情報実験第４
ビッグデータ

全体スケジュール

- 第１回 ガイダンス・環境構築
- 第２回 Python基礎
  - 文法、データ型
  - テーブルデータ操作(pandas),グラフ
- 第３回 テキストマイニング基礎
  - 形態素解析、係り受け解析
- 第４回 文書類似度、個人課題発表
- 第５回 個人課題
- 第６回 チーム作り / チーム課題設定
- 第７回～第９回 チーム課題
- 第１０回 チーム課題発表
- Extra:共起（共起グラフ,論文紹介,word2vec)
- 第１１回まとめ(チーム課題フィードバック等）

前準備①

- プログラム保存用のフォルダを作る

  書類/exp4bd/2nd/

  #Finderで 書類 を表示して、右クリックメニュー(orメニュー [ファイル]）で[新規フォルダ]
- http://www.itpro.titech.ac.jp/exp4/からダウンロードした 品詞体系.xlsx を2ndフォルダに格納
- 2ndの下にdataフォルダを作る

#フォルダのパス・名前は任意だが、以降の説明では上記パスを利用するので、適宜読み替えること

前準備②

- 青空文庫 (http://www.aozora.gr.jp)から
  福沢諭吉 「学問のすすめ」
  のテキストファイルをダウンロードする

  青空文庫トップページ
  →右上の検索ボックスで「福沢諭吉」を検索
  →作家別作品リスト：福沢 諭吉
  →6.学問のすすめ
  → 47061_ruby_28378.zip

- 解凍してできた gakumonno_susume.txt を
  書類/exp4bd/2nd/data/
  に格納

テキストマイニング基礎

テキストマイニング

- 文字列を対象としたデータマイニング
- 通常の文章からなるデータを単語で区切り、単語の出現頻度や共出現数（共起）などを解析することで有用な情報を取り出す。

例1） 大量のアンケートの自由記述で
みんながどんなことを書いているのか知りたい
→ 多く使われている単語をリストアップ
Q: 大学生活の印象を教えてください
→頻出語：楽しい/大変/テスト/バイト
例2) ホテルの口コミデータを分析して改善に役立てたい
→ よく使われる単語の組み合わせ(共起）を調べる
→ 頻出共起 : 価格-高い, ベッド-狭い, メニュー - 少ない

テキストマイニング

<<本実験で扱う技術>>

- 形態素解析
  →テキストマイニングの基礎。広く使われている
  Mecabを使う
- 係り受け解析
  → 取り扱いが難しいので、どんなものかCaboChaを
  触ってみる程度。

係り受け解析

文A:「私は発表した」
文B:「 私は発表したA君と話した」

それぞれ 「発表した」のは誰か？
「私」は何をした？

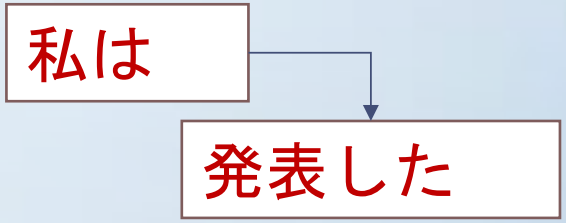
Let's try it

係り受け解析

- ターミナル で cabohca と入力
- 以下を入力

私は発表した

```
cokeMBP13-2016:~ coke$ cabocha
私は発表した
    私は-D
  発表した
EOS
```



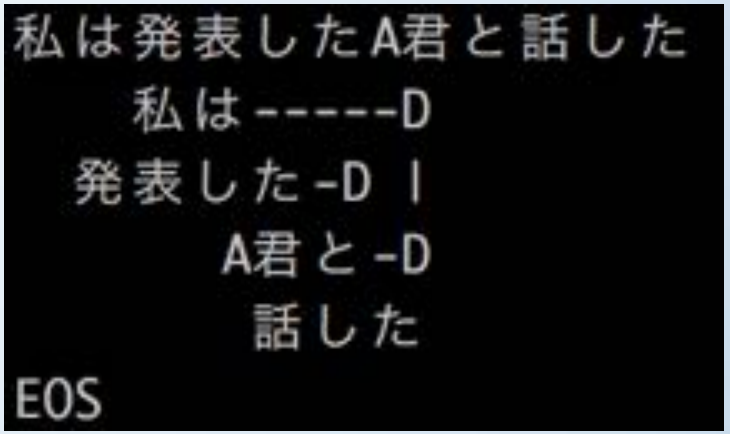
「発表した」の「私」
「私」は「発表した」

Let's try it

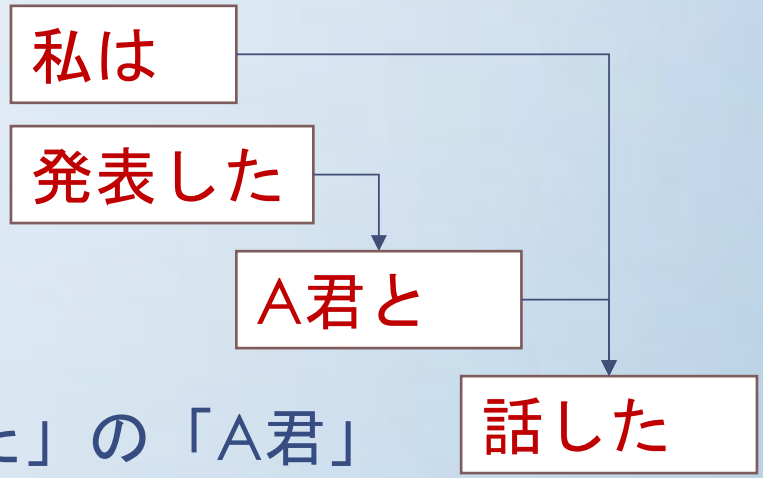
係り受け解析

- 続けて以下を入力

私は発表したA君と話した



- 結果が表示されたらcontrol+cでmecabを終了させる



「発表した」の「A君」
「私」は「話した」

形態素解析

- 形態素解析：テキストデータ（文）を単語（言語で意味を持つ最小単位,形態素,Morpheme）に分割し、それぞれの形態素の品詞等を判別する
  →単語がスペースで区切られる英語に比べて、
  日本語は単語の区切りが不明確。
  「東京都」は「東京　都」？「東　京都」
  →本質的に曖昧だが、普通は「東京　都」だと判断
  →Mecabなどの形態素解析器では、
  語の使われ易さや前後との繋がり易さをもとに
  機械的に判断する
  ※「東京」「都」とするか
  「東京都」と１語とするかは利用する辞書による

Let's try it

形態素解析の実行①

- ターミナル で
  cabohca
  と入力
- 以下を入力し、結果が表示された control+cで終了させる

```
cokeMBP13-2016:~ coke$ mecab
私は発表した
私	名詞,代名詞,一般,*,*,*,私,ワタシ,ワタシ
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
発表	名詞,サ変接続,*,*,*,*,発表,ハッピョウ,ハッピョー
し	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
EOS
```

Let's try it

形態素解析の実行①

```
cokeMBP13-2016:~ coke$ mecab
私は発表した
私	名詞,代名詞,一般,*,*,*,私,ワタシ,ワタシ
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
発表	名詞,サ変接続,*,*,*,*,発表,ハッピョウ,ハッピョー
し	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
EOS
```

表層形	品詞	品詞細分類1	品詞細分類2	品詞細分類3	活用形	活用型	原形	読み	発音
私	名詞	代名詞	一般	*	*	*	私	ワタシ	ワタシ
は	助詞	係助詞	*	*	*	*	は	ハ	ワ
発表	名詞	サ変接続	*	*	*	*	発表	ハッピョウ	ハッピョー
し	動詞	自立	*	*	サ変・スル	連用形	する	シ	シ
た	助動詞	*	*	*	特殊・タ	基本形	た	タ	タ

Let's try it

品詞体系

- Mecabの品詞体系は「IPA品詞体系」(をいくつか簡略化したもの)

→ 資料：品詞体系.xlsx

・テキストマイニングでは、多くの場合「名詞」「動詞」「形容詞」に着目して、文書の特徴を抽出する
　# 助詞、助動詞、接続詞等はどの文書にも頻出するため特徴にならないため。

Let's try it

形態素解析の実行②

- Anconda Navigaterからjupyter notebookを起動
- Documents/exp4bd/2nd/を開く
- [New | Python3]を選択
- [File | Rename]で適当な名前で保存する

→以降は
jupyter notebook で説明