知能機械と自然言語処理
第４回：
自然言語処理の応用４：
文の意味の解析システム
機械翻訳システム

David Ramamonjisoa

目次

- 自然言語理解
- 質問応答システム
- F尺度
- 詳細流れ

自然言語の理解

人工知能の完全版、完璧版

機械が翻訳できる

機械が質問に対して正しい回答

どのような入力にしても音声、文書、画像、映像など

絵でわかる人工知能（三宅）

文の理解の問題

- どのようにして自然言語の意味を表現し、コンピュータが処理できるようにするか。
- 制約のない文に対してどのように意味表現を関連付けるか。
- 文の意味表現を知識源に結びつけるプログラムをどのようにして利用するか。

文書分類（Document Classification）

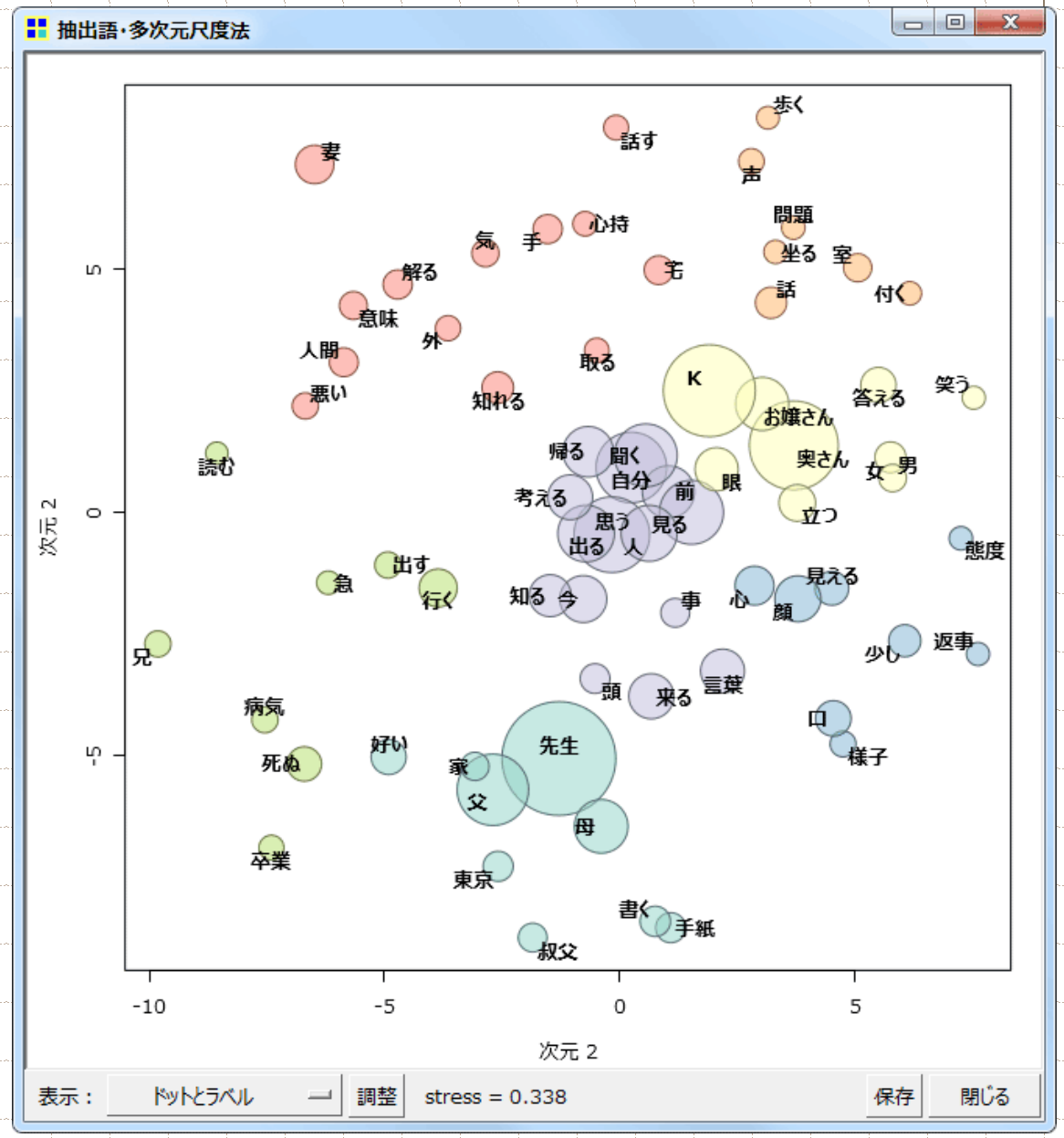
- 問題：電子文書をその内容に基づいて、1つ以上に分類する
- 解決方法：
  - 教師あり文書分類：
    - ベクトル空間モデル(tf-idf), 単純ベイズ分類器,潜在意味解析, 機械学習（サーポートベクトルマシンSVM、決定木、k近傍法）, word2vec, doc2vecなど
  - 教師なし文書分類：
    - 階層的クラスター分析, 文書のクラスター分析
- 例：メールのスパムかハム(spam or ham)の分類
- 評価：分類精度

テキスト分析

- 多次元尺度構成法（MDS）
- 対応分析
- 共起ネットワーク
- 文書中でコードが多く出現している箇所
- ワードクラウド

多次元尺度構成法(MDS)

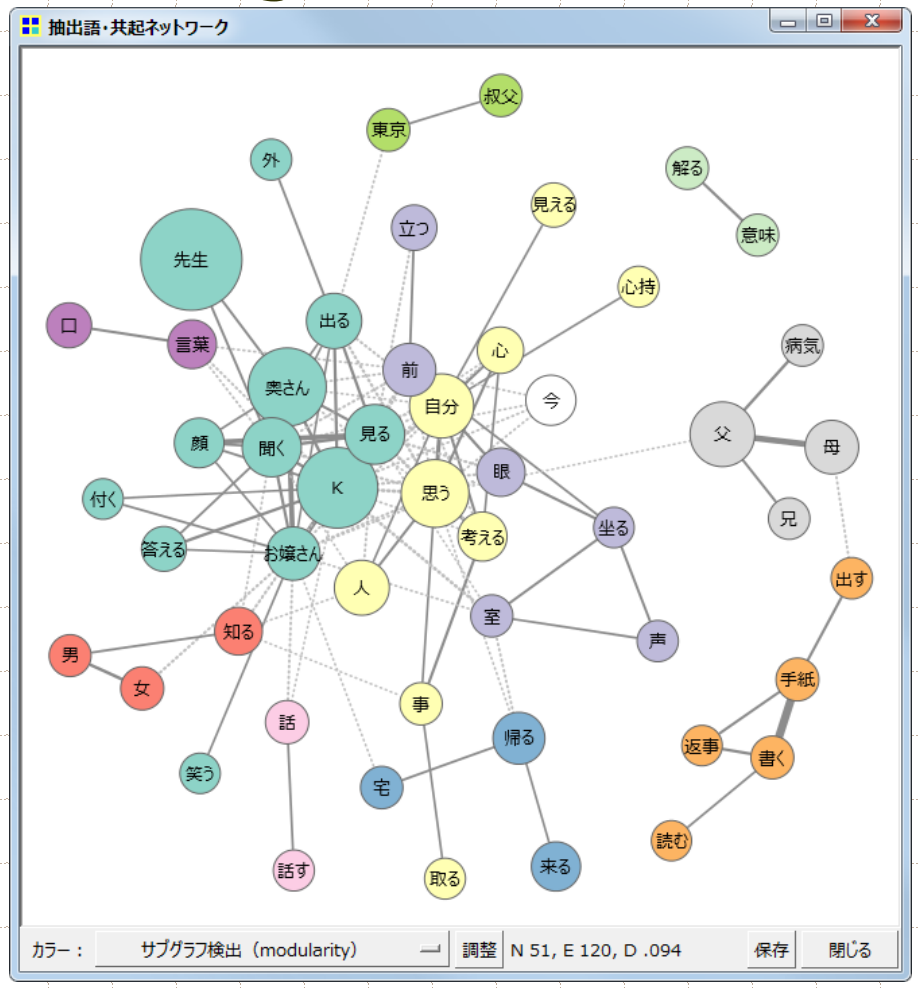
夏目漱石の小説:
こころの単語とMDS



KHCODER

共起ネットワーク

- 出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く



ワードクラウド

- ワードクラウドとは文章などの文字データからよく使われる言葉を抽出し、出現回数に応じて文字の大きさや文字を色分けして、どんな言葉が多く使われているかを見える化したものです。

質問応答システム

- データベースに対する問い合わせ
  a.Which country is Athens in?
  b.Greece.

このようなシステムはどれくらい難しいのだろうか。

- 知りたい内容を質問文でそのまま入力する
- 大量な文書リスト中で回答を抜き出す必要はない。

SQLを用いて以下のデータベースを問い合わせる→質問を答える

city_table: A table of cities, countries and populations

City	Country	Population
athens	greece	1368
bangkok	thailand	1178
barcelona	spain	1280
berlin	east_germany	3481
birmingham	united_kingdom	1112

自然質問文からSQLクエリへ

- データベースに対する問い合わせ

  a.Which country is Athens in?

- SQLの質問は以下に示す

```
SELECT Country FROM city_table WHERE
  City = 'athens'
```

→英語からSQLへの翻訳が容易になる

質問文の意味表現はどのようにするか

- 構文解析の結果は文法を生成される。
- あらかじめ文法を作成してから質問文を解析する。
- 各句構造規則には、要素SEMに対する値を組み立つ。
- 文字列連結計算+を用いて、子の構成素に対応する値をつなぎ合わせることで親の構成素の値を作っている。

```
>>>nltk.data.show_cfg('grammars/book_grammars/sql0.fcfg')
 % start S
S[SEM=(?np + WHERE + ?vp)] -> NP[SEM=?np] VP[SEM=?vp]
    VP[SEM=(?v + ?pp)] -> IV[SEM=?v] PP[SEM=?pp]
    VP[SEM=(?v + ?ap)] -> IV[SEM=?v] AP[SEM=?ap]
    NP[SEM=(?det + ?n)] -> Det[SEM=?det] N[SEM=?n]
    PP[SEM=(?p + ?np)] -> P[SEM=?p] NP[SEM=?np]
    AP[SEM=?pp] -> A[SEM=?a] PP[SEM=?pp]
    NP[SEM='Country="greece"'] -> 'Greece'
    NP[SEM='Country="china"'] -> 'China'
Det[SEM='SELECT'] -> 'Which' | 'What'
N[SEM='City FROM city_table'] -> 'cities'
IV[SEM=''] -> 'are'
A[SEM=''] -> 'located'
P[SEM=''] -> 'in'
```

- これにより、クエリを構文解析してSQLに変換できる。
- 質問：'What cities are located in China?'

```
>>> from nltk import load_parser
>>> cp = load_parser('grammars/book_grammars/sql0.fcfg')
>>> query = 'What cities are located in China'
>>> trees = cp.nbest_parse(query.split())
>>> answer = trees[0].node['SEM']
>>> q = ' '.join(answer)
>>> print q
```

SELECT City FROM city_table WHERE Country="china"

- 最後にデータベースcity.dbに対してクエリを実行し、結果を得る。

```
>>> from nltk.sem import chat80
>>> rows = chat80.sql_query('corpora/city_database/city.db', q)
>>> for r in rows: print r[0],
canton chungking dairen harbin kowloon mukden peking shanghai sian tientsin
```

自然言語、意味論、論理

- 翻訳の別の例：

a. What cities are in China and have populations above 1,000,000?

b. SELECT City FROM city_table WHERE Country = 'china' AND Population > 1000

- AND＝andの意味が本当正しいのか。

ホランダ語：Margrietje houdt van Brunoke.

英語：Margrietje loves Brunoke

翻訳は意味があるのか。

自然言語、意味論、論理

- 意味論における2つの基本的な概念を導入している。
- 1つ目は、表明文が「ある状況において真もしくは偽」というものである。
- 2つ目は、限定された名詞句や固有名詞は「世界における実体に対応する」
- 文の集合を考えた際には、真偽値の概念を採用し、状況において真になるかどうかを問うことだできる。推論も可能になる。

意味論、論理

- 文は整合である

a. Sylvania is to the north of Freedonia.

b. Freedonia is a republic.

- 文は不整合である

a. The capital of Freedonia has a population of 9,000.

b. No city in Freedonia has a population of 9,000.

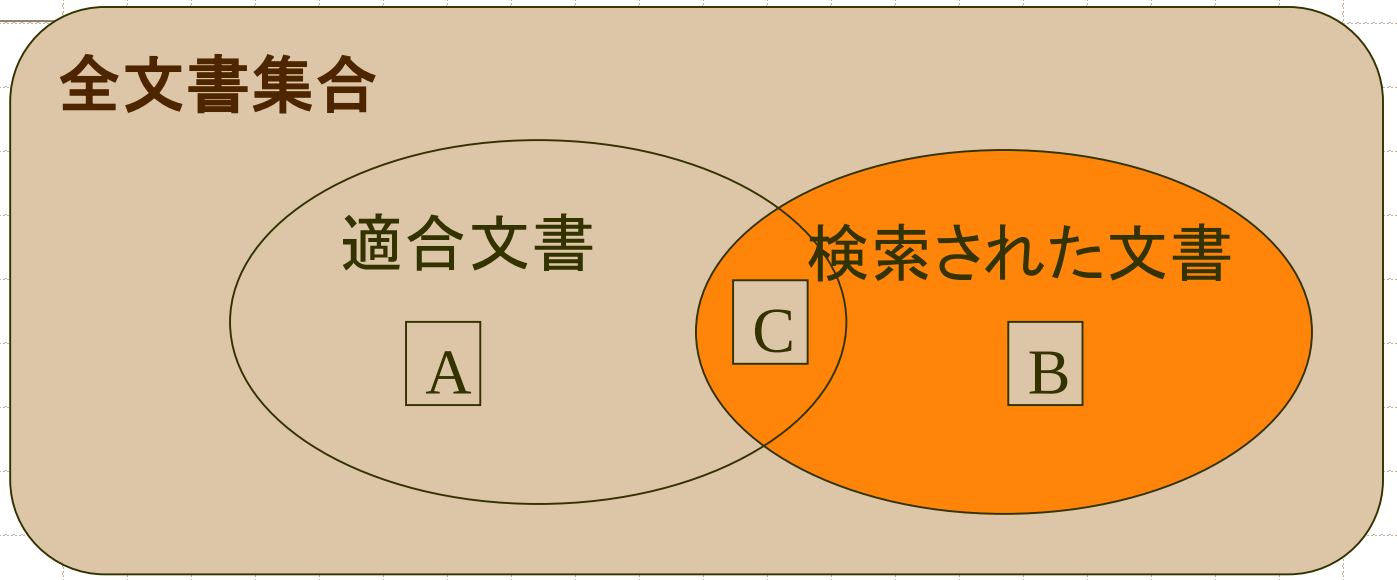
a. Sylvania is to the north of Freedonia.

b. Freedonia is to the north of Sylvania.

情報検索システムの評価尺度

- 情報検索システムの有効性(effectiveness)
  - 適合性(relevance)
  - 適切性(pertinence)
  - 有用性(usefullness)
- 再現率と適合率
  - 完全性
  - 正確性
  - 再現率(recall):完全性を評価するための尺度
  - 適合率(precision):正確性を評価するための尺度

情報検索システムの評価尺度



$$\text{再現率} = \frac{\text{検索された文書中の適合文書の数}}{\text{全文書中の適合文書の数}} = \frac{|C|}{|A|}$$

$$\text{適合率} = \frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}} = \frac{|C|}{|B|}$$

F尺度、平均適合率

- F尺度

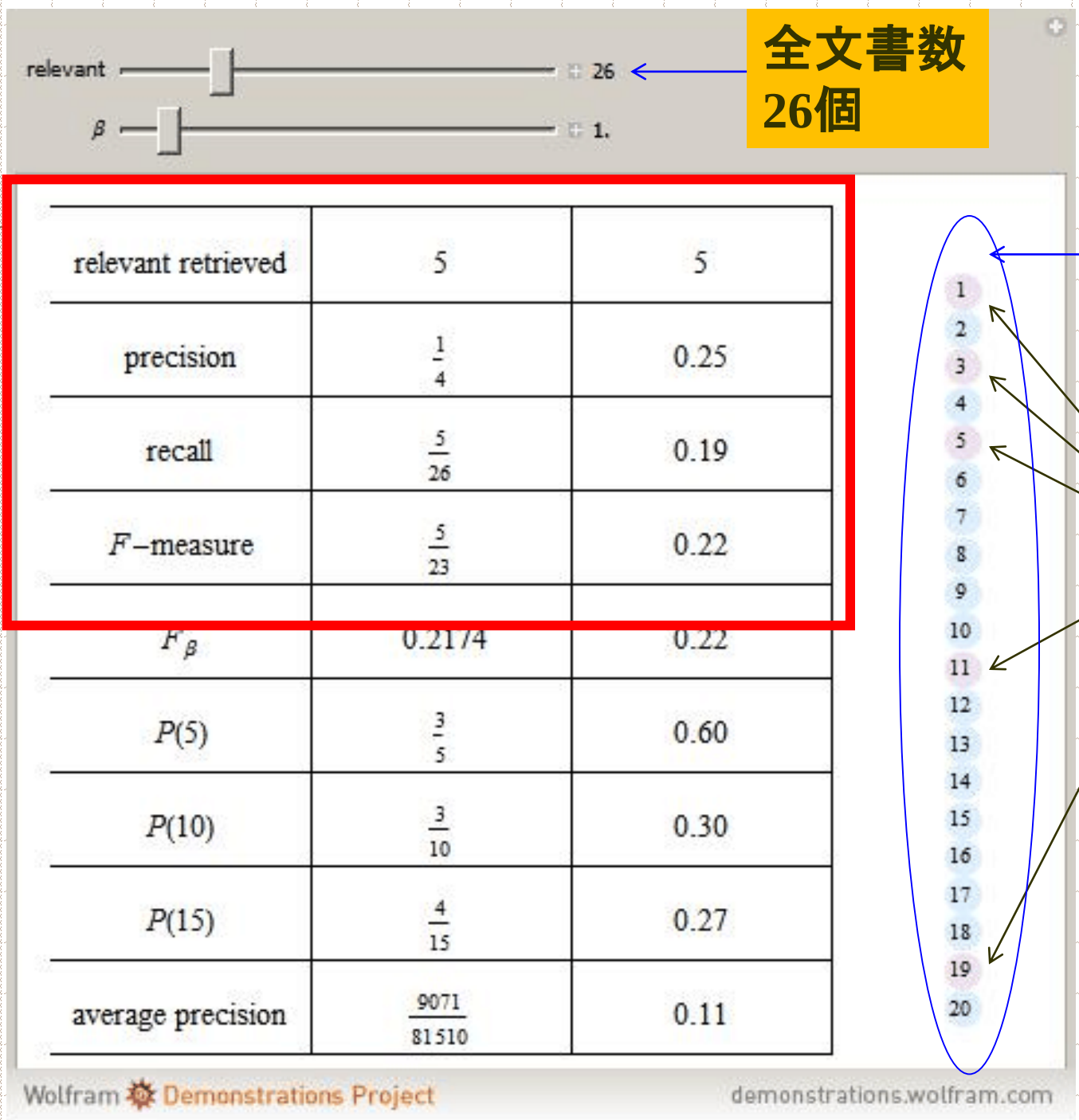
$$F = \frac{2}{(1/R + 1/P)}$$

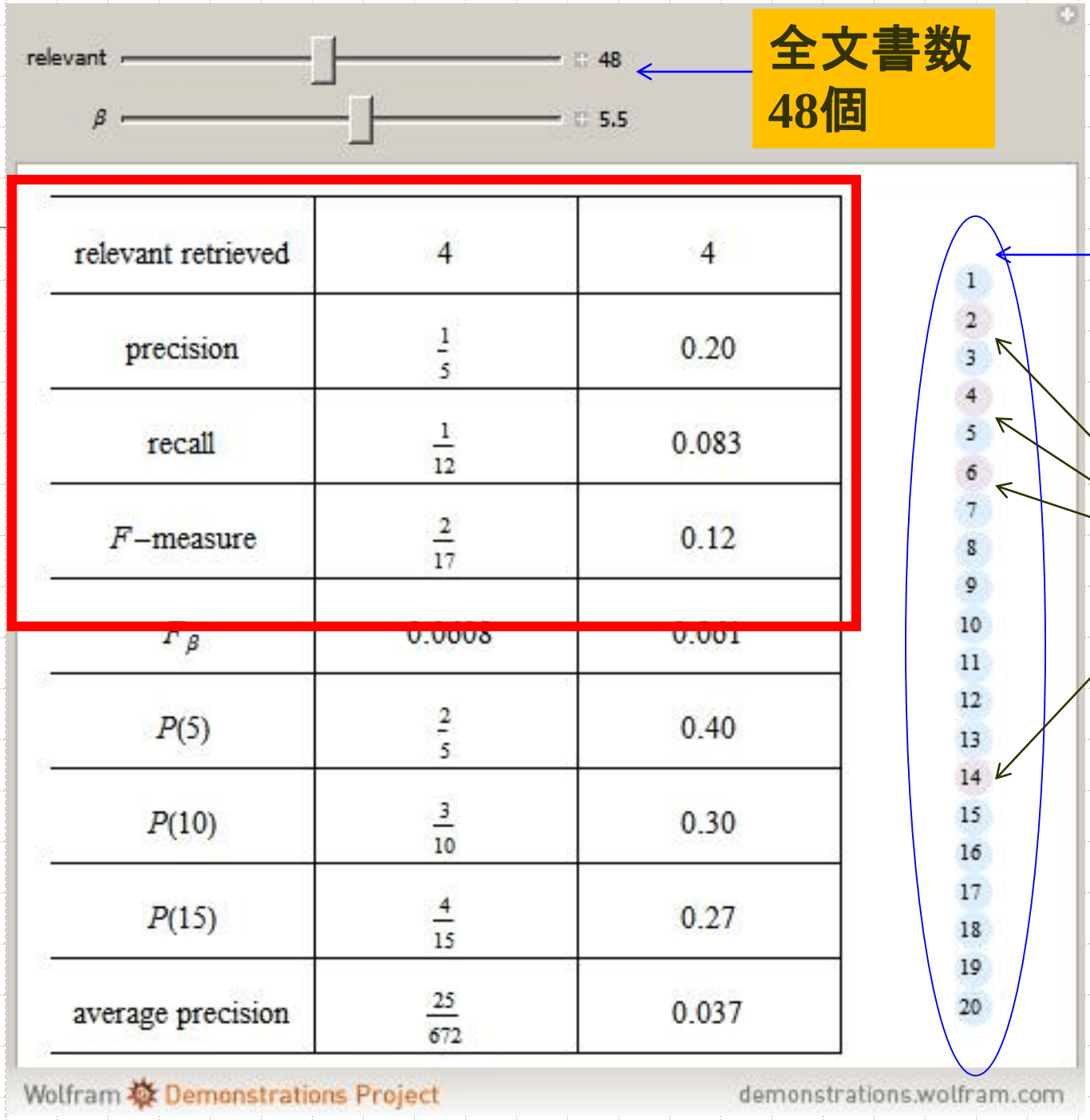
R: 再現率、P:適合率

- 平均適合率

$$AP = \frac{\Sigma \; P(x)}{R}$$

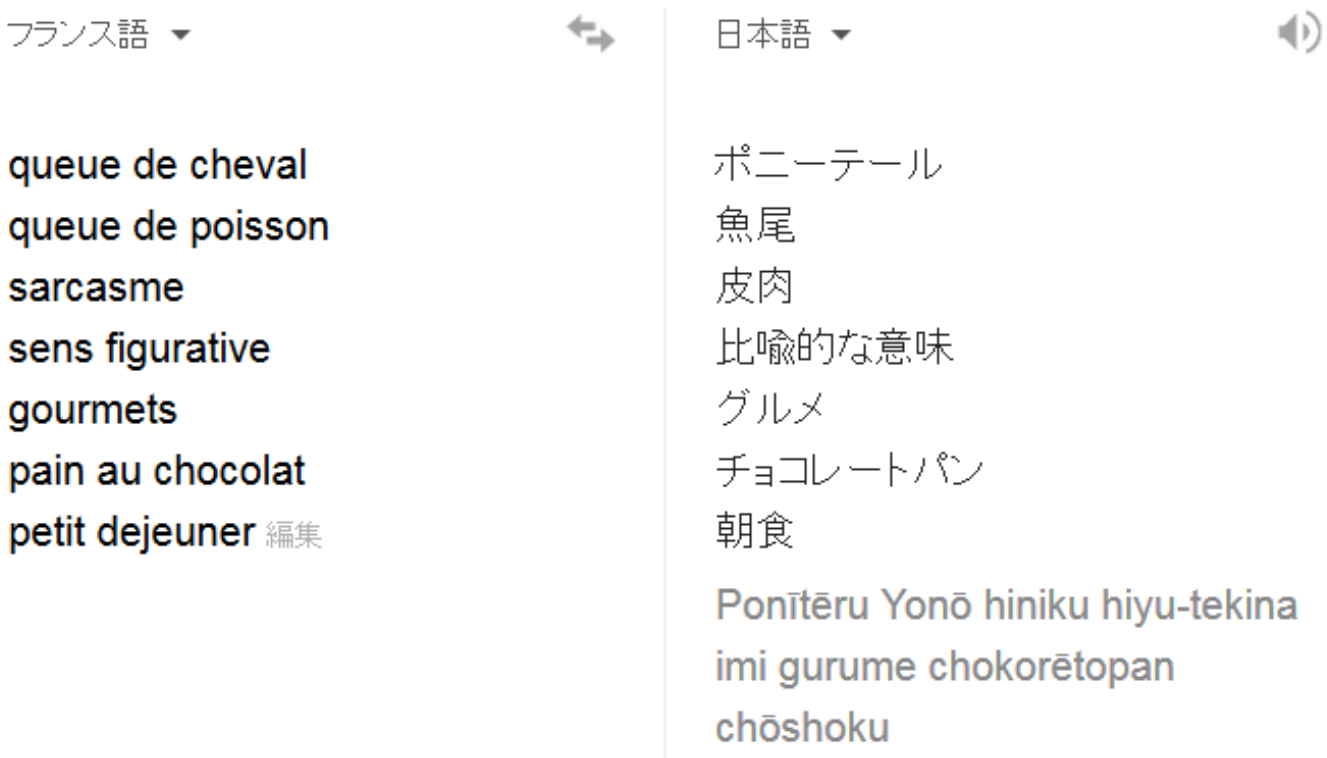
$$\text{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{\text{number of relevant documents}}$$


relevant
26
β
1.
全文書数
26個
relevant retrieved 5 5
precision 1/4 0.25
recall 5/26 0.19
F−measure 5/23 0.22
F_β 0.2174 0.22
P(5) 3/5 0.60
P(10) 3/10 0.30
P(15) 4/15 0.27
average precision 9071/81510 0.11
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
検索された文書数
20個
適合
文書
数5個
Wolfram Demonstrations Project
demonstrations.wolfram.com


relevant
48
β
5.5
全文書数
48個
relevant retrieved
4
4
precision
1/5
0.20
recall
1/12
0.083
F−measure
2/17
0.12
F β
0.0608
0.061
P(5)
2/5
0.40
P(10)
3/10
0.30
P(15)
4/15
0.27
average precision
25/672
0.037
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
検索された文書数
20個
適合
文書
数4個
Wolfram Demonstrations Project
demonstrations.wolfram.com

機械翻訳システム

- ある言語の文を別の言語に自動的に翻訳すること
- 原言語 (source language)→目標言語 (target language)



機械翻訳の問題点

2言語間のずれ

イディオム

非単調性

文脈依存性

中間言語方式

- 中間言語(interlingua)
  - 文の内容の表現形式
  - すべての言語について共通
- 処理の流れ
- 特徴
  - 多言語翻訳への対応が容易
  - 中間言語の設計が非常に難しい

その他の応用

- 情報抽出の自動学習（巨大知識ベースの構築）
- ソシアルネットワークの分析
- テキストマイニング
- 商品レビューのオピニオンマイニング
- トピックモデル
- 対話システムの構築

質問応答システム

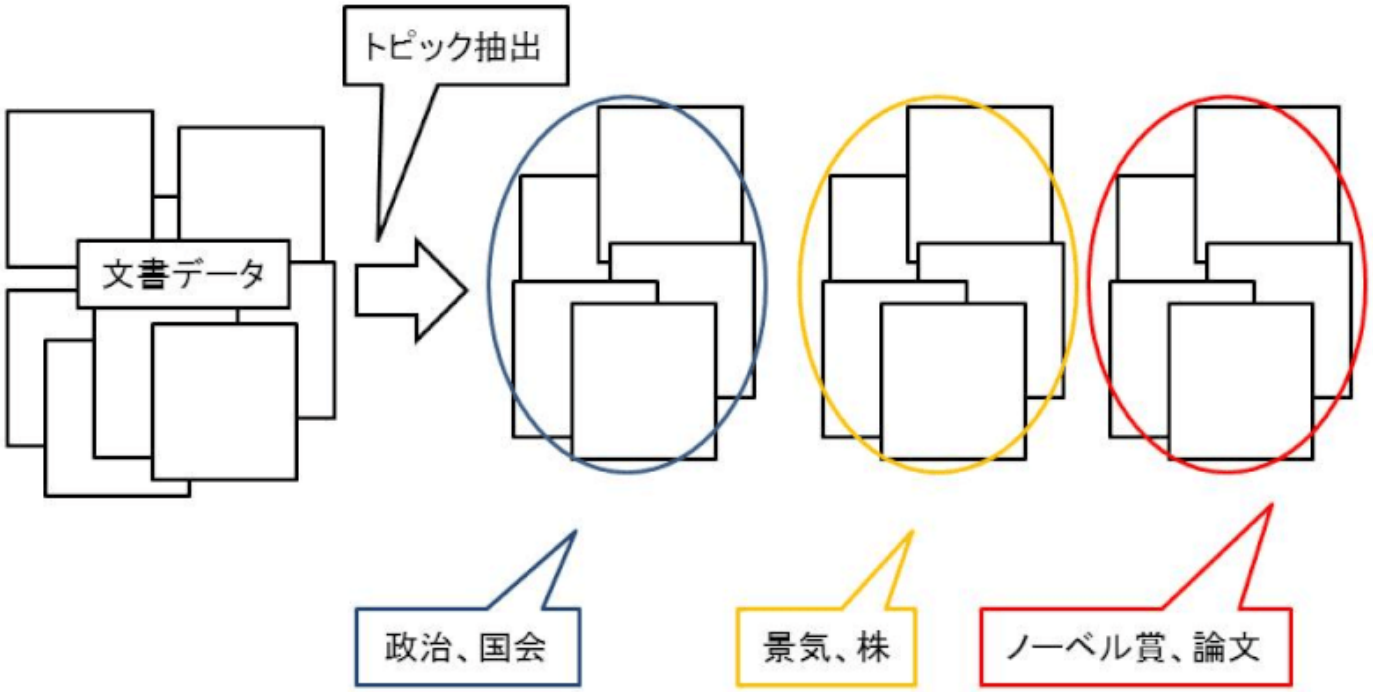
トピックモデルとは？

- 言語学における話題（わだい：Topic）は、主題（しゅだい、ドイツ語: Thema、英語: theme）、題目（だいもく）などともいい、文によって陳述される中心的対象をいう。

- 文中に明示された場合には話題語（主題語）ともいう。

トピック分析

- トピックモデル：文章からそれが何かを説明するためのモデル



引用：http://qiita.com/GushiSnow/items/8156d440540b0a11dfe6

トピック分析の3つの手法

- 潜在的意味解析（LSI (Latent Semantic Index)）
- 文章中の潜在的なトピックを推定（LDA Latent Dirichlet Allocation）

文章分類や、文章ベクトルの次元削減等に用いられる技術

まとめ

- 情報検索と質問応答システムは自然言語処理の応用である。
- 評価するための尺度
  - 再現率、適合率、F尺度、平均適合率などある
- テストコレクション
- NTCIR, TREC コンテストのワークショップ

参考文献

- 「自然言語処理ー基礎と応用ー」田中穂積, 電子情報通信学会編, 1999.