KH　Coderを利用したテキストマイニング

宮城教育大学教職大学院

平　真木夫

概要

- テキストマイニングとは？
- KH　Coderにおけるテキストデータ（HTML）
- 自然言語処理
- 共起関係とは？
- 対応分析
- 共起ネットワークからクラスター分析

自然言語処理の分類

- 形態素解析
  - 自然語の文を単語や複合語に分割し、各単語の品詞を求める → 茶筅、**TermExtract**など
- 構文解析
  - 文を文節に分割し、文節間の依存関係を求める

- 意味文脈解析
  - 文章における文章間の意味関係を求める
- 応用処理 （**テキストマイニング**）
  - 文章クラスタリング・分類、情報抽出など

テキストマイニングの下ごしらえ

- 類義語辞書の作成（置換 or コーディング・ルールの活用※チュートリアル参照）　Synonym-Subjects.txt参照
  - 予習復習、予習して復習する、復習・予習…→予習・復習
  - お母さん、お父さん、家の人、兄、姉…→家族
  - 友達、ともだち、友だち→友達
- **複合語をチェック**しながら分析する
- 疑似HTML化：テキストの**階層構造**
  - 分析の区切りを示すために、HTMLのタグを利用
  - 大きなまとまり：＜h1＞・・・＜/h1＞
  - 中間的なまとまり：＜h2＞・・・＜/h2＞
  - ミニマムなまとまり：＜h3＞・・・＜/h3＞

複雑に考えずに、H1だけで十分　段落でも可

参照：Rev-FiveSubjects.txt 数学.txt

テキストの説明（国語と理科が混在）

<H1>国語</H1>

←科目のレベル　H1

<H2>国語 2222</H2>

←各回答者のレベル H2

小学校から中学校へ進学したとき、国語の勉強方法や教え方は変わると思います。…… 中学生時代の勉強はその人の将来を左右すると言えます。

<H2>国語 3333</H2>

小学校と中学校の勉強方法、教え方は大きく異なると考える。中学校においては、学習指導要領に沿って、……授業の確認、漢字、熟語の練習をしっかりさせたい。

<H2>国語 000000</H2>

私は、小学生の時は、国語の基礎としてまずは平仮名、片仮名、漢字や、文のつくり方、作文の基礎などを学び……中学校のときは、ノートを見直し、学校のワークを解いただけで、主だった勉強はせず暗記科目のようになっていた気がする。

<H1>理科</H1>

<H2>理科 555555 </H2>

私は小学校から中学校へと進学する際に私が担当している理科は…

他にも、<H1> 一学期</H1> 、<H1> 二学期</H1>
<H1> 男子</H1>、<H1> 女子</H1>といった設定も可能

KH Coderの**基本的**な使い方

1. プロジェクト→データとなる文書の登録
2. 前処理：ファイルのチェック，前処理の実行
   →**複合語**の検出（TermExtractがお勧め）
   1. 複合語の状況をチェックしながら**同義語**を判断
3. ツール→抽出語　：複合語を含んだ抽出語を元に文章の特徴を判断する
   1. H1のレベルで分析するのか？段落のレベルで分析するのか？

対応分析

- 類似の手法：双対尺度法、コレスポンデンス分析、数量化理論Ⅲ類、クロス分析など
- 解釈方法
  - **H1のレベル**で**バブルプロット**として表示するように指定すると・・・円の大きさが生起頻度の大きさを意味している。
  - **軸に意味づけ**をした方が、結果の解釈がしやすい。
  - 異なる項目のカテゴリの位置関係は、**原点からの方向**で判断する。原点から見て同じ方向にあれば、一見して距離があっても、同様の意味づけが可能である。
  - 関連の強い項目どうしは、**原点からみて同一方向に布置**される。
  - 関連の強いカテゴリは近くに、弱いカテゴリは遠くにプロットされるが、これはあくまでカテゴリ間の相対的な関係で、絶対的なボリュームを表わすものではない。
  - 布置された位置の距離によって近さや遠さを知ることはできても、回答した人数は分からない→**サンプルサイズは結果に反映されない**。
  - ２軸の寄与率（その軸だけで元のデータの何割を説明することができているかを表した数字）を合計した「**累積寄与率**」が80％以上であれば、元データをかなり反映している。

共起ネットワーク

- **共起関係**とは：　1つのテキストの中に、複数の語がそれぞれ共に頻出している状態
  - （例）「勉強方法」「丸暗記」が**小中の勉強の違いに関する質問**の中で共に出現
- テキストの中で用いられた**単語をノード**とし、単語と単語の**共起性をリンク**とするネットワーク。そして、リンクの強さを Jaccard 係数で表している（**強さ≒太さ**）。
- お勧めのオプション
  - 1.強い関係ほど太字、2.出現頻度が多いほど大きい円、3.ラベルが重ならない、4.**サブグラフ検出**、5.保存はPDFかPNG

Jaccard係数

- 集合*X*と*Y*の共通要素を要素の総数で割って集合間の類似度を比較する。
- ベクトル*X*(X1,X2,,,Xn)，ベクトル*Y* (Y1,Y2,,,Yn)と定義し，X∪Yの要素をZ1,Z2,,,Znとするとと Jaccard係数は以下のように定められる。

$$\text{Jaccard 係数} = \frac{|x \cap y|}{|x \cup y|} = \frac{x \cdot y}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - x \cdot y}$$

たとえば，X={1,4,7}　Y={1,3,4,8,9}　とすると，
J(X,Y)＝ |X∩Y| ÷ |X∪Y| =|{1,4}|÷|1,3,4,7,8,9}|=2÷6 = 0.33…
Jaccard距離＝1－(Jaccard係数)＝(|X∪Y| －|X∩Y|)÷ |X∪Y| ←**非類似度**
他にもユークリッド距離，コサイン係数も選択できるようになっているが，Jaccard係数が無難。

階層クラスター分析

- 階層クラスター分析を行うとデンドログラム（樹形図）が出力される
  - 全対象の類似度を計算し，最も類似の高いものから順次グルーピングする。最終的に1つのクラスターになるまで繰り返す。
- デンドログラムでは、図の**左の方で結合**すればするほど**近い関係**にあるといえる（グラフの下の数値は結合距離）。クラスタリングは**Ward法**が最も一般的。（Jaccard距離が小さいものをグルーピングする）
- 点線の部分の高さまでで結合されているクラスターで色分けされている
- どのようにクラスターが結合されていくかの過程も確認可能
- クラスター分析の結果、どれくらいの数にグループ分けするべきかについては、**完全に分析者に委ねられている**。分析したら終わり、ではない。→ **共起ネットワークとあわせて実施**

参考資料

- 「社会調査のための計量テキスト分析―内容分析の継承と発展を目指して」 樋口耕一 ナカニシヤ出版
- KH Coder作者によるチュートリアル http://www.slideshare.net/khcoder/kh-coder-28776074
- 簡単だけれでもとっても重要な統計学の話 http://staff.miyakyo-u.ac.jp/~m-taira/Lecture/simple-but-important.html