



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 講師紹介,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①
 - 形態素解析、係り受け解析
- 第4回 **テキストマイニング基礎② / 個人課題**
 - **文書類似度**
- 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
- 第12回 まとめ(課題フィードバック等)

前準備

- プログラム保存用のフォルダを作る

書類/exp4bd/3rd/

#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

dataフォルダを3rdフォルダに格納

#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

テキストマイニング基礎 ～文書類似度～



文書の特徴量

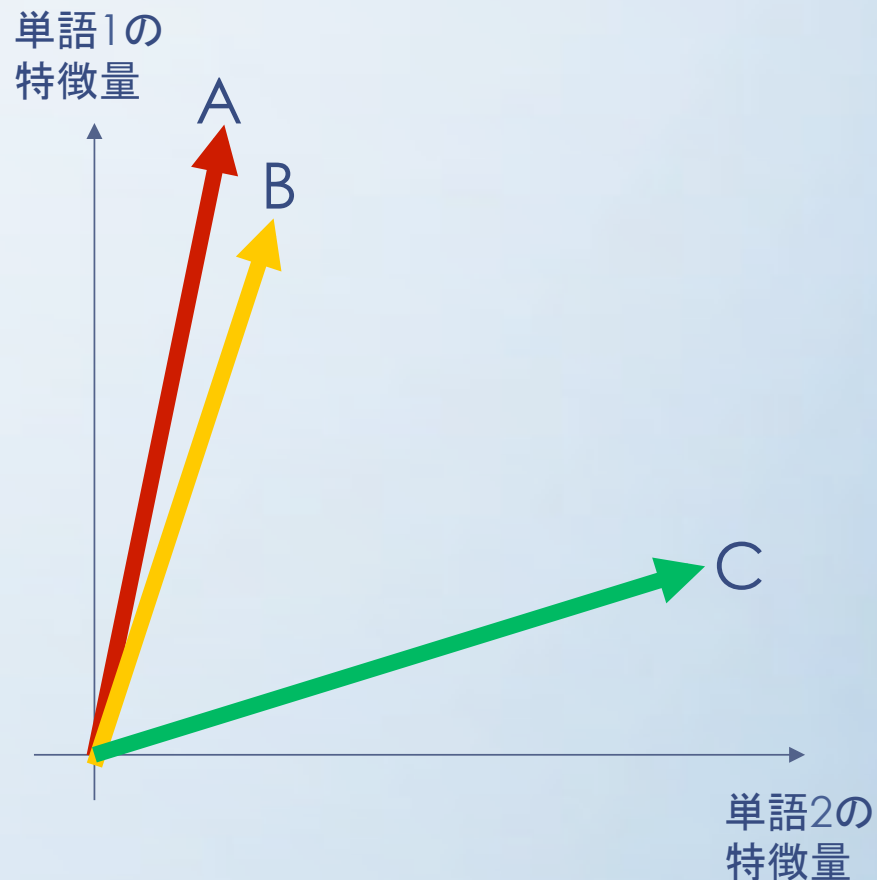
- 前回、単語の出現回数を調べることで文書の特徴を把握した
- 複数の文書を比較するとき、単語の出現回数で比較することは適切なのか？
 - どの文書でも共通的によく使われる単語より、ある特定の文書にしか出現しない語のほうが、より、その文書の特徴を表しているのでは？

文書の特徴量：TF-IDF

- 文書 $d \downarrow j$ 中の単語 $t \downarrow i$ の特徴量には、TF-IDFがよく用いられる
 - $tfidf \downarrow i, j = tf \downarrow i, j \cdot idf \downarrow i$
- TF(Term Frequency:単語出現頻度)
 - その文書の中で特定の単語が出現した回数
 - $tf \downarrow i, j = n \downarrow i, j / \sum \downarrow k n \downarrow k, j$ =文書 $d \downarrow j$ における単語 $t \downarrow i$ の出現回数/文書 $d \downarrow j$ におけるすべての単語の出現回数の和
- IDF(Inverse Document Frequency:逆文書頻度)
 - $idf \downarrow i = \log |D| / |\{d: d \ni t \downarrow i\}|$ =総文書数/単語 $t \downarrow i$ を含む文書数
- idfがにより、多くの文書に出現する語（一般的な語）は重要度が下がり、特定の文書にしか出現しない単語の重要度を上げる役割を果たす

文書の特徴量ベクトル

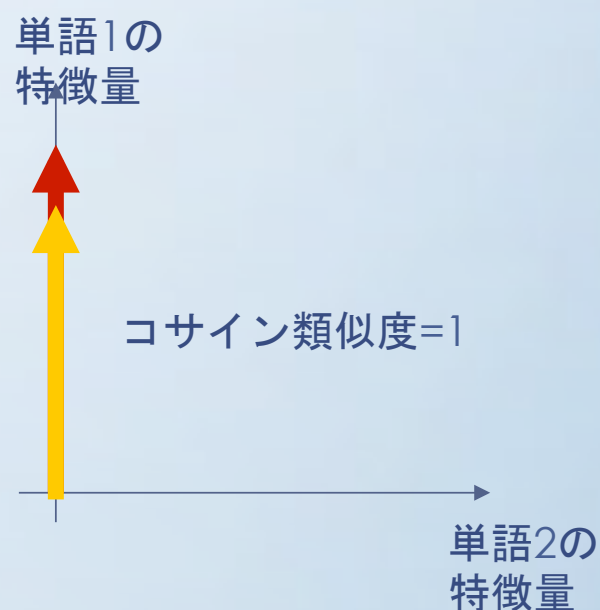
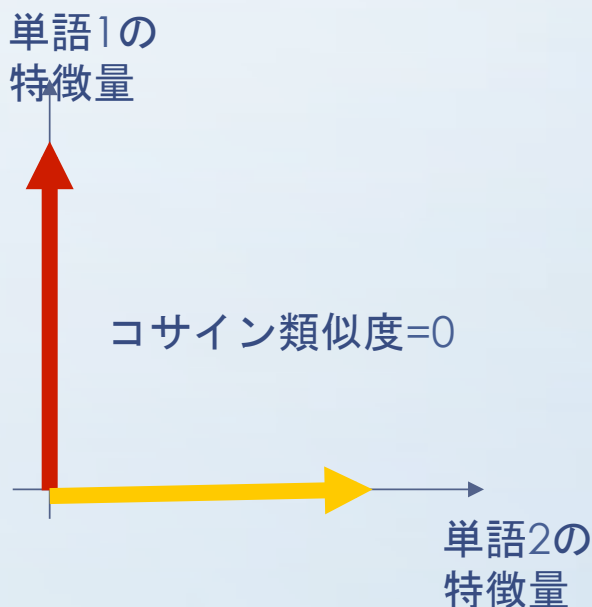
文書	単語1	単語2
A	1	0.3
B	0.8	0.4
C	0.3	0.9



文書Bは文書Cよりも文書Aに似ている

文書の類似度

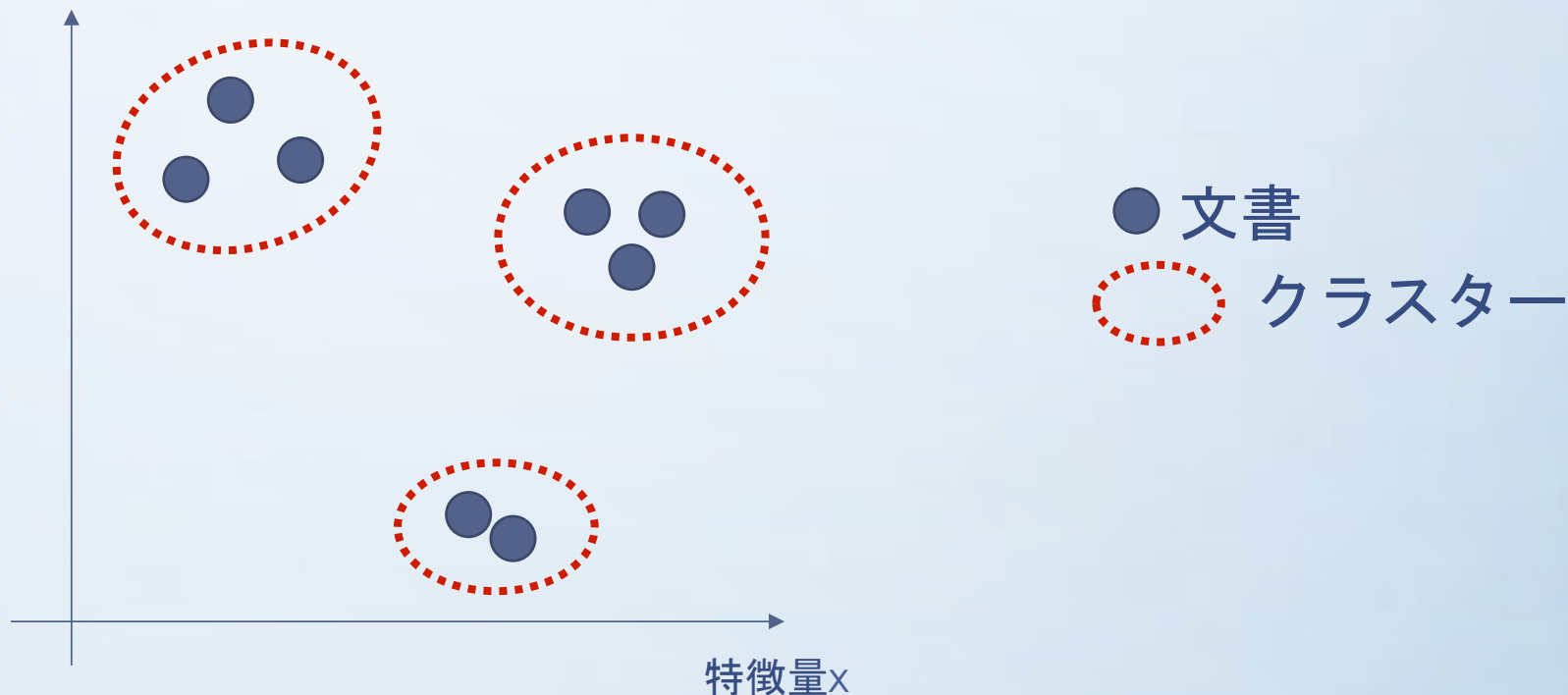
- 文書の類似度計算では、コサイン類似度がよく使われる
 - ベクトル同士のなす角度の近さ。
1に近ければ類似しており、0に近ければ似ていない



文書クラスタリング (教師なし学習)

特徴量y

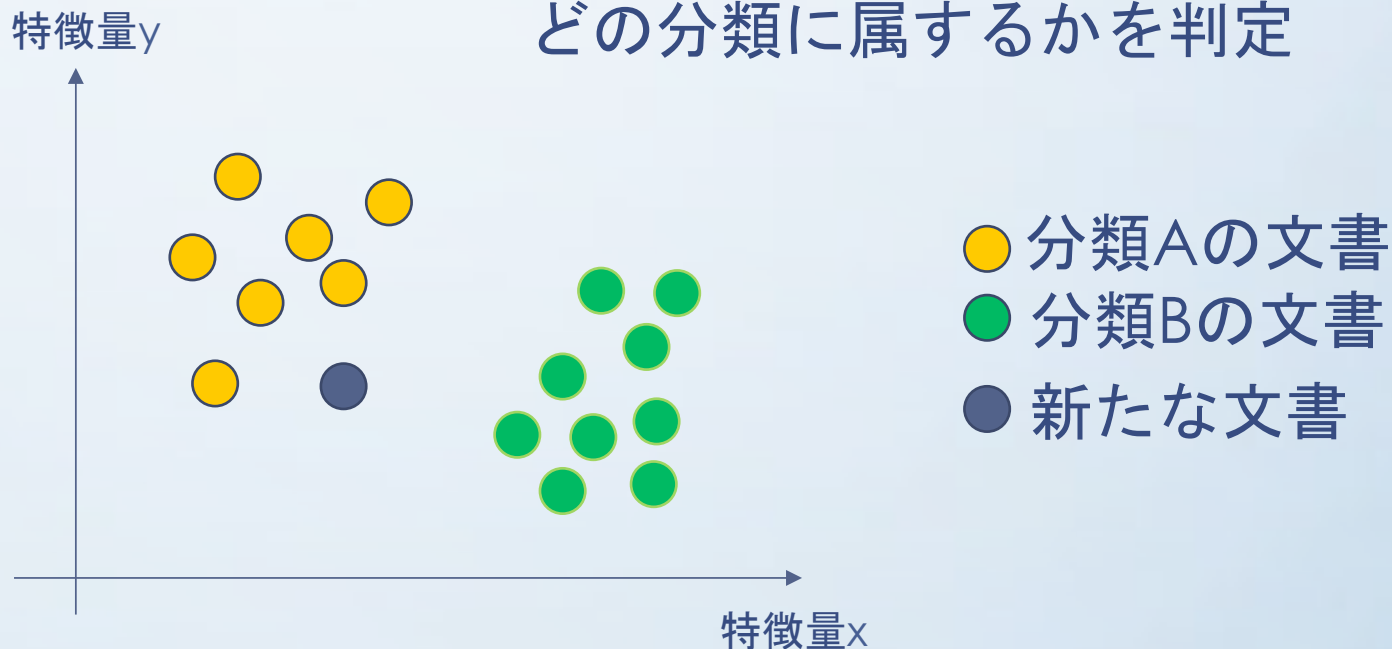
文書群を似ている部分集合に分ける



クラスタリング手法にはk-means, Ward法など、
様々なものがある

文書分類(教師あり学習)

分類済みの文書に対して、新たな文書がどの分類に属するかを判定



分類器にはナイーブベイズ、SVM(サポートベクターマシン)など、様々なものがある。

文書の特徴量抽出と 文書類似度・分類

Let's
try it

- Anaconda Navigator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