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

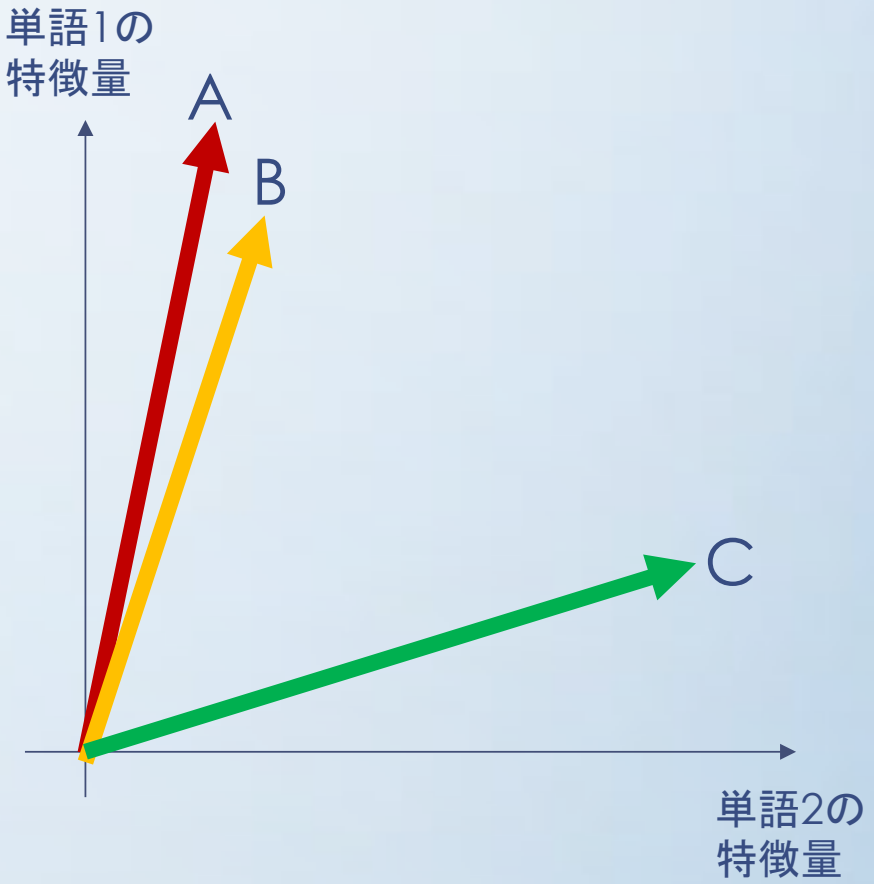
- 前回、単語の出現回数を調べることで
  文書の特徴を把握した
- 複数の文書を比較するとき、単語の出現回数で比較することは適切なのか？
  - どの文書でも共通的によく使われる単語より、
    ある特定の文書にしか出現しない語のほうが、
    より、その文書の特徴を表しているのでは？

文書の特徴量：TF-IDF

- 文書d_j中の単語t_iの特徴量には、TF-IDFがよく用いられる
  - $\mathrm{tfidf}_{i,j} = tf_{i,j} \cdot idf_i$
- 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
  - $tf_{i,j} = \dfrac{n_{i,j}}{\sum_k n_{k,j}}$ $= \dfrac{\text{文書}d_j\text{における単語}t_i\text{の出現回数}}{\text{文書}d_j\text{におけるすべての単語の出現回数の和}}$
- IDF(Inverse Document Frequency:逆文書頻度)
  - $idf_i = log \dfrac{|D|}{|\{d:d \ni t_i\}|}$ $= \dfrac{\text{総文書数}}{\text{単語}t_i\text{を含む文書数}}$
- 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

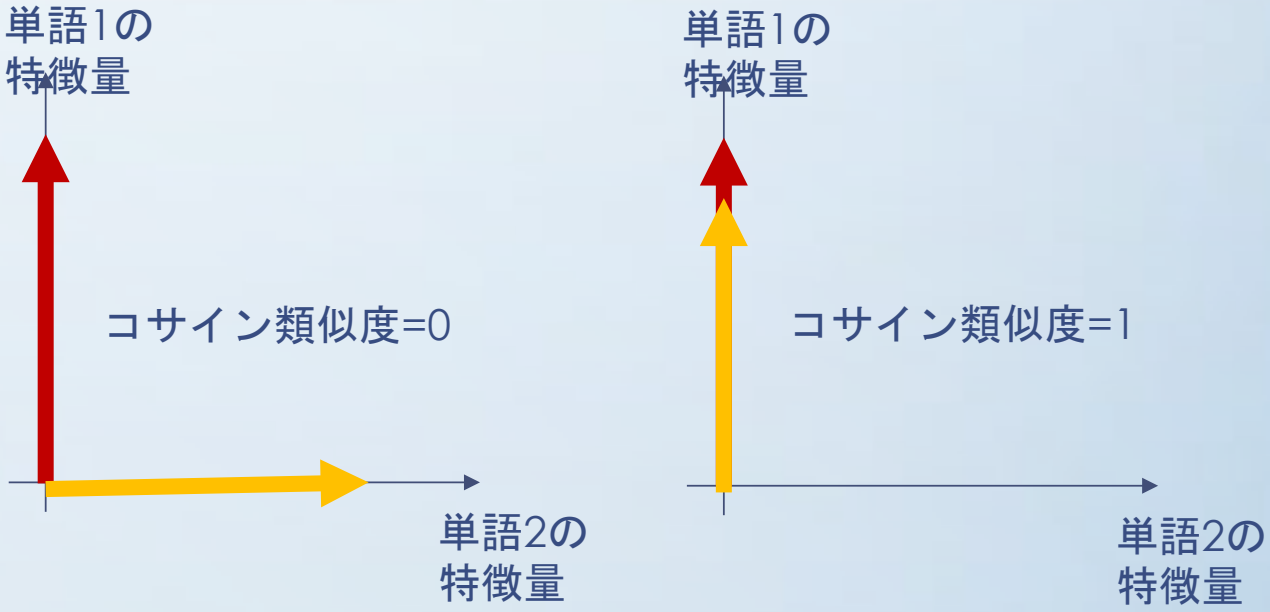
文書	単語1	単語2
A	1	0.3
B	0.8	0.4
C	0.3	0.9



文書Bは文書Cよりも文書Aに似ている

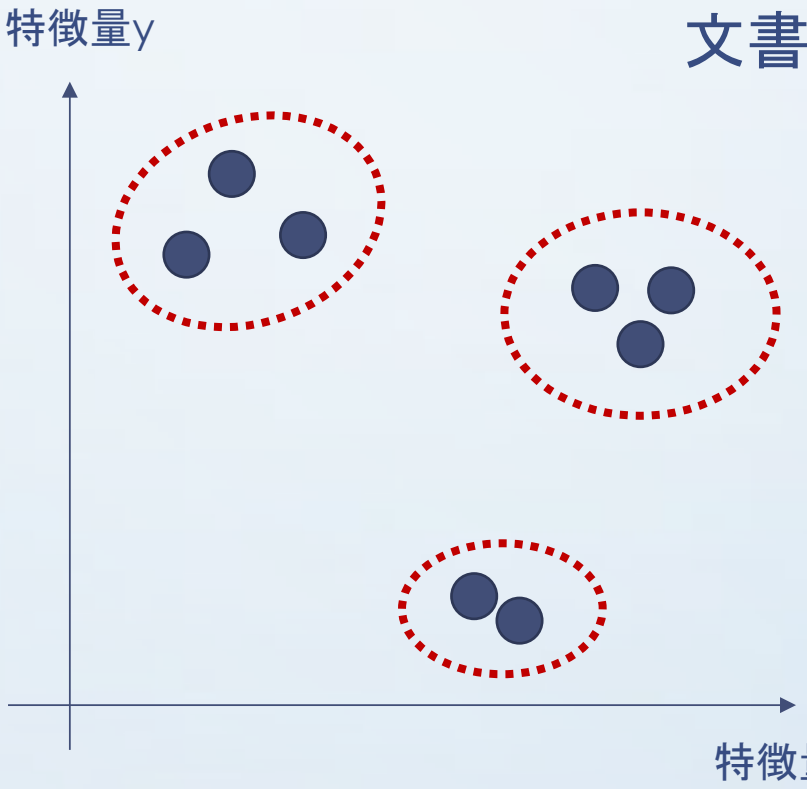
文書の類似度

- 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)

文書群を似ている部分集合に分ける



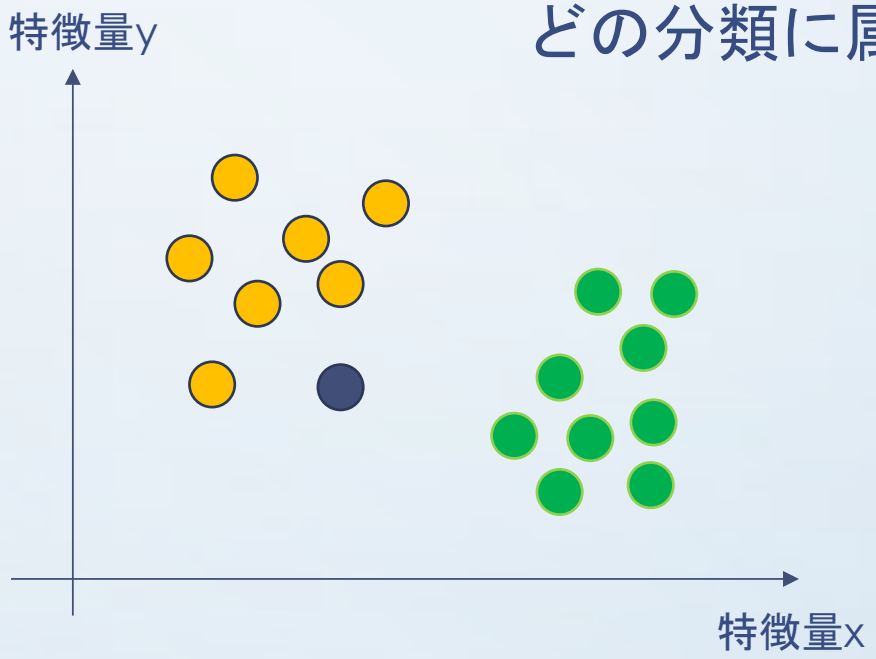
文書

クラスター

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

- Anconda Navigaterからjupyter notebookを起動
- Documents/exp4bd/3rd/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