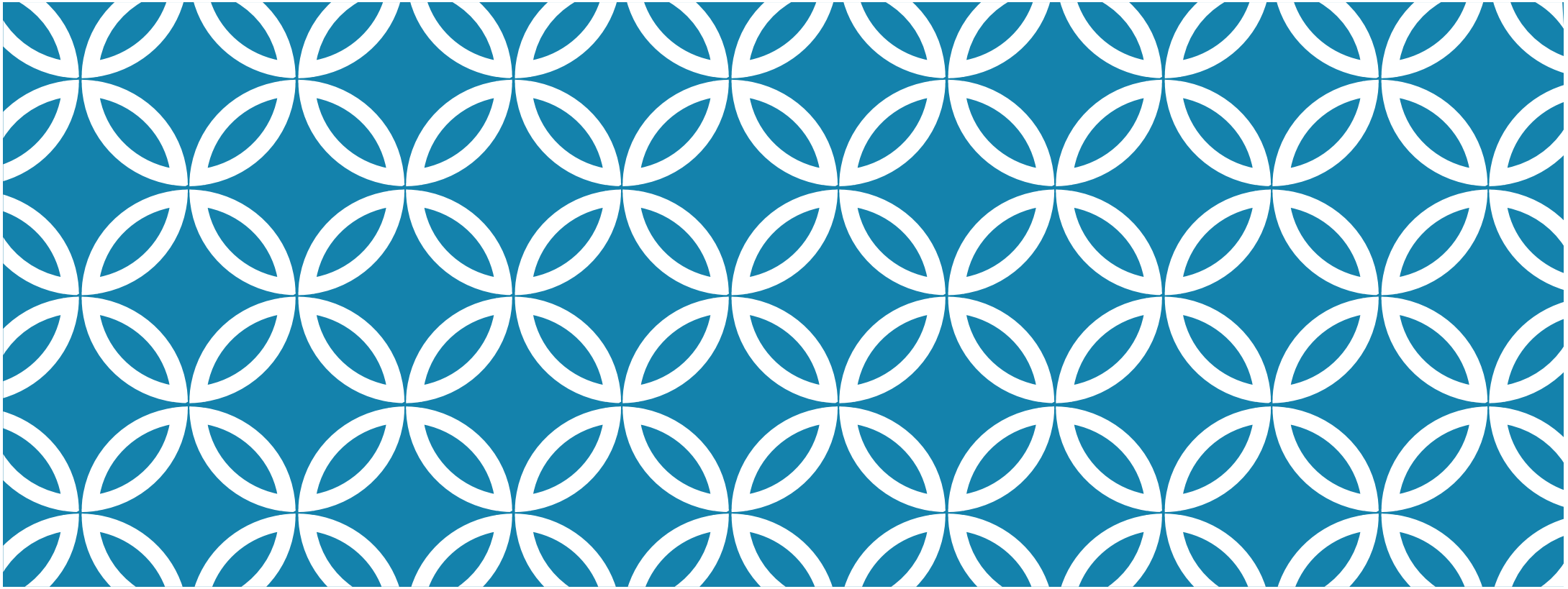


情報アクセス論 第12回 テキストマイニング(1)

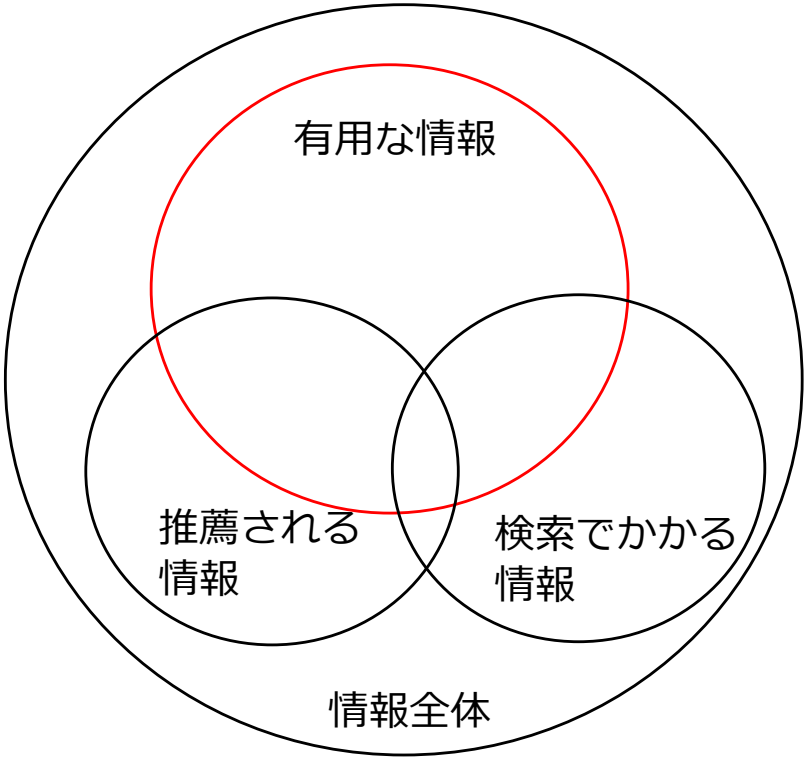
西原陽子

検索にかからない情報を獲得するにはどうすればよいか？

情報にアクセスする利用者は、多くの情報の中から有用な情報を得たいと思っている。

検索技術を用いると、問い合わせに合致する情報が獲得でき、
推薦技術を用いると、利用者の興味に合致する情報を獲得できる。

しかし、問い合わせや興味の外にあり、かつ有用な情報は検索や推薦では獲得が難しい



テキストマイニングとは

大量のテキストデータの中から、有用な情報を掘り起こす技術
- テキスト（text, 言語データ）とマイニング（mining, 掘り起こす）からなる造語

テキストデータとは文字のみで記されたデータ
- Webページのテキストや、図書のテキストなど

大量のデータの中から、有用な情報を掘り起こす技術である
データマイニングから派生してできた技術である。

テキストマイニングの利用例
- レビューの分析、意見の抽出、評価表現抽出、特徴的な単語の抽出
- 文書分類（文書を極性ごとに分類し、各極性での意見を抽出するなど）

情報アクセス技術としてのテキストマイニングの使い方

レストラン検索サイトで、新歓のお店を探しているとする。

- 場所とレストランの種類、予算などを入力し絞り込むが

お店の特徴がよく分からず、適切なお店を選びにくいことが多い。

お店の特徴を表す情報がWEBページや、レビュー、口コミなどに含まれている

テキストマイニング技術を用いると、例えば以下のようなことができる

- お店の特徴を表す単語を抽出する（「アットホーム」「ロマンチック」など）
- お店の評価に関わる単語のペアを抽出する（「夜景がきれい」「駅から近い」）
- レビューの極性（ポジネガ）を判定し、極性ごとに意見を抽出する（ポジティブなレビューでは「アットホーム」が多いが、ネガティブなレビューでは「狭い」が多いなど）

テキストからの単語の抽出

テキストは単語により構成されている。

テキストマイニングを行うためには、
テキストを構成する最小単位である単語の抽出が不可欠である。

単語は１つ以上の形態素により構成されている。

- 単語は１つ以上の形態素により構成されている。
- 形態素とは意味を持つ最小の単位：情報アクセス論（単語）、情報（形態素）

日本語のテキストから単語を抽出するには、形態素解析を行う

- 形態素解析とは、テキスト中の形態素とその品詞を特定する処理を指す。

形態素解析を行うには、形態素解析器を用いる

- 多くのフリーソフトがある：Chasen, MeCab, Igoなど

形態素解析器の実行

MeCabを用い、「さくらさくら」のテキストから形態素を抽出した例を示す。

１列目が入力したテキスト中の形態素（表記のまま）
２から7列目が品詞情報
８列目が形態素の原形、
９列目がふりがな、10列目が読み仮名

```
霞か雲かにほひぞいづる
霞　　名詞,一般,*,*,*,*,霞,カスミ,カスミ
か　　助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ
雲　　名詞,一般,*,*,*,*,雲,クモ,クモ
か　　助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
ほ　　動詞,自立,*,*,五段・ラ行,体言接続特殊２,ほる,ホ,ホ
ひぞ　動詞,自立,*,*,五段・ラ行,体言接続特殊２,ひぞる,ヒゾ,ヒゾ
いづる　動詞,自立,*,*,下二・ダ行,体言接続,いづ,イヅル,イズル
```


形態素解析後の単語整形処理

形態素解析器で抽出される形態素は辞書に登録されているものに限定され、
辞書にないものは正しく抽出がなされない。

```
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
ほ　　動詞,自立,*,*,五段・ラ行,体言接続特殊２,ほる,ホ,ホ
ひぞ　　動詞,自立,*,*,五段・ラ行,体言接続特殊２,ひぞる,ヒゾ,ヒゾ
```

正しく抽出するには、より多くの単語が登録された辞書を用いるや、
形態素をつなげて一つの単語とする方法がある。

- IPADICよりも、Neologdの方が登録単語数が多く、新しい単語が多い
- 名詞が連続して出現するならば、一つの単語を形成する可能性が高いとしてつなげてしまう（情報、アクセス、論 → 情報アクセス論）

単語の重要度の評価

テキスト中の全ての単語はそれぞれ重要度が異なっている

- 魚釣りの学生の例を考えると、魚の名前や釣りポイントの単語はそれ以外の単語より重要度が高い

単語の重要度を評価する指標として、以下の指標がある

- TF-IDF
- Okapi BM25

TF-IDFによる単語重要度評価

教科書第４章の4.3節を復習すること

文書dにおける、単語tの重要度は以下の式により評価される

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

TF(t,d): 文書dにおける単語tの頻度

IDF(t): 文書集合における単語tが含まれる文書の逆数

TF-IDFでは、ある文書dにおいて頻度が高く、他の文書には含まれない単語tが、その文書dを特徴付ける重要な単語と評価する

OKAPI BM25による単語重要度評価

一般的なTF-IDFでは文書内の総単語数を考慮していない

- 文書内の総単語数が多いものだとTF-IDFの重要度は適切なものとなるが、少ないものだと不適切なものになることがある

Okapi BM25を用いると文書内の総単語数を考慮した上で、単語の重要度を評価することができる

$$Okapi\ BM25(t,d) = \frac{TF(t,d) \times IDF(t) \times (K1+1)}{K1 \times \left(1-b+b \times \frac{WN(d)}{AWN}\right) + TF(t,d)}$$

分子のTF(t,d)×IDF(t)に、重みがかけられている

- WN(d)：文書に含まれる総単語数。大きいほど重要度が低くなる
- AWN：文書の平均単語数。小さいほど重要度が低くなる
- TF(t,d)：文書d内の単語tの頻度。大きいほど重要度が低くなる

K1とbは定数（1<K1<2, 0<b<1)

単語の共起関係の評価

テキストは単語がつなぎ合わされて作られており、
一定の範囲内にある単語は何かしらの関連があると考えられる。

- 上の例では「テキスト」と「単語」は関連があると言える

一定の範囲で２つの単語が共に出現する時、共起関係にあると呼ぶ

共起関係にある単語の組が分かると、単語だけよりも高度な情報を獲得できる

- （例）新しいゲームを表す単語について、共起する単語を調べることができれば、
  そのゲームがどのような評価を受けているかが分かる

共起関係を評価するときは、共起する範囲を設定する。

- 文、段落、テキストなどの単位が考えられる

共起頻度による共起関係の評価

一定範囲で共起する頻度が高ければ、その2つの単語は関連がある

共起頻度に基づく評価値$(t1, t2)$ = 単語$t1$と$t2$が共に含まれる頻度

- 範囲は文、段落、テキストなど、それぞれ変えることができる
- 単語重要度におけるTF(t)と類似する考え方に基づく

共起頻度はシンプルな評価値であるが、各単語の出現頻度に依存するため、適切な評価ができないことがある。

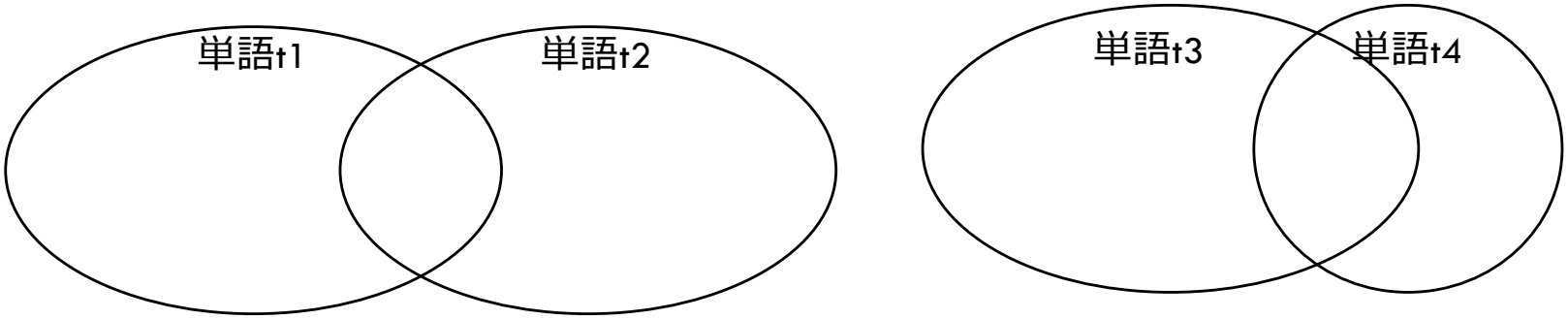
- 単語t1とt2の共起頻度が100、単語t3とt4の共起頻度が100 => どちらも同じ関連の強さ
- 単語t1とt2の出現頻度が10000、単語t3とt4の出現頻度が1000 => ？

相互情報量による共起関係の評価

各単語の出現頻度を考慮した共起頻度により、共起関係を評価する

$$\text{相互情報量}(t1, t2) = log\frac{P(t1,t2)}{P(t1)P(t2)}$$

- P(t1,t2): 単語t1とt2が共起する確率
- P(t1), P(t2): 単語t1とt2の出現確率



2つのベン図で共通部分の面積が同じだが、各単語の面積が異なる

- 単語t4は他の単語よりも面積が小さい＝そもそもの頻度が低い
- 単語t1とt2よりも、単語t3とt4の関係の方が強い。これを相互情報量は評価できる

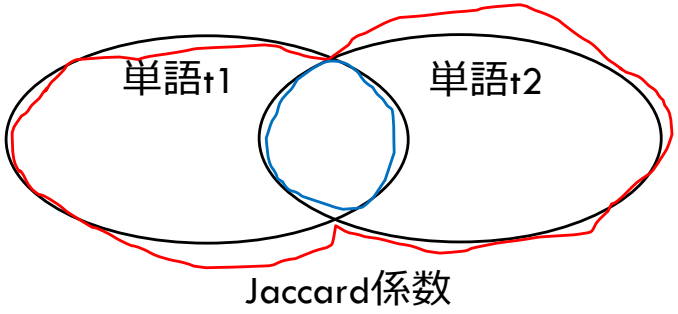
その他の方法による共起関係の評価

共起関係を基にしたその他の評価方法として、以下のものがある。

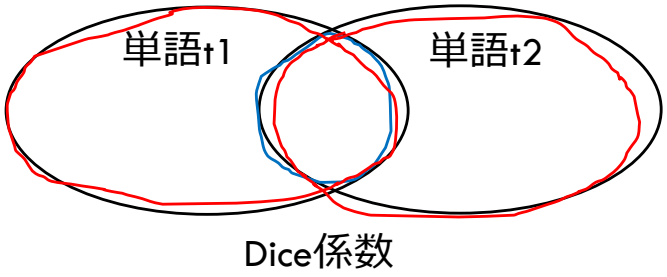
$$Jaccard(t1,t2) = \frac{\text{単語}t1\text{と}t2\text{を両方とも含む文書数}}{\text{単語}t1\text{または単語}t2\text{を含む文書数}}$$

$$Dice(t1,t2) = \frac{2\times\text{単語}t1\text{と}t2\text{を両方とも含む文書数}}{\text{単語}t1\text{を含む文書数}+\text{単語}t2\text{を含む文書数}}$$

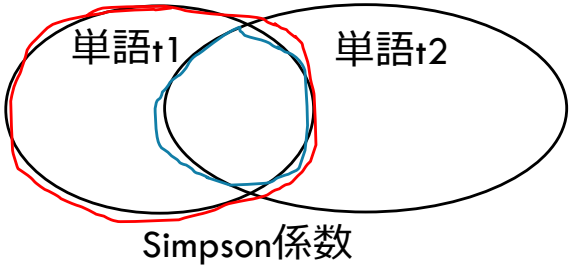
$$Simpson(t1,t2) = \frac{\text{単語}t1\text{と}t2\text{を両方とも含む文書数}}{\min(\text{単語}t1\text{を含む文書数},\text{単語}t2\text{を含む文書数})}$$



Jaccard係数



Dice係数



Simpson係数

テキストからの単語の役割の抽出

テキストは単語がつなぎあわされてできており、それぞれに役割がある

- 修飾語、被修飾語の関係
- （例）書籍のレビューデータから修飾語と被修飾語の共起関係が抽出されると、「個性的なキャラクター」や「独創的なストーリー展開」などと書籍の特徴が明らかになる

テキストから単語の役割を抽出する際には係り受け解析器を用いる

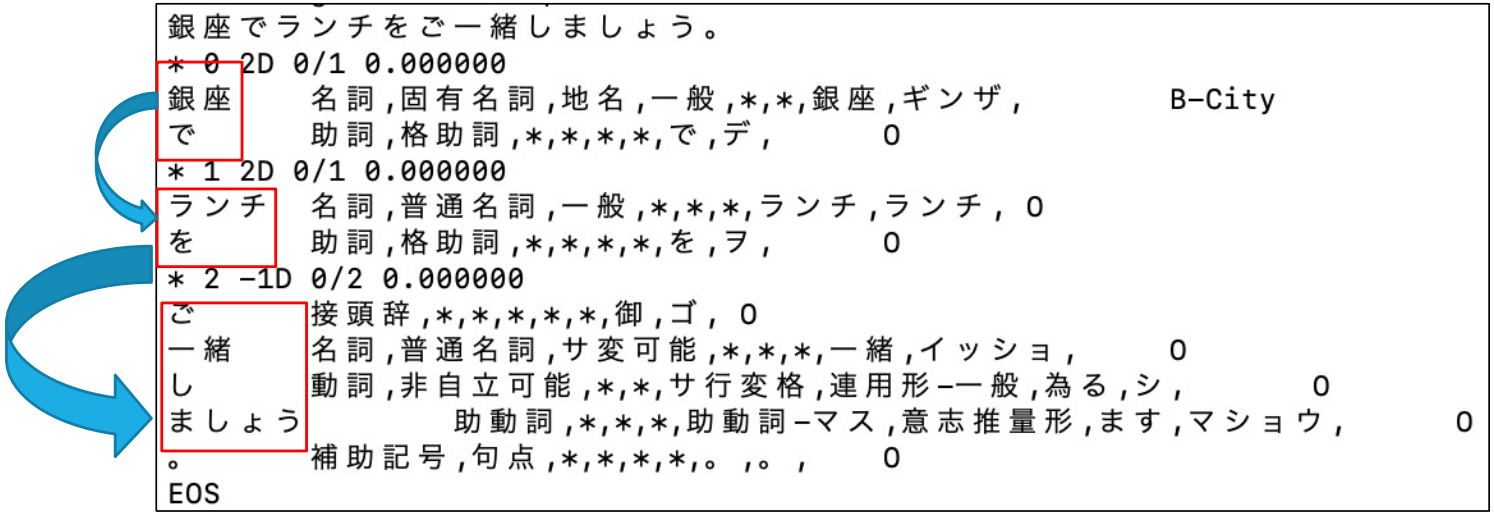
- ソフトウェアとしてはCaboChaやKNP、GiNZAなどがある
  - KNPはWEB上で試すことが可能（http://lotus.kuee.kyoto-u.ac.jp/nl-resource/cgi-bin/knp.cgi）

係り受け解析器の実行

GiNZAを用い、「銀座でランチをご一緒しましょう」を係り受け解析した例

2列目の数字Dの部分が係り受けを示す。

- 2D「銀座で」は2D「ランチを」に係る
- 2D「銀座でランチを」は1D「ご一緒しましょう」に係る



単語や文章の極性の評価

極性とは、ある単語や文章が読み手に与える印象を表すもの。

大きくはポジティブ、ネガティブ、ニュートラルの3つに分けられる。

- 「今日は良い天気」：ポジティブな印象を与える文
- 「今日は悪い天気」：ネガティブな印象を与える文

単語や文章の極性を評価することにより、ある対象に対する全体的な評価を知ることや、詳細な意見を知ることが可能となる。

- ある商品に対しての全体的な評価（レビューサイトでは星による評価と合わせて利用可能）
- ある商品に対しての詳細な評価（機能Aは良いが、機能Bは悪いなど）

極性の評価には、極性が示された単語の辞書を用いることが多い

- 日本語評価極性辞書（https://is.gd/5fznKs）
- 単語感情極性対応表（https://is.gd/l0T4v3）

極性が示された辞書の例

1列目が単語、2列目が読み、3列目が品詞、4列目が極性を示す数値

- 数値がプラスであればよりポジティブ、数値がマイナスであればよりネガティブ

極性値が高い単語の例

優れる:すぐれる:動詞:1
良い:よい:形容詞:0.999995
喜ぶ:よろこぶ:動詞:0.999979
褒める:ほめる:動詞:0.999979
めでたい:めでたい:形容詞:0.999645

極性値が低い単語の例

ない:ない:助動詞:-0.999997
酷い:ひどい:形容詞:-0.999997
病気:びょうき:名詞:-0.999998
死ぬ:しぬ:動詞:-0.999999
悪い:わるい:形容詞:-1

文章の極性判定の例

文章中に含まれる単語を抽出し、単語の極性値の合計や平均により判定することができる

- 「今日は良い天気」：今日(0.24)+良い(0.99)+天気(0.24) = 1.47 → 平均は0.49
- 「今日は悪い天気」：今日(0.24)+悪い(-1)+天気(0.24) = -0.52 → 平均は-0.17

ただし、複雑な文章になると、極性値の合計や平均だけでは極性の評価が困難となる。

- 「今日は**良い**天気でも**ない**ことも**ない**」：否定が複数回含まれる文
- 「**適当**に勉強をする」と「分類は**適当**であった」：文脈により極性が異なる場合
- 「料理は**良かった**が、雰囲気は**悪かった**」：1文で複数のことを言及する場合

上記の問題に対応するには、文脈を考慮した上での判定が必要となる

- 係り受けを考慮する、時系列を評価するなど

まとめ

テキストマイニングで使われる技術を紹介した

単語の抽出、重要度の評価

単語の共起関係の評価

単語の役割・極性の評価

