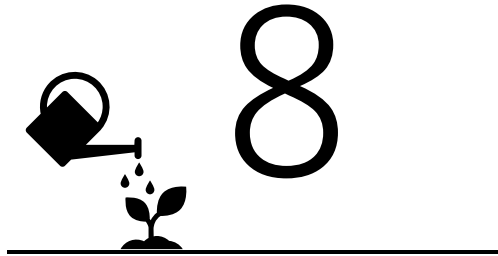
Pride and Prejudice, Jane Austen

4. 機能語と内容語

- 機能語 -- > 文体 (書き手の癖) 等
 - どんな文書にも (同じ言語で書かれたテキストであれば) 一定数現れる
 - どのような文書の集まりでも, 語の頻度を集計する際には上位に上がってきやすい
- 内容語 -- > 作品内のテーマ, 重要な内容, 作者の考え等
 - 文書の種類 (ジャンル, 内容, トピック等) によって使用される内容語の種類, 頻度は異なる

テキスト分析演習

6
(頻度)



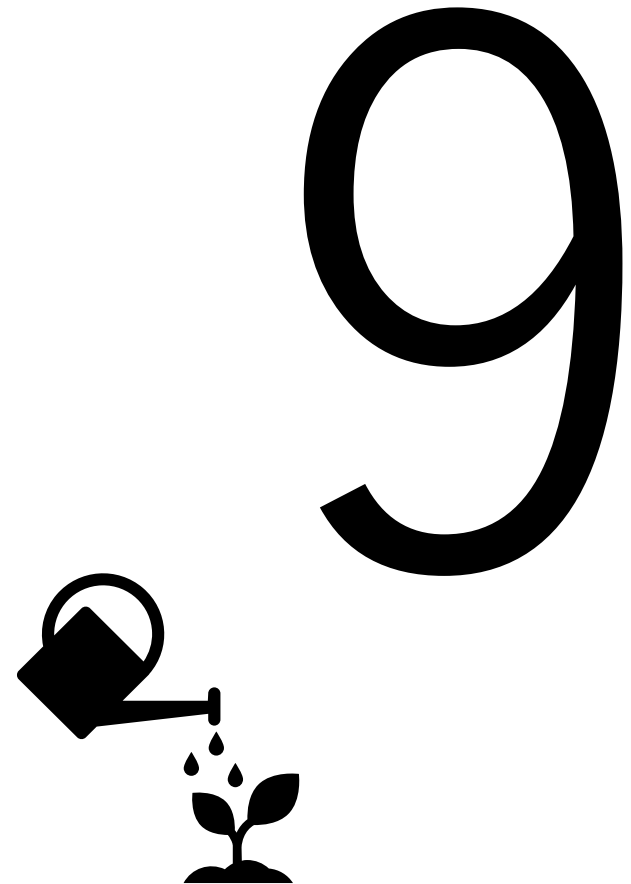
引用文献・参考文献

《参考文献等》

テキスト分析演習

7

(頻度)



1. TF-IDF (Term Frequency-Inverse Document Frequency ; 単語頻度 – 逆文書頻度)

- 「ある文書で特によく使われていて、かつ全体ではあまり出現しない語」を見つけるための重みづけ指標
 - その文書にとって重要な語を数値化して抽出
 - よく出てくるが他の文書でもよく使われる語 (= ありふれた語) はスコアを下げる
 - 検索エンジン・情報検索・トピック分析・クラスタリングなどでよく使われる
- TFIDF.ipynb

3. TF-IDF (Term Frequency-Inverse Document Frequency ; 単語頻度 – 逆文書頻度)

- TF (Term Frequency) = 「その文書内で何回使われているか」

$$TF(t, d) = \frac{t \ d}{d} \quad TF(t, d) = \frac{\text{語 } t \text{ の文書 } d \text{ における出現回数}}{\text{文書 } d \text{ の全単語数}}$$

- 単語の出現頻度を文書内で正規化したもの (相対頻度)

- IDF (Inverse Document Frequency) = 「その語がどれだけ珍しいか」

$$IDF(t) = \log\left(\frac{N}{n_t+1}\right) \quad IDF(t) = \log\left(\frac{\text{文書全体の数}(N)}{\text{語 } t \text{ を含む文書の数 (出現文書数)} (n_t)+1}\right)$$

- よく出現する語 (例えば「です」「する」など) はスコアが下がる
- +1 はゼロ除算を防ぐための調整

- TF-IDF = その語がその文書でどれくらい重要かを示す指標

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

2. ジャンル (レジスター) 間の比較

対象コーパス (みたいもの) vs 参照コーパス (比較対象)

例 : Austen に対して, Dickens の使用語彙はどのように特徴的 ?

樋口一葉 に対して, 夏目漱石 の使用語彙はどのように特徴的 ?

たとえば sea が :

対象 (Dickens) [1万語] : 80回

参照 (Austen) [2万語] : 5回

→ sea は Dickens に特徴的な語

- 対数尤度比 (Log-likelihood (ratio))
- MI-score
- Mann-Whitney の U 検定

統計検定を使って, 対象コーパス (の語彙) が参照コーパスと比較して有意に異なる語を特徴語

2-1. ジャンル (レジスター) 間の比較 : 対数尤度比

対数尤度比(Log-Likelihood ratio; LLR) : ある統計モデルが観測データをどれだけうまく説明できるかを表す指標の一つ。尤度 (likelihood) の対数を取ったもの

- 尤度 (likelihood) : モデルのパラメータのもとで, 観測データが得られる「確からしさ」
- 対数尤度 (log-likelihood) : 尤度の値はととても小さくなりがちであるため, 計算や解析をしやすいするために対数をとったもの

2-1. ジャンル (レジスター) 間の比較 : 対数尤度比

「LLRとは、2つのコーパスの中で統計的に強い偏りがある特徴語を抽出するためのもの」

表1 統計量を求めるための分割表

	対象コーパス	参照コーパス	計
見出し語 W の頻度	a	b	a+b
見出し語 W 以外の頻度	c	d	c+d
計	a+c	b+d	a+b+c+d= (n)

c = 対象コーパスの総語数 - a

d = 参照コーパスの総語数 - b

<計算式>

- ① $\text{Chi}^2 = n(ad-bc)^2 / ((a+b)(c+d)(a+c)(b+d))$
- ② $\text{Yates} = n(|ad-bc| - n/2)^2 / ((a+c)(b+d)(a+b)(c+d))$
- ③ $\text{LLR} = 2(a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) + n \log(n))$

寺嶋 (2009: 72, 79)

2-1. ジャンル (レジスター) 間の比較 : 対数尤度比

- `Loglikelihood.ipynb`

9-3. Dickens vs Austen の Log-Likelihoodの結果を見て, Austenの作品に特徴的な語にはどのような傾向があるといえそうでしょうか。

9-4. 夏目漱石と芥川龍之介のLog-Likelihoodの結果を見て, どちらかの作家に特徴的な語にはどのようなものがあるのかを挙げてください。

2-2. ジャンル (レジスター) 間の比較 : MI-Score

MI-score (Mutual Information Score; 相互情報量) : 2つの語 (または変数) の間にどれだけ強い関連があるか (共起の度合い) を測る統計的な指標。語A(x)と語B(y)が一緒に出現する頻度が偶然か, それとも意味的に関連しているのかを判断するのに使える

(e.g., Church and Hanks, 1990)

MI-score.ipynb

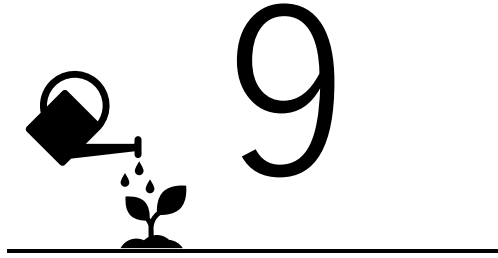
$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

2-3. ジャンル (レジスター) 間の比較 : マン・ホイットニーのU検定

- 「2つの独立したグループの中央値に差があるか」を検定するためのノンパラメトリック検定
- 「この2つのグループは、値の大小の傾向に違いがあるのか？」を検定する... t 検定のノンパラメトリック版 (カテゴリカルデータにも使える)
 - グループAとグループBのスコアを比べたいが、正規分布とは限らないとき
 - 平均ではなく、中央値の差を見たいとき
 - サンプルサイズが小さい / 外れ値があるとき
 - パラメトリック (手法) t 検定等 : 正規分布の仮定が必要
 - カテゴリカルデータ (質的データ) : 性別, 職業等
- Mann-Whitney.ipynb

テキスト分析演習

7
(頻度)



引用文献・参考文献

《参考文献等》

- Dunning, Ted. (1993). "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics*. 19 (1): 61–74. (<https://aclanthology.org/J93-1003.pdf>)
- 寺嶋弘道. (2009). 「日本語教育語彙を選定するための統計的指標— 尤度比検定、カイ2乗検定、イエーツの補正公式の特徴 —」 『ポリグロシア』 17巻: 71–83. (https://www.apu.ac.jp/rcaps/uploads/fckeditor/publications/polyglossia/Polyglossia_V17_Terajima.pdf)
- Church, Ward Kenneth and Hanks, Patrick. (1990). "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics*. 16(1):22–29. (<https://aclanthology.org/J90-1003.pdf>)

テキスト分析演習

8

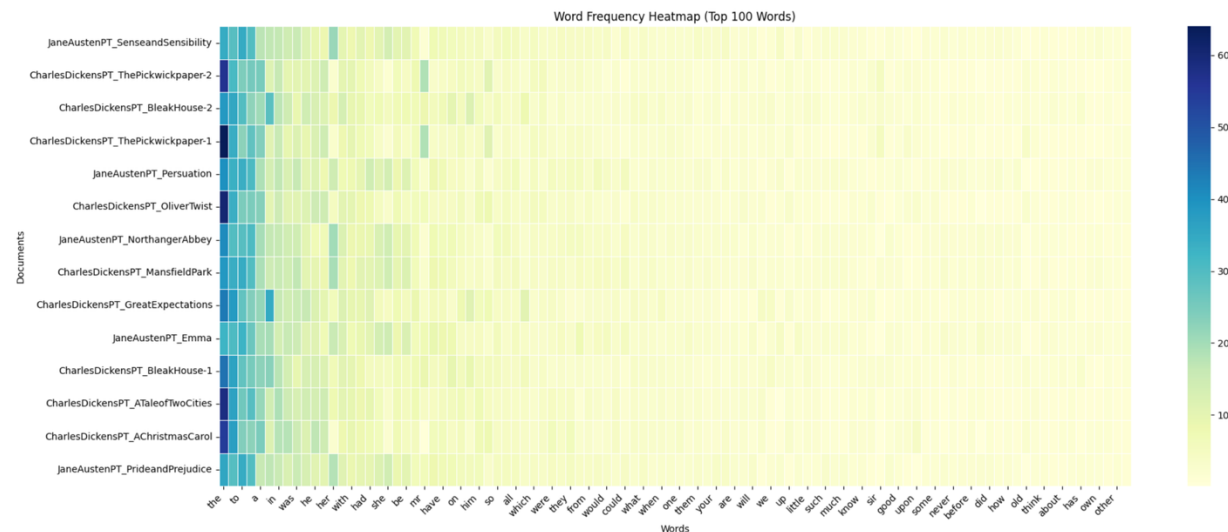
(視覚化)

10



2. ヒートマップ

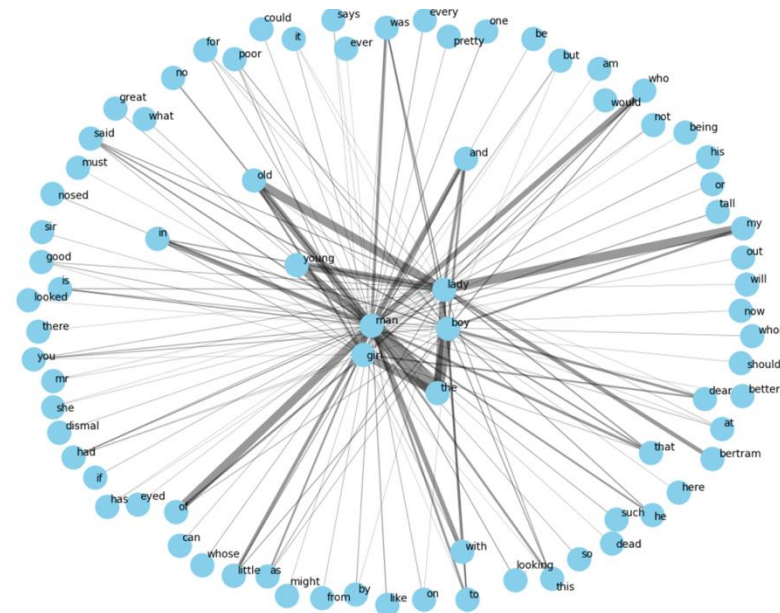
Heatmap.ipynb



- 数値データを色の濃淡やグラデーションで視覚的に表現する図
- 各セルの色で値の大小を表現し、パターンや傾向を一目で把握できる
 - 色の違い（明暗や色相）で数値の高低が直感的に分かる
 - 行や列の比較に便利
 - 大量のデータでも全体像の把握がしやすい

3. 共起ネットワーク

Network.ipynb



- テクスト中で一緒に現れやすい語同士を結んだグラフです。
 - ノード：語（または形態素・表現）
 - エッジ：語Aと語Bが**共起**した関係
 - 重み：共起の強さ（頻度や統計的尺度）
- 主に「語彙間の関係」を可視化する

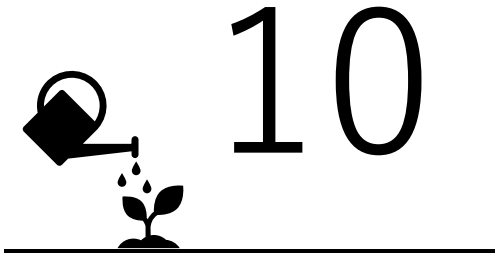
3. 共起ネットワーク

• 共起の定義

- ウィンドウ共起：スライディング窓（例：前後±5語）内に同時出現
→ 局所的な連語傾向を捉えやすい
- 文内共起：同一文に出たら共起 → 文レベルの意味連関
- 文書内共起：同一文書に出たら共起 → テーマ（話題）単位の関連
- 依存関係共起（係り受け）：「主語-述語」「修飾-被修飾」など
構文関係に限定 → 意味的にシャープ
- コロケーション：ある語と語が一緒によく使われる慣用的な語の組み合わせ

テキスト分析演習

8
(視覚化)



引用文献・参考文献

《参考文献等》

テキスト分析演習

9

(機械学習)

1 1



11. テキスト分析演習 9 (階層的クラスタリング)

階層的クラスタリングとは

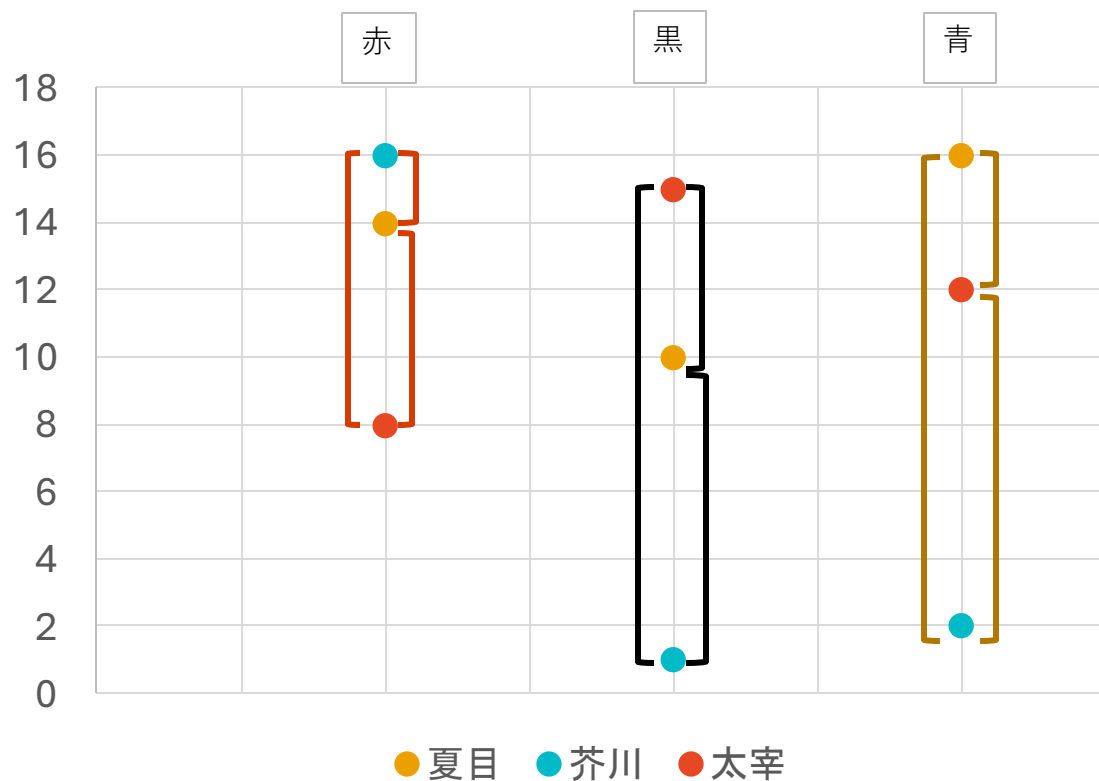
- 各要素の距離を測り, その近さを元にグルーピング
- 近い = 類似している
- 相対頻度を元に距離を測るため, 意味が類似しているのではなく, あくまで使用 (相対) 頻度の値が近い

文献	色彩語		
	赤	黒	青
夏目	14	10	16
芥川	16	1	2
太宰	8	15	12

- 夏目と芥川の距離?
- 夏目と太宰の距離?
- 芥川と太宰の距離?

11. テキスト分析演習 9 (階層的クラスタリング)

階層的クラスタリングとは



赤	夏目-芥川	-2
	夏目-太宰	6
	芥川-太宰	8
黒	夏目-芥川	9
	夏目-太宰	-5
	芥川-太宰	-14
青	夏目-芥川	14
	夏目-太宰	4
	芥川-太宰	-10



一例：ユークリッド距離

それぞれの項目における差の2乗和の平方根

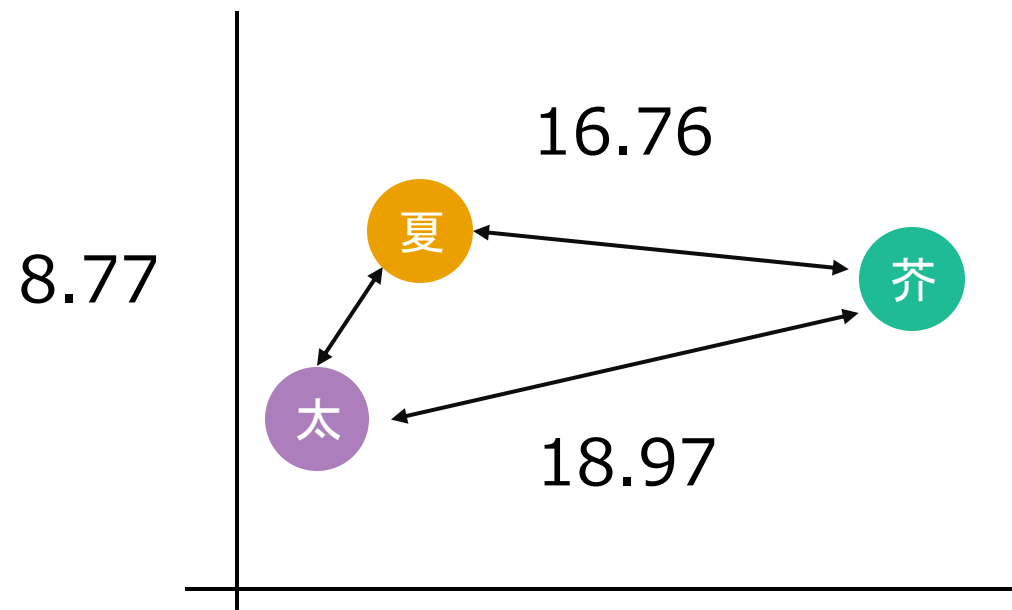
$$d(A, B) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

11. テクスト分析演習 9 (階層的クラスタリング)

階層的クラスタリングとは

赤	夏目-芥川	-2
	夏目-太宰	6
	芥川-太宰	8
黒	夏目-芥川	9
	夏目-太宰	-5
	芥川-太宰	-14
青	夏目-芥川	14
	夏目-太宰	4
	芥川-太宰	-10

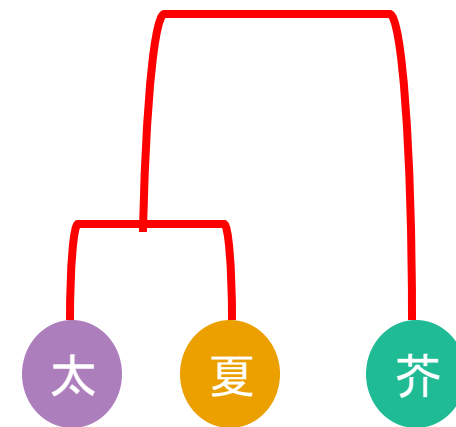
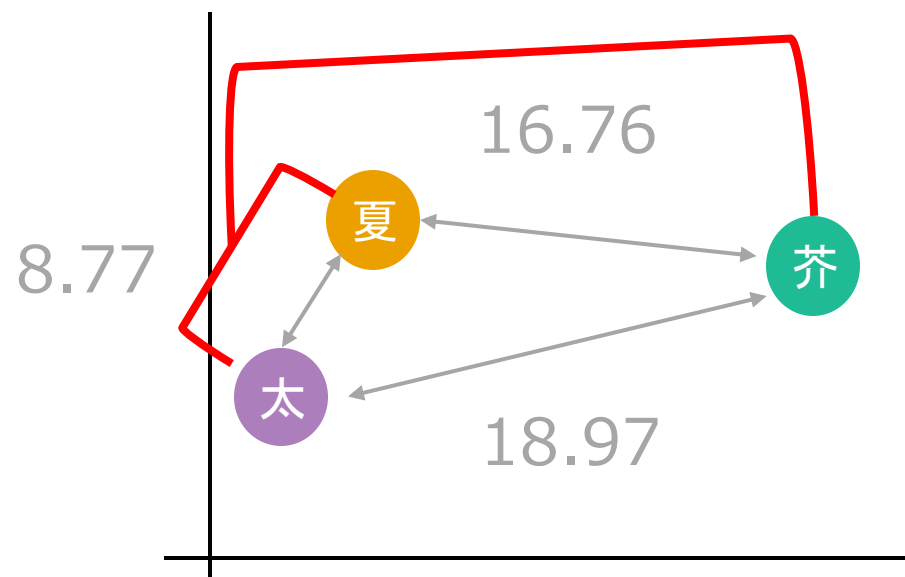
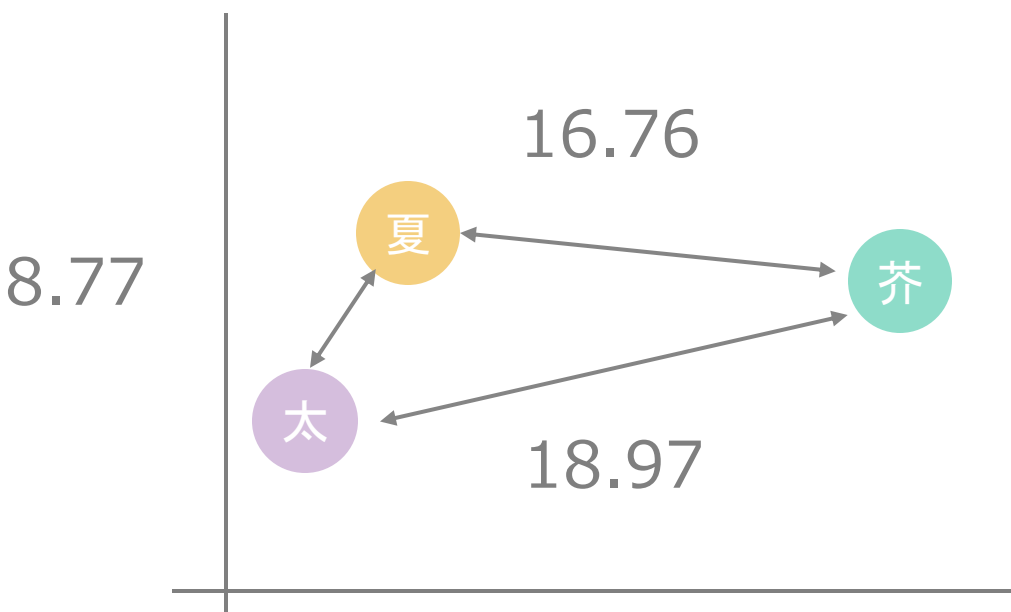
- 夏目と芥川の距離 = 16.76
- 夏目と太宰の距離 = 8.77
- 芥川と太宰の距離 = 18.97



11. テキスト分析演習 9 (階層的クラスタリング)

階層的クラスタリングとは

- 夏目と芥川の距離 = 16.76
- 夏目と太宰の距離 = 8.77
- 芥川と太宰の距離 = 18.97



11. テキスト分析演習 9 (階層的クラスタリング)

階層的クラスタリングの応用

- Cluster_Analysis.ipynb

11-1. 英語・日本語コーパスの階層的クラスタリングの結果（語でzスコア化, ユークリッド距離, ウォード法）をpdfとして保存し, MS Wordに挿入。その下に, 結果から読み取れることを記述してください。ワードファイルを保存する際には以下の名前で保存し, ワードファイルを提出してください。

“ClusterAnalysis_IkuFujita.docx”

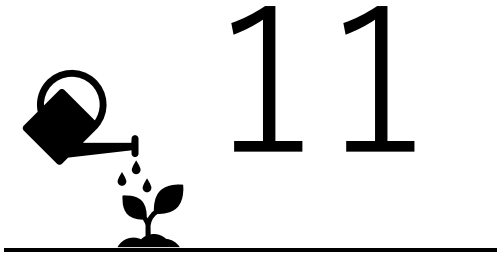
11. テキスト分析演習 9 (階層的クラスタリング)

階層的クラスタリングの応用

- `Cluster_Analysis.ipynb`
- ヒートマップを階層的クラスタリングを使用して描画

テキスト分析演習

9
(機械学習)



引用文献・参考文献

《参考文献等》

- 村上征勝. (1994). 『真贋の科学—計量文献学入門—』. 東京：朝倉書店.
- 石川慎一郎, 前田忠彦, 山崎誠 (編). (2010). 『言語研究のための統計入門』. 東京：くろしお出版.
- 小杉考司. (2018). 『言葉と数式で理解する多変量解析入門』. 京都：北大路書房.

テキスト分析演習

10

(機械学習)

12



コレスポネンス分析とは…？

- 表形式のデータの行と列の関係性を低次元（通例多くても3次元）のグラフ上に描画することで，行と列（例：作品と単語）の関係性を捉えやすくする手法

	黒	赤	緑	黄	白
芥川A		✓			
芥川B	✓	✓	✓	✓	
夏目A				✓	✓
夏目B					✓
太宰A		✓		✓	
太宰B	✓				

	黒	赤	緑	黄	白
芥川A		✓			
芥川B	✓	✓	✓	✓	
夏目A				✓	✓
夏目B					✓
太宰A		✓		✓	
太宰B	✓				

並べ替え1

並べ替え2

	黒	赤	緑	黄	白
太宰B	✓				
芥川A		✓			
芥川B	✓	✓	✓	✓	
太宰A		✓		✓	
夏目A				✓	✓
夏目B					✓

	黒	赤	緑	黄	白
太宰B	✓				
太宰A		✓		✓	
芥川B	✓	✓	✓	✓	
芥川A		✓			
夏目B					✓
夏目A				✓	✓

	黒	赤	緑	黄	白
芥川A		7			
芥川B	13	3	5	6	
夏目A				4	4
夏目B					4
太宰A		6		6	
太宰B	14				

並べ替え
計算1

	黒	赤	緑	黄	白
1回目	3	2	4	-1	-5

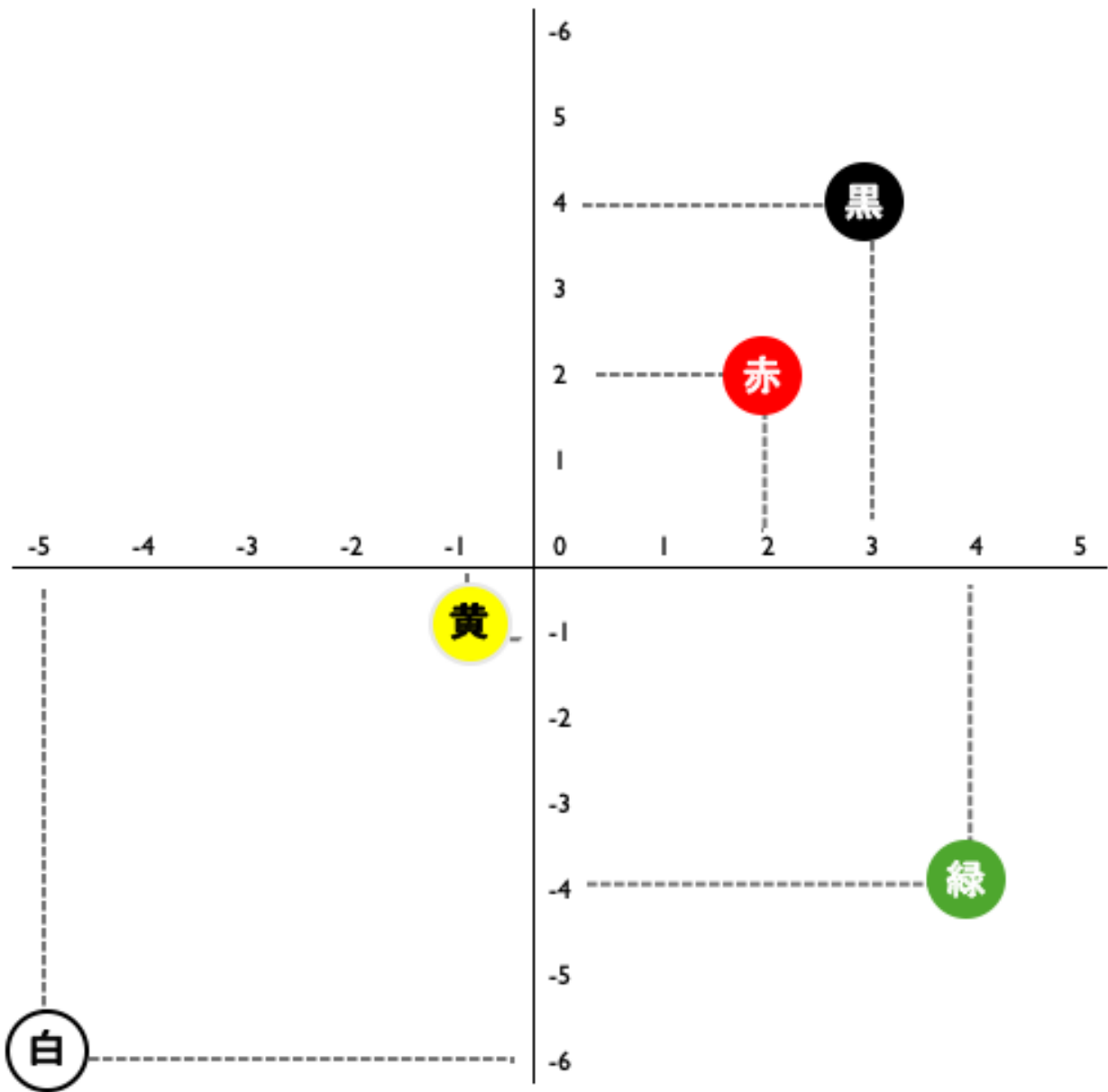
	太宰A	太宰B	芥川A	芥川B	夏目A	夏目B
1回目	-3	-1.5	1	2	-2	-4

並べ替え
計算2

	黒	赤	緑	黄	白
2回目	4	2	-4	-1	-6

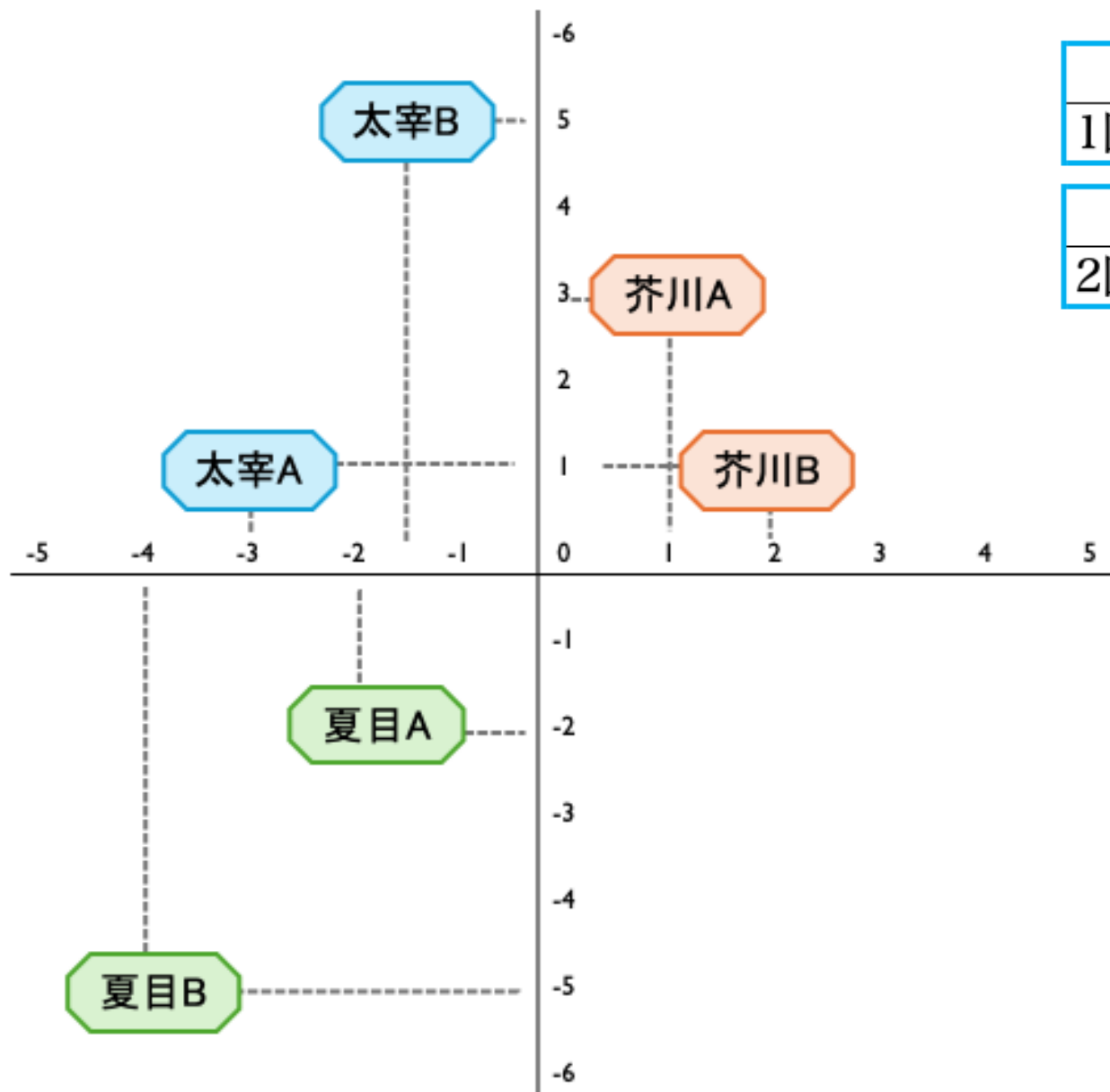
	太宰A	太宰B	芥川A	芥川B	夏目A	夏目B
2回目	1	5	3	1	-2	-5

12. テクスト分析演習 10 (コレスポネンス分析)



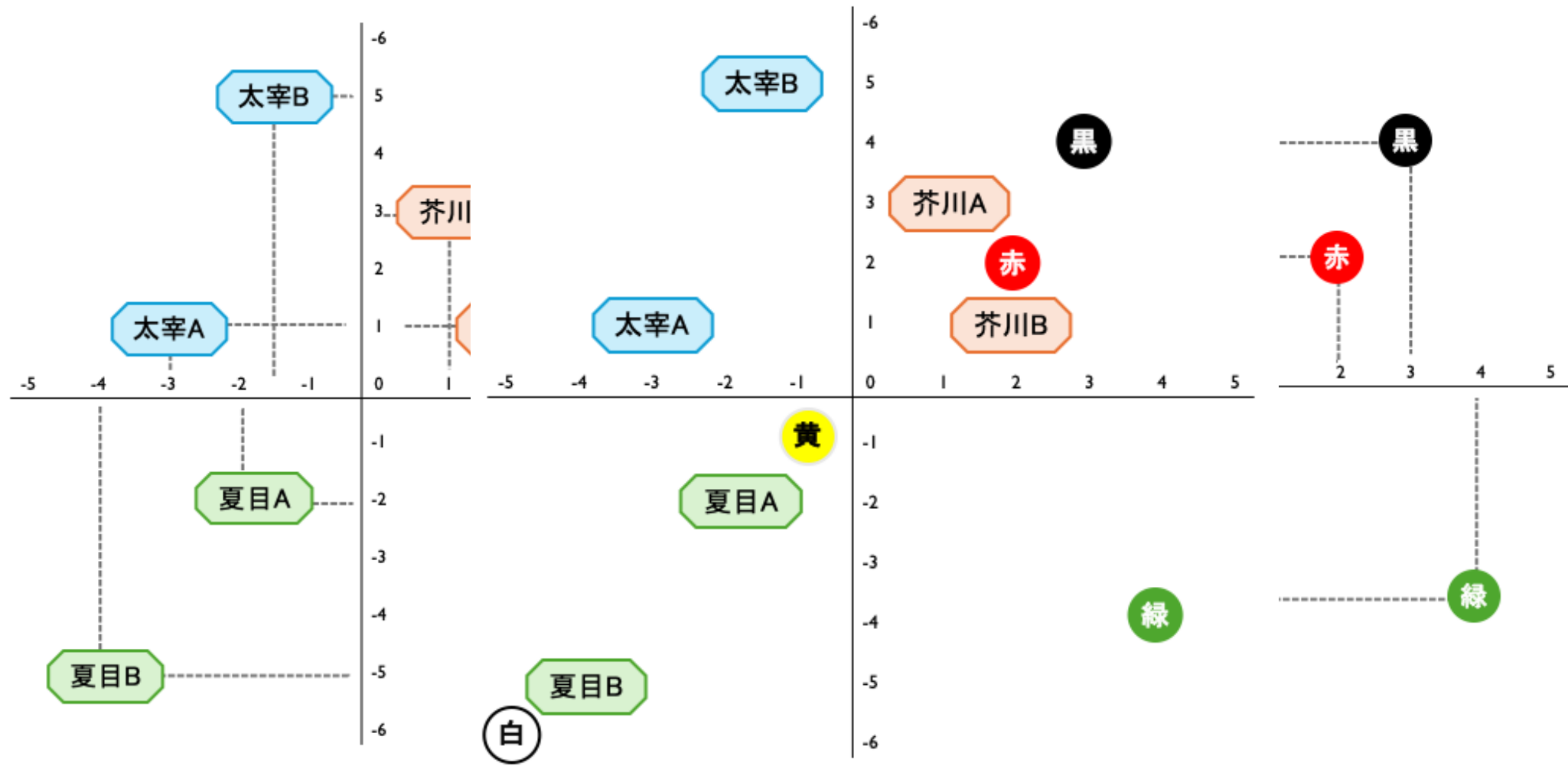
	黒	赤	緑	黄	白
1回目	3	2	4	-1	-5

	黒	赤	緑	黄	白
2回目	4	2	-4	-1	-6



	太宰A	太宰B	芥川A	芥川B	夏目A	夏目B
1回目	-3	-1.5	1	2	-2	-4

	太宰A	太宰B	芥川A	芥川B	夏目A	夏目B
2回目	1	5	3	1	-2	-5



コレスポネンス分析

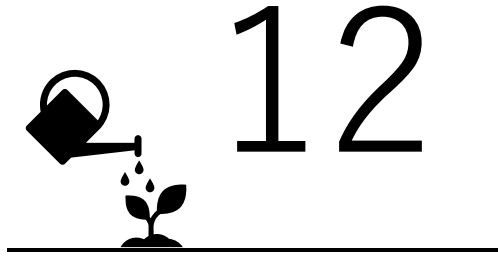
- Correspondence_Analysis.ipynb

12. 英語・日本語コーパスのコレスポネンス分析の結果（をpdfとして保存し、MS Wordに挿入。その下に、結果から読み取れることを記述してください。ワードファイルを保存する際には以下の名前で保存し、ワードファイルを提出してください。

“CorrespondenceAnalysis_IkuFujita.docx”

テキスト分析演習

10
(機械学習)



引用文献・参考文献

《参考文献等》

- 村上征勝. (1994). 『真贋の科学—計量文献学入門—』. 東京：朝倉書店.
- 石川慎一郎, 前田忠彦, 山崎誠 (編). (2010). 『言語研究のための統計入門』. 東京：くろしお出版.

テキスト分析演習

11

(機械学習)

13



- 複数の文書 (セグメント) に跨って現れる語を, その頻度を基に共起傾向の確率を計算しグループ分けする手法
- トピックモデルで言及されるトピックとは, 語群 (話題や主題といった他分野で使用されている意味を示すものではない)
- 学習モデルや辞書なしで潜在的な意味にアプローチできる
- ここでは潜在的ディリクレ配分法(LDA; latent Dirichlet allocation)

“計量的分析からの Tennyson における immortality 省察” (研究発表レジュメより)

日本英文学会中国四国支部第76回大会

- 共起：セグメント（文書）単位--
 > BoW (Bag of Words)
- 綴りは同じ，意味が違ふものはトピック（語群）は分かれる

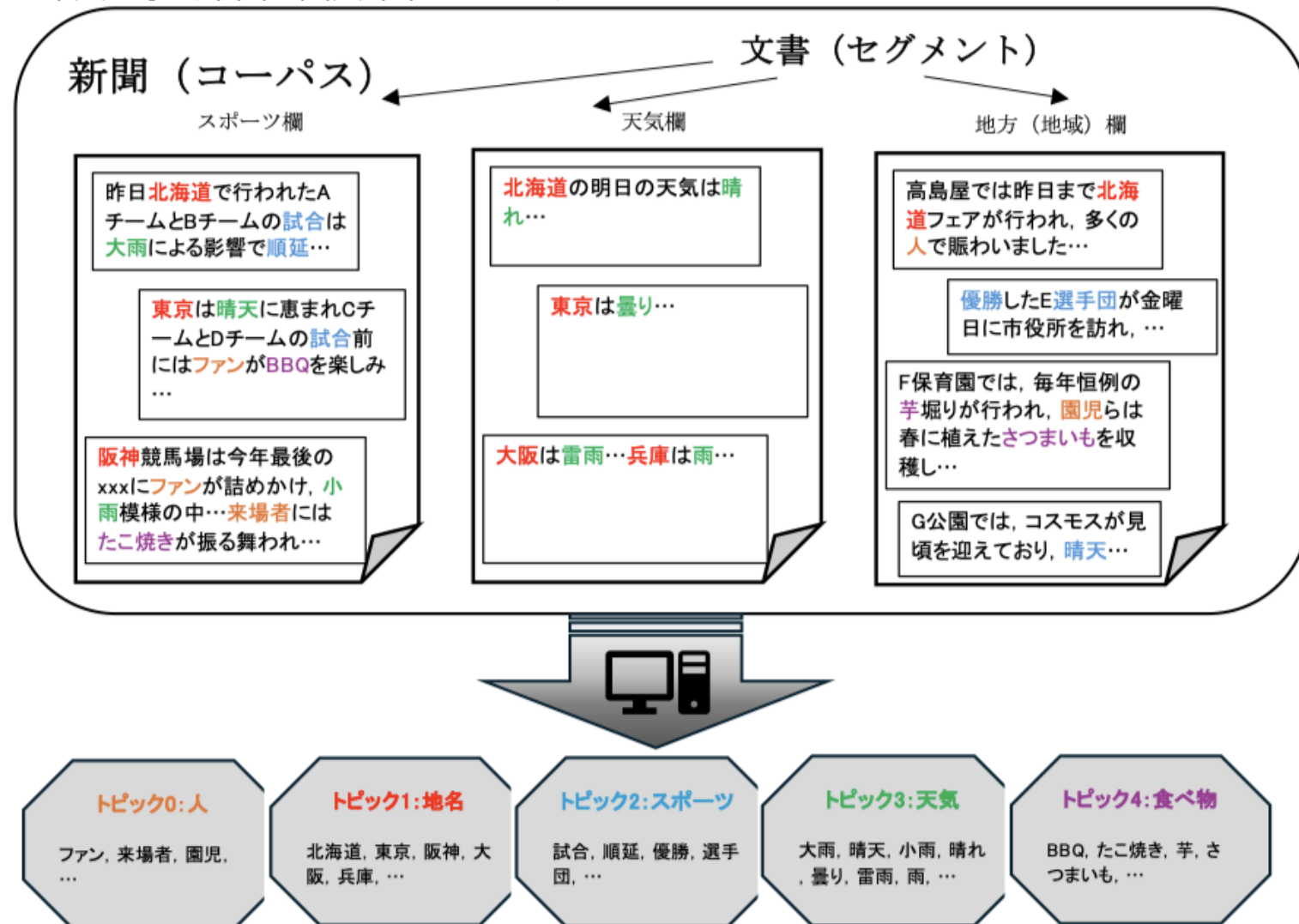
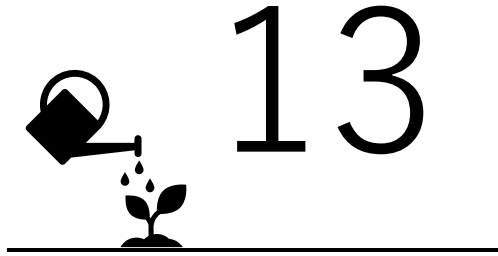


図 1 LDA の概念図

テキスト分析演習

11
(機械学習)



引用文献・参考文献

《参考文献等》

藤田郁. (2024). 「計量的分析からの Tennyson における immortality 省察」. 日本英文学会中国四国支部第76回大会. 於就実大学.

まとめ

人文情報学と
従来の人文学の融合

15



[再掲]

デジタル人文学（デジタルヒューマニティーズ・人文情報学）：

- **蓄積系**：データベースの作成，アーカイブ等
 - 例：コーパス作成等（本講義のデータベースとアノテーションで少し触れます）
- **解析系**：史資料を分析・解析し，新たな知見を得る
 - 例：画像解析，テキスト分析等
- **可視化系**：（解析系で得られた）知見や研究成果を可視化
 - 例：地図やVR(バーチャルリアリティー；Virtual Reality;仮想現実)等

[再掲]

従来の「人文学」研究

- 研究者（ら）の知識，経験をフル活用した質的な研究
 - データの量的な傾向ではなく，意味・文脈・主観的な判断を重視
 - 比喩，隠喩，（暗示された）主題など、定量化しづらい要素に注目
- 個別作品の精読（close reading）
 - 文学作品や歴史文書，美術作品などを時間をかけて丹念に読み解く
 - 文体，語彙，修辞技法，主題などに注目
 - 解釈や批評を通じて作品の意味や価値を探る

[再掲]

従来の「人文学」研究

- 歴史的・哲学的・文化的文脈の重視（精読に含まれる場合も有）
 - 作品が生まれた時代や社会，思想背景を深く掘り下げる
 - 作者の伝記や政治・宗教的影響なども考慮に入れる
- 手作業による資料収集・分析
 - 古文書，写本，原典資料などを図書館やアーカイブで実際に調査
 - 注釈作成，資料比較，翻訳なども手作業で行う
- 解釈の多様性と主観性
 - 明確な「正解」があるわけではなく，解釈に幅がある
 - 研究者の視点や理論（例：フェミニズム批評、ポストコロニアル理論など）が研究の立脚点

[再掲]

人文学と情報学を掛け合わせることの意義

- 客観的, 科学的視点を取り入れることができる
 - 「〇〇だと思う」に（科学的）な根拠をプラスできる
 - 説得力が増す

「科学」：

- **実証性 (Verifiability)**: ある仮説が（実験などで）実証できるか
- **客観性 (Objectivity)**: 得られた結果が客観的に認められるか
- **再現性 (Replicability/Reproducibility)**:
（実験など）同一条件下であれば誰が行なっても同じ結果になるか

15-1. 人文学研究に情報学の手法等を取り入れることのメリットとデメリットは何で、それらを踏まえた上で、自分だったらどのように人文学研究に情報学の手法を取り入れていきたいですか。

15-2. 本講義では人文情報学の中でも特にテキスト分析に焦点を当て、扱ったデータも文学作品という限定的なものでした。ご自身の興味・関心のある対象は何かを明記した上で、そのデータにどの手法をあてがったら面白い結果が出そうだと思うか、こんな結果が見れたら良いな、という想像でも構いませんので少なくとも1つ手法を挙げ、可能な範囲で期待される結果を書いてください。

まとめ

人文情報学と
従来の人文学の融合



15

引用文献・参考文献

《参考文献等》