質的データの分析方法

文章データを数値的に評価するには

広島大学 AI・データイノベーション教育研究センター

村上 祐子

目標

記述アンケートなどの質的な分析対象を量的に分析する方法を理解する

この授業で紹介すること

- 文章を量的分析するための**形態素解析**
- 文章分析結果を視覚的に表す**ワードクラウド**

キーワード

形態素解析、ワードクラウド

こんなことはありませんか？

アンケートの分析結果を報告する時に…

授業の感想を記述してください。
- 面白かった。
- 分かりやすかった。
- 先生の話すスピードが速い。
- スライドが分かりにくい。
- 興味深かった。

この授業は学生にとって良い効果があった。

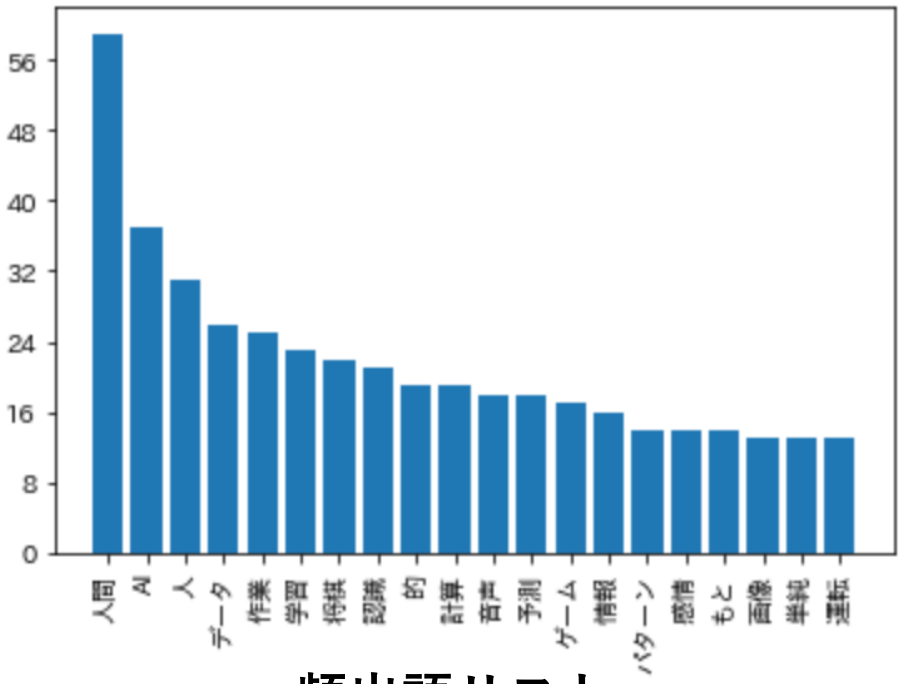
授業への不満を述べている人もいますが…。

文章の分析結果を数値的に表せると説得力が増す？

テキストマイニング

文章を分析対象として、単語や文節で区切り、出現頻度や、単語の相関などを解析することで量的な情報を取り出す方法

「AIができると思うこと」



頻出語リスト

ワードクラウド

文章から量的情報を抽出するために必要な操作

文章から量的な情報を取り出すには、文章の内容を**構造化データ**にしなければならない。

形態素解析で文章を単語に区切ることで、単語ごとに集計したり、単語を検索できる構造化データにできる。

頻出順位	単語	出現回数
1	人間	59
2	AI	37
3	人	31
4	データ	26
5	作業	25

形態素解析
文章を形態素（言語で意味を持つ最小単位）に分割し、形態素の品詞を判別する作業

形態素解析

文章を形態素（言語で意味を持つ最小単位）に分割し、形態素の品詞を判別する作業

吾輩は猫である。

吾輩 / は / 猫 / で / ある /。

[名詞] [助詞] [名詞] [助動詞] [助動詞] [記号]

形態素解析の活用例

- 検索エンジン

西条 / ~~の~~ / ラーメン / ~~屋~~

- SNSのトレンド

「☆大学に合格しました！」

「この春から☆大生です」

「☆大受かった」

「☆大学入学おめでとう」

「☆大学」がトレンド〇位！

出現頻度解析

- テキストマイニングの手法の一つ
- 形態素解析によって得られた単語の群についてそれぞれの単語の出現回数を数える

例：アンケート

「AIができると思うこと」

AIができることとして

「作業」「学習」「将棋」の印象がある

順位	単語	出現回数
1	人間	59
2	AI	37
3	人	31
4	データ	26
5	作業	25
6	学習	23
7	将棋	22

例題

中谷宇吉郎著「スポーツの科学」の文章の内容をテキストマイニングしてみましょう。

https://www.aozora.gr.jp/cards/001569/files/57459_63338.html

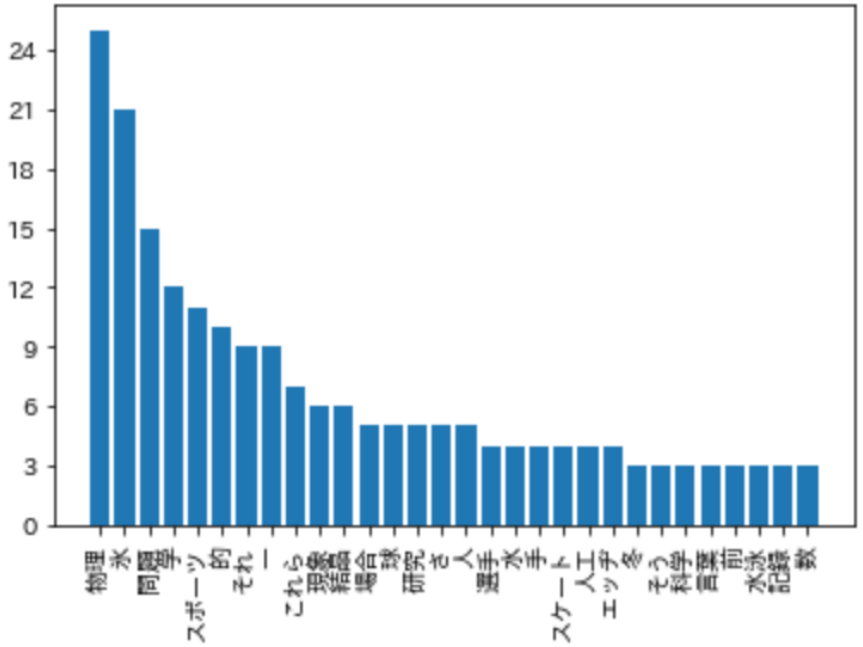
名詞の頻出語を表にまとめて、内容を推定してみましょう。

解説

形態素解析は様々なプログラム言語で開発されています。

ここではJanomeを用いて名詞の頻出語を抽出してみました。

https://mocobeta.github.io/janome/



名詞頻出語30位のうち特に以下に注目

- 氷
- スケート
- エッヂ
- 冬

ウィンタースポーツについての文章と推測

解説

実際に、中谷宇吉郎著「スポーツの科学」の中身を読んでみましょう。

*…この頃ある機会に東京の**スケート**リンクというものに初めて行ってみた。そして十年振りに**スケート**を穿いて人工の**氷**の上を滑ってみるという新しい経験を得たのである。…(中略)…**スケート**の物理学に対応して、スキーの物理学にもより以上に困難な問題がいくらも山のように聳えている。スキーの問題には昨**冬**少しばかり手をつけてみて初めて驚いたのである。…*

筆者の**スケートの経験**から物理学について考えている

テキストマイニングの結果と整合する

「テキストマイニングをする」とは？

分析対象の文章

↓ （前処理）

サンプル

↓ 形態素解析

データ

↓ 統計処理、**可視化**

情報

アンケート「AIができると思うこと」の分析結果

↓

「AIはゲーム（将棋）ができると思う。」

ワードクラウド

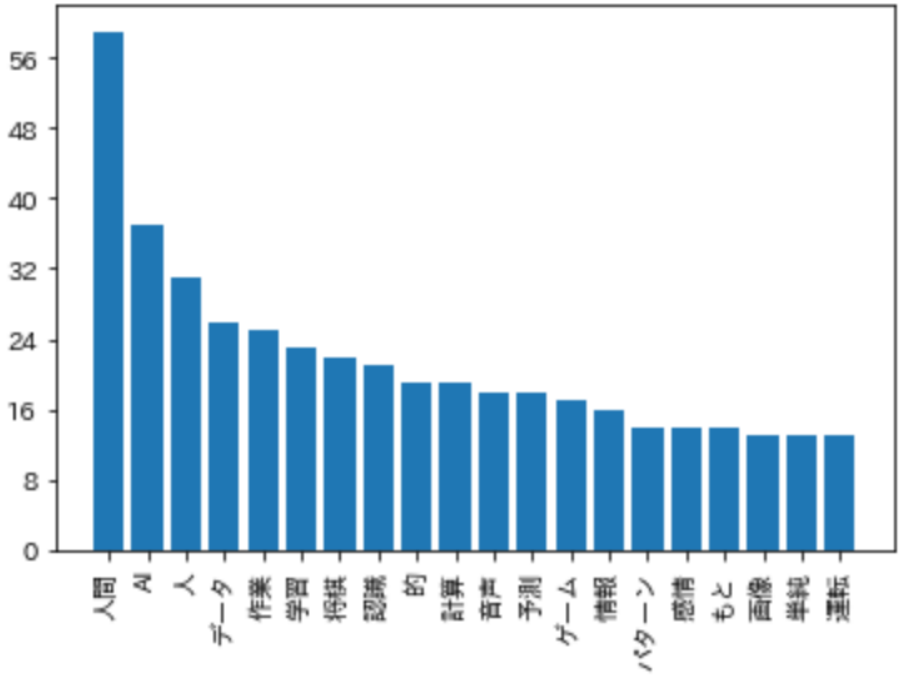
文章の出現頻度を文字の大きさで視覚化したグラフ

文字が大きい

→分析対象の文章の中で出現数が多い

比べてみましょう

頻出語のグラフ



単語の出現数が一目でわかる。

ワードクラウド

よく使われている単語が一目でわかる。

問題

1. 例題で取り上げた、中谷宇吉郎著「スポーツの科学」の文章中の名詞についてワードクラウドで表してみましょう。図示した結果から文章の傾向について何かわかりますか。
2. 例題の頻出語のグラフとワードクラウドの結果を比較してみましょう。分析から考察できることに違いがあるか考えてみましょう。