



データサイエンス・リテラシー（１）

AIテキストマイニング演習

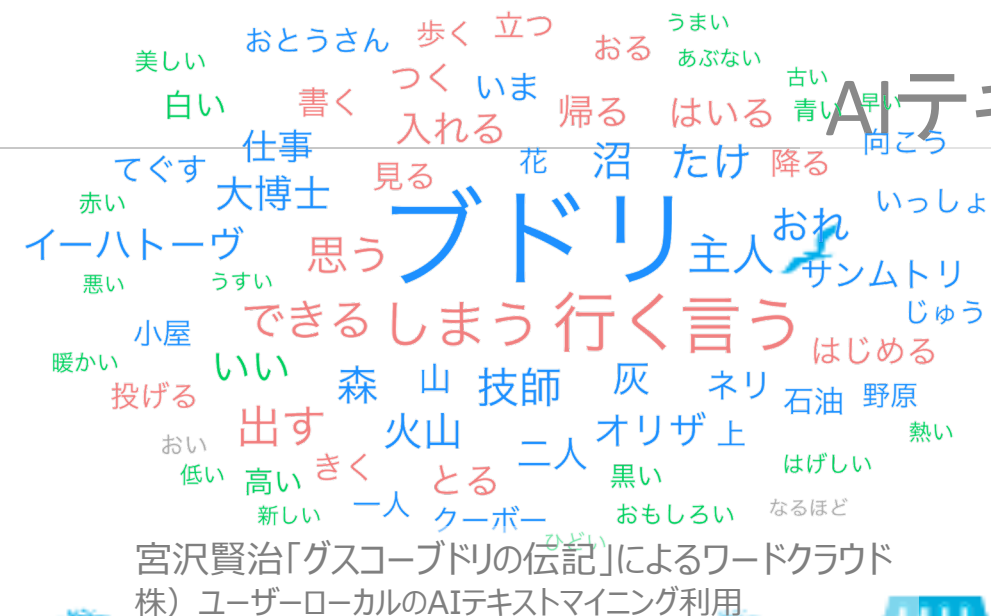


Table of Contents

オープンデータを使った分析演習

AIテキストマイニング

- ワードクラウド、共起キーワードなど

グループワーク：テキストデータの分析

分析結果の発表



テキストマイニング

テキストマイニング: **自然言語処理技術**を用いて,
自然言語で書かれた文章の解析を行い,
知識を抽出する方法

例



日記



年	月	日	場所	事項

自然言語処理

自然言語: 人間が意思疎通をするために自然発生的にできた言語
日本語, 英語, 中国語, スペイン語, ...

自然でない言語: プログラミング言語, 形式言語など,
人工的に作られた言語

自然言語をコンピュータで扱いやすい形にする処理を自然言語処理という。
形態素解析, 構文解析, 固有表現認識など様々な技術が含まれる。

主な手法

ルールベース

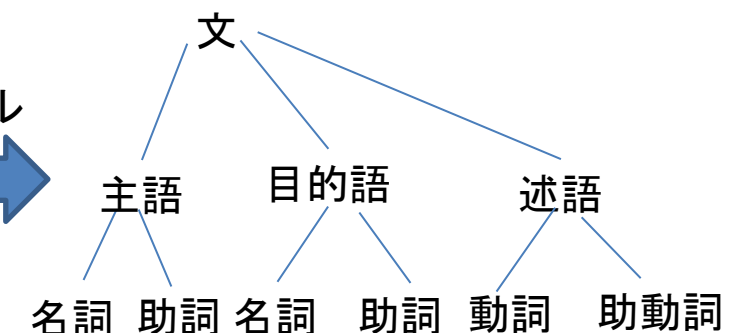
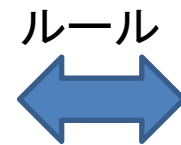
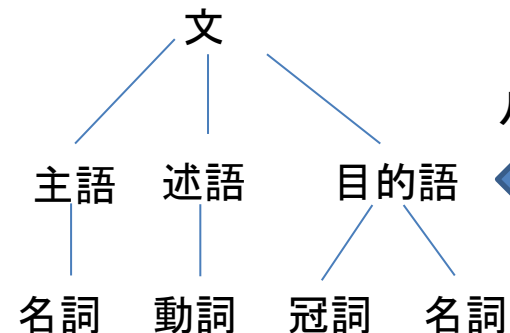
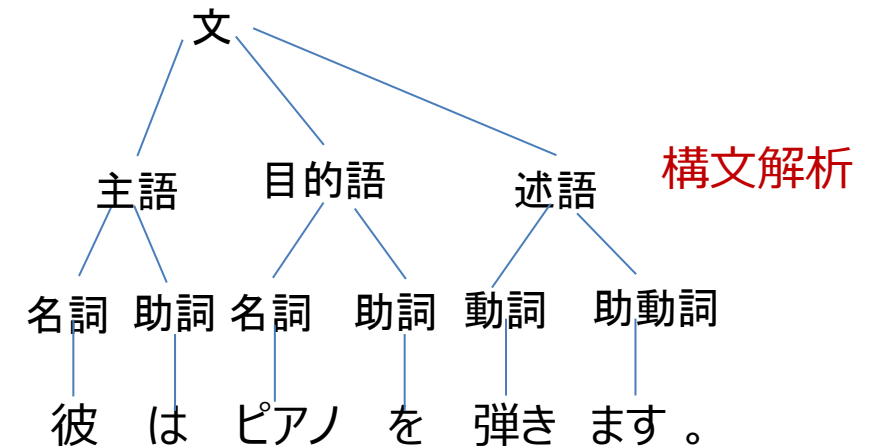
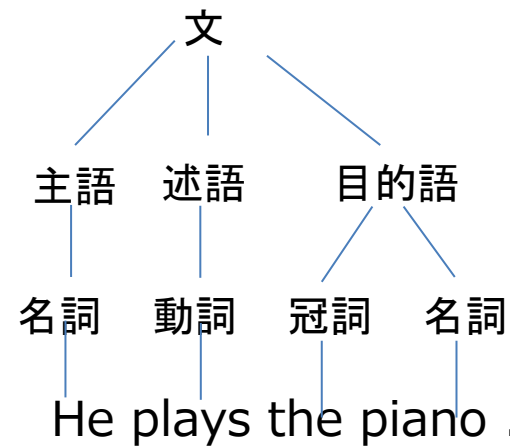
文章を単語に分解
単語・熟語の役割を確認
構文（ルール）を決定

He plays the piano.

彼はピアノを弾きます。

He plays the piano .

彼 は ピアノ を 弾き ます 。 形態素解析



AIテキストマイニング演習

分析したいテキストデータを見つける

□青空文庫

- 著作権の消滅した作品と、「自由に読んでもらってかまわない」とされた作品を、テキストとXHTML（一部はHTML）形式に電子化
- 公開されている作品を、「作家別」「作品別」「分野別」で検索可能
- 収録作品数：16075（著作権なし：15738、著作権あり：337）
(2020/07/17)
- <https://www.aozora.gr.jp/>

青空文庫 —オープンデータ

青空文庫

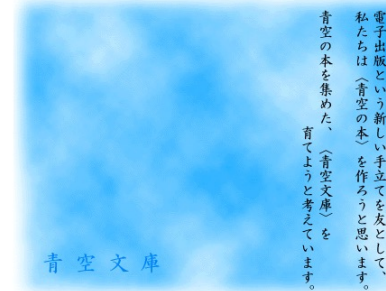
CITY
UNIV.

7C PD Locked...

www.aozora.gr.jp 内を検

インターネットの電子図書館、青空文庫へようこそ。

「青空文庫収録ファイルを用いた朗読配信をお考えのみなさまへ」



青空文庫

初めての方はまず「[青空文庫早わかり](#)」をご覧ください。ファイル利用をお考えの方は、[こちら](#)をご覧ください。ブラウザでは読みにくいと思った方は、「[青空文庫のXHTML・TEXT](#)」の方をどうぞ。「誤植かな？」とお気づきの方は、[こちら](#)を参考に報告いただけます。

メインエリア

青空文庫早わかり	青空文庫の使い方と約束事を紹介しています。初めての方、ファイルのダウンロード方法、使い方、こちらへどうぞ。									
総合インデックス	作家名、作品名の50音別に、公開作品と入力・校正作業中の作品を一覧できるインデックスです。作品を探すときは、下の近道もご利用ください。									
公開中 作家別：	あ行	か行	さ行	た行	な行	は行	ま行	や行	ら行	わ行
公開中 作品別：	あ	か	さ	た	な	は	ま	や	ら	わ
	い	き	し	ち	に	ひ	み	ゆ	る	を
	う	く	す	つ	ぬ	ふ	む	ゆ	る	ん
	え	け	せ	て	ね	へ	め	れ	る	他
	お	こ	そ	と	の	ほ	も	よ	ろ	他
作業中：	作家別・作品別									
青空文庫 分野別リスト	分野別に公開作品を一覧できる、インデックスです。									

青空文庫早わかり

総合インデックス

公開中 作家別：

公開中 作品別：

作業中：

青空文庫 分野別リスト

青空文庫

早わかり

総合インデックス

作家名、

作品を探

あ行 か行 さ行 た行 な行 は行

ま行 や行 ら行 わ行 他

あ

い

う

え

お

か

き

く

け

こ

さ

し

す

せ

そ

た

ち

つ

て

と

な

に

ぬ

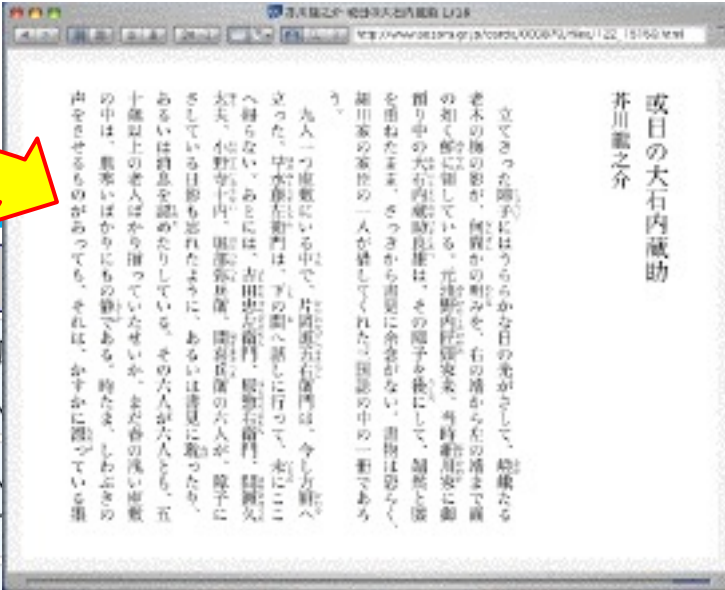
ね

の

作家別・作品別

分野別に公開作品を一覧できる、インデックスです。

作家別、作品別、分野別リストから分析する著作を
選定し、「HTMLで読む」で表示された文章から
文章を1万字程度コピーする。



AIテキストマイニング演習

 新規作成  文書一覧  カスタマイズ  前処理ツール  ヘルプ

解析したいテキストを入力する

 1つの文書を解析  2つの文章を比較  Excelクロス集計  音声入力

サンプル： 太宰治『走れメロス』

 フォーム入力  ファイルをアップロード  履歴から選択

①コピーした文章を
枠内にペースト

②1万字以下になるように
文章を部分的に削除・調整

0 / 10,000文字

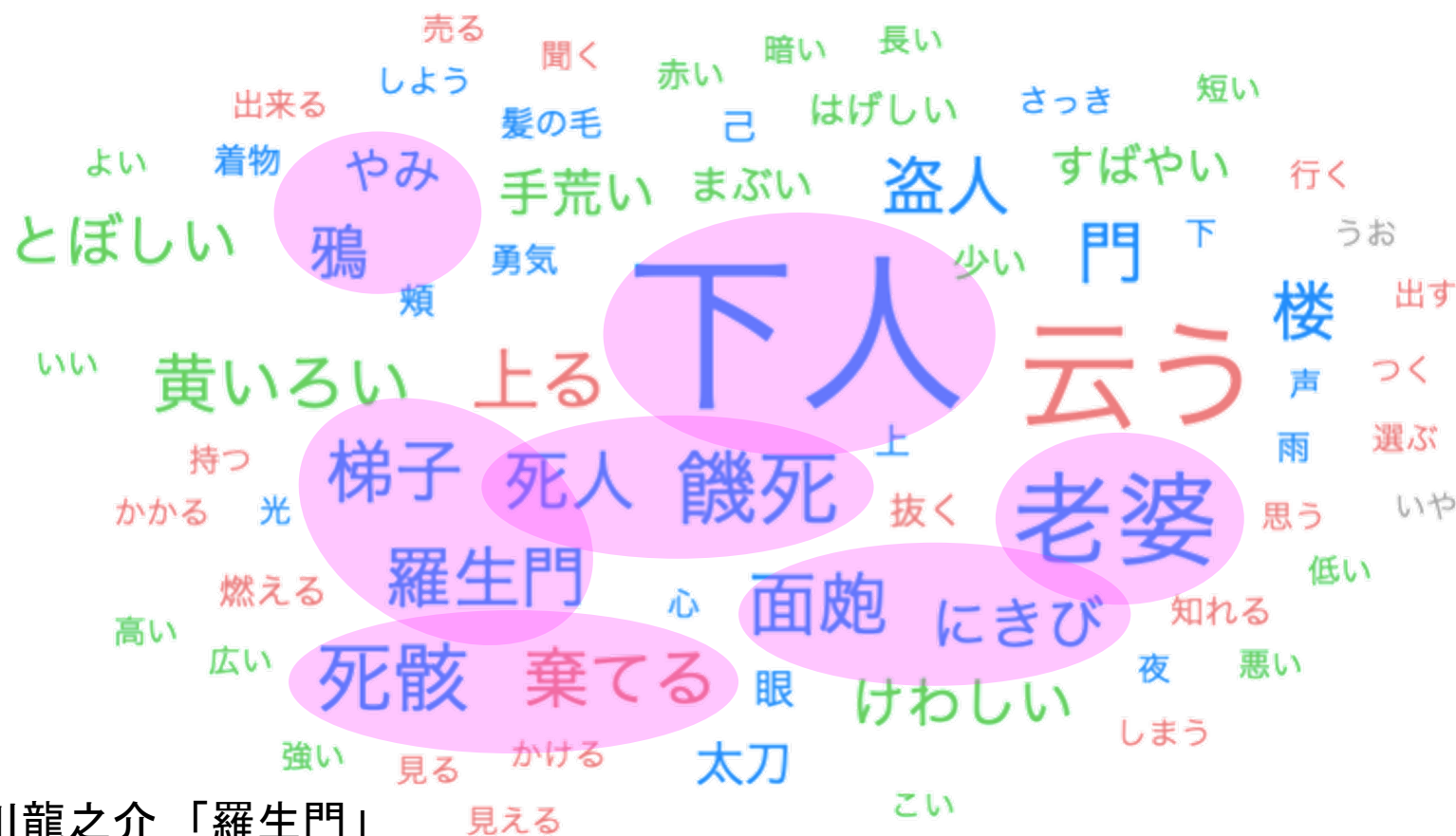
 文字数上限を増やす

テキストマイニングする

③「テキストマイニングする」ボタンを
クリックして分析実行

ワードクラウド

出現頻度が高い単語を選び、頻度に応じた大きさで図示
青色：名詞、赤色：動詞、緑色：形容詞、灰色：感動詞



- 主人公である「下人」や「老婆」の頻度が非常に高い。
- 下人の特徴である「面皷（にきび）」が頻度が高く強調されている。
- 舞台である羅生門と梯子、さらに状況を示す陰鬱な印象を与える言葉が多く使用されている。

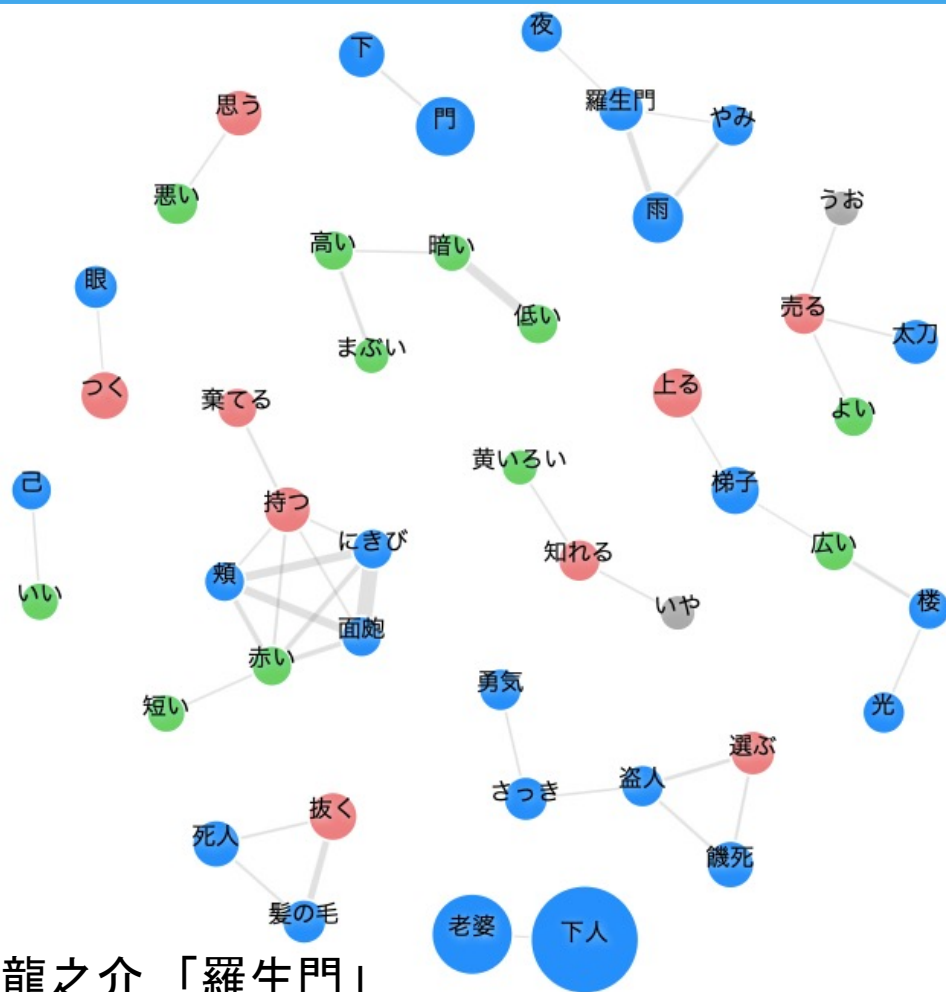
単語出現頻度

文章中出现する品詞ごとの単語の頻度表

■ 名詞	スコア	出現頻度	■ 動詞	スコア	出現頻度
下人	424.41	44	棄てる	84.94	28
老婆	164.12	15	啼く (なく)	12.84	8
餓死 (がし)	85.16	8	見開く	6.85	3
死骸 (しがい)	57.78	14	ともす	5.96	2
面皷 (にきび)	38.58	4	挿す	2.79	2
門	35.76	8		1.93	2
梯子 (はしご)	32.99	8		1.49	2
● 主人公である「下人」や「老婆」の頻度が非常に高い。 ● 下人の特徴である「面皷 (にきび)」が頻度が高く強調されている。 ● 舞台である羅生門と梯子、さらに状況を示す陰鬱な印象を与える言葉が多く使用されている。			● 一般的な「云う」、「上る」の他に鴉が「啼く」や死骸が「棄て」られ流などの退廃的な状況を示す言葉が上位に現れた。		
■ 形容詞	スコア	出現頻度	■ 感動詞	スコア	出現頻度
黄いろい	7.65	1	うお	0.09	1
とぼしい	4.26	1	いや	0.01	1
けわしい	3.93	1	---	---	---
手荒い	3.46	1	---	---	---
すばやい	2.19	1	---	---	---

芥川龍之介 「羅生門」

文章中に出現する単語の出現パターンが似たものを線で結んだ図



- メインキャストの「下人」と「老婆」とのやり取りで物語が成り立っている.
- 下人の「頬」にある「赤い」「面皰（にきび）」が何度も現れ強調される.
（「面皰」と振り仮名の「にきび」は同時に出てくるため、関係性が強く出ている.）
- 老婆の「死人」の「髪の毛」を「抜く」という動作が注目されている.
- 「羅生門」の風景として「夜」、「やみ」、「雨」が特徴的に現れる.

芥川龍之介「羅生門」

階層的クラスタリング

文章中での出現傾向が似た単語をまとまりとしてとらえられるよう樹形図で表示

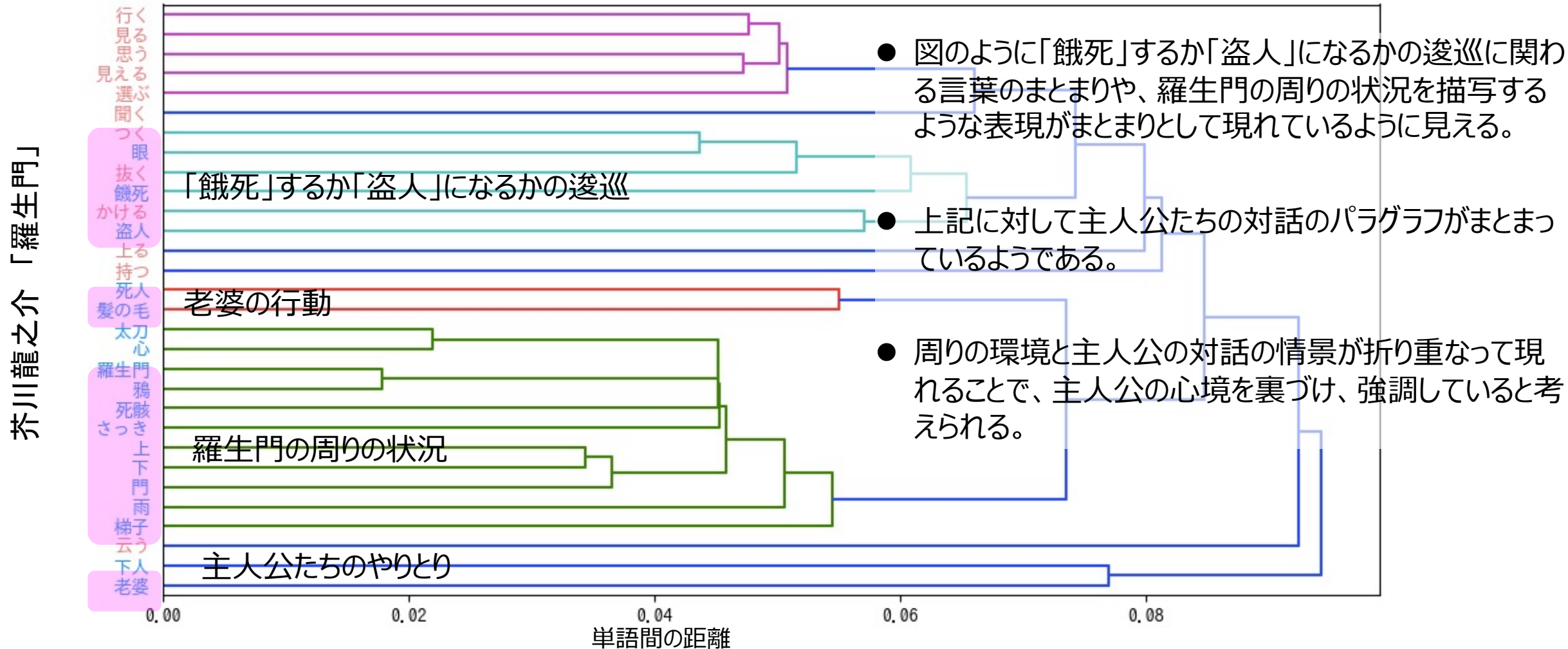


Table of Contents

オープンデータを使った分析演習

AIテキストマイニング

- ワードクラウド、共起キーワードなど

グループワーク：テキストデータの分析

分析結果の発表



AIテキストマイニング演習（グループ）

全体の流れ

- 20分
 - 1. 青空文庫から分析する**著者・作品**を選定
 - 2. テキストマイニングの**Webツール**ページに**文書**をペーストし**分析実施**
 - 3. **分析結果の表やグラフ**を画面キャプチャ等して**スライドにペースト**、スライドごとの**分析の分担を決定**
- 5分
 - 4. **それぞれ考察**を**スプレッドシートに記載**（**個人ワーク**）
- 15分
 - 5. 各メンバーの**分析結果**を**グループで発表・議論**し→**スプレッドシート**
 - 6. **議論の結果**を**スプレッドシートにまとめる**
- 30分
 - 7. **グループリーダーが発表**

詳細は次ページ



AIテキストマイニング演習（グループ）

「青空文庫」等から分析するテキストを選定

リーダーと書記を決定

（まだやったことのない人）

- Classroomにある第6回講義「AIテキストマイニング」分析シートに記載

分析する文書を「青空文庫」

<https://www.aozora.gr.jp/>の
文書から選定

- 上記シートに著者名、作品名を記載



青空文庫

PD Locked...

www.aozora.gr.jp 内を検索

Google bing goo

インターネットの電子図書館、青空文庫へようこそ。

「青空文庫収録ファイルを用いた朗読配信をお考えのみなさまへ」

電子出版という新しい手段を友として、
私たちは（青空の本）を作ろうと思ひます。
青空の本を集めた、（青空文庫）を
育てようと思ひます。

青空文庫

初めての方はまず「[青空文庫早わかり](#)」をご覧ください。

ファイル利用をお考えの方は、[こちら](#)をご覧ください。

ブラウザでは読みにくいと思った方は、「[青空文庫のXHTML・TEXTの読み方](#)」をどうぞ。

「誤植かな？」とお気づきの方は、[こちら](#)を参考に報告いただけると助かります。

メインエリア	
青空文庫早わかり	青空文庫の使い方と約束事を紹介しています。初めての方、ファイルやキャプチャーの取り扱いについて知りたい方も、こちらへどうぞ。
総合インデックス	作家名、作品名の50音別に、公開作品と入力・校正作業中の作品を一覧できるインデックスです。公開中の作品を探索するときは、下の近道もご利用ください。
公開中 作家別：	あ行 か行 さ行 た行 な行 は行 ま行 や行 ら行 わ行 他
公開中 作品別：	あ か さ た な は ま や ら わ い き し ち に ひ ゐ り ゐ う く す つ ぬ ゑ ゑ ゑ え け せ て ね へ ゑ れ お こ そ と の ほ も よ ゑ 他
作業中：	作家別・作品別
青空文庫 分野別リスト	分野別に公開作品を一覧できる、インデックスです。

#aozoraHack

AIテキストマイニング演習（グループ）

「青空文庫」等からのテキストを**10,000字以内**でコピーしてペースト

□選定した文書をコピーして
「分析ページ」にペースト

- 分析HP にアクセスして文書をペースト

<https://textmining.userlocal.jp/>

- 分析HPにアクセスするのは書記に限定
- 10,000文字の制限を満たすように適宜文書を編集

The screenshot shows the UserLocal AI Text Mining web application. At the top, there's a navigation bar with links like '新規作成' (New Creation), '文書一覧' (Document List), 'カスタマイズ' (Customize), '前処理ツール' (Preprocessing Tools), and 'ヘルプ' (Help). Below this, a header says '解析したいテキストを入力する' (Enter the text you want to analyze). There are four buttons: '1つの文書を解析' (Analyze one document), '2つの文章を比較' (Compare two sentences), 'Excelクロス集計 (ベータ版)' (Excel Cross-tabulation (Beta)), and '音声入力' (Voice input). A sample text '太宰治『走れメロス』' is shown. Below the buttons, there are three tabs: 'フォーム入力' (Form input), 'ファイルをアップロード' (Upload file), and '履歴から選択' (Select from history). The 'フォーム入力' tab is active, showing a large text area with the instruction '①ここにペースト' (Paste here). To the right of the text area, there's a note '②1万字以下になるように調整' (Adjust so it's 10,000 characters or less) with a yellow arrow pointing down to a character count '0 / 10,000文字' and a link '文字数上限を増やす' (Increase character limit). At the bottom right, there's a green button 'テキストマイニングする' (Text Mining) with a yellow arrow pointing up to it, and a note '③「テキストマイニングする」のボタンを押す' (Press the 'Text Mining' button).

AIテキストマイニング演習（グループ）

各メンバーの分析結果を共有し、グループとしてまとめる

□ 各メンバーの分析結果を共有

- ① 分析結果に対する各自の考察 ← 個人ワーク
- ② 共通の考察，独自の視点 ← グループ議論
等について議論し、シートに記載
- ③ グループワークでの貢献内容を議論後に記載（小テストの代わり）

□ グループのまとめ

- 上記、考察のまとめをシートに記載

7-1A 「テキストマイニング」 分析シート ☆ 共有

ファイル 編集 表示 挿入 表示形式 データ ツール アドオン ヘルプ 最終編集: 2分前 (tkawai_さん)

100% ¥ % .0 .00 123 Arial 24 B I S A

	A	B	C	D	E	F	G	H	I
1	ルーム01				①	②			
2	リーダー・書記 氏名	著者名	作品名	選択した分析技術	各自の考察	グループ議論での考察	グループのまとめ（書記が記載）	グループワークでの貢献内容（各自で記載）	
3									
4									
5									
6									
7									
8									
9									

ルーム01 ▼ ルーム02 ▼ ルーム03 ▼ ルーム04 ▼ ルーム05 ▼ ルーム06 ▼ ルーム07 ▼ ルーム08 ▼ 全体まとめ（編集不可） ▼

テキストマイニング分析結果の発表

□各グループでの分析結果を発表

●各グループのリーダーが説明（1分/グループ）

- 選択したテキスト（著者・作品名）
- 特徴的な分析結果や分析結果に共通するポイントなどを説明
 - －（ワードクラウド・単語出現頻度・共起キーワード・階層的クラスタリング）
- 議論した内容（結果についての考察）

●Classroomの質問用フォームに質問を記入（1人1回以上！）

- 発表時にできるだけ取り上げたいと思います。
- 時間がない場合は触れられないかもしれません。悪しからず。

データの収集

オープンデータを使った分析演習

AIテキストマイニング

- ワードクラウド、共起キーワードなど

グループワーク：テキストデータの分析

分析結果の発表



データの収集

オープンデータを使った分析演習

AIテキストマイニング

- ワードクラウド、共起キーワードなど

グループワーク：テキストデータの分析

分析結果の発表



テキストマイニング分析結果の発表

□各グループでの分析結果を発表

- 各グループのリーダーが説明（1分/グループ）
 - 選択したテキスト（著者・作品名）
 - 特徴的な分析結果や分析結果に共通するポイントなどを説明
 - －（ワードクラウド・単語出現頻度・共起キーワード・階層的クラスタリング）
 - 議論した内容（結果についての考察）