プログラミング応用
http://bit.ly/kosen02

Week15@後期(week30 in 2016)
2016/1/26

本日の内容

- 講義(自然言語処理)
  - 自然言語処理の概要(第２回)
    - 文書間の類似度と応用例
    - 文書間の類似度を測る指標
- 演習
  - Jaccard係数を計算するシェルスクリプト

残りは試験に向けた質問の時間にします。
終わった人から帰って良いです。

質問

- 文書D_1とより似ているのは文書D_2, 文書D_3どちら？

D_1

虫歯の原因や治療法、予防法、子供の虫歯、治療費などについて分かりやすく解説します。

D_2

子供の虫歯は早期発見が重要で、最終的な治療費も安くなります。

D_3

うちの子供はいちごパフェが好きでよく食べます。

文書間の類似度

- 文書
  - １文以上の文がまとまったもの
    例) 新聞記事、Twitterへの投稿
- 文書間の類似度
  - 文書同士が「どれだけ似ているか?」定量化した指標

    → 類似度を定量化できればどのようなソフトウェアに応用可能?

応用例) 記事の自動分類

ニュース配信サービス等では、”似た記事”を1つのページにまとめて表示

応用例) スパムメール判定

多くのメールサービスでは、ユーザに迷惑メールを
報告してもらい迷惑メールの事例を収集。
収集した迷惑メールと”似ている”メールを
迷惑メールフォルダに振り分ける。

例) 類似レシピ検索

レシピやその他の多くの検索サービスでは、現在閲覧中のページと関連する(=“似ている”)ページを表示。

本日の内容

- 講義(自然言語処理)
  - 自然言語処理の概要(第２回)
    - 文書間の類似度と応用例
    - 文書間の類似度を測る指標

文書間の類似度を図る指標

1. 集合演算を用いる指標: Jaccard係数
   - 文書を(日本語の場合)形態素の集合として表現。
   - 2つの文書間でどの程度の割合の形態素を共有しているかを基準に類似度を測る
2. ベクトル演算を用いる指標: コサイン類似度
   - 文書をベクトルとして表現。2つのベクトルのコサインの値を類似度として用いる。
     (本日は詳しく扱わない。)

Jaccard係数による類似度計算

1. 文書を集合として表現
2. Jaccard係数を計算

1. 文書を集合として表現

D_1

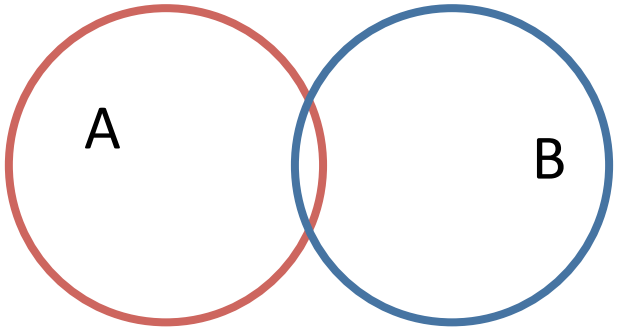
集合で表現

D_1={吾輩, 猫, 名前,}

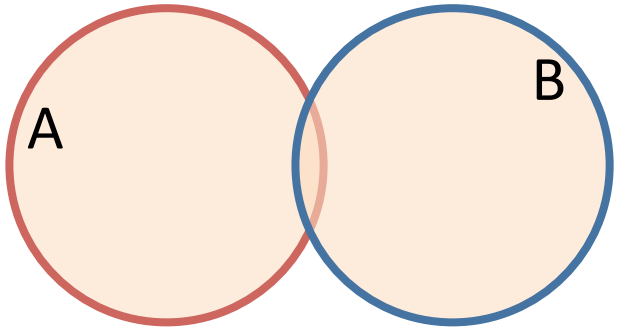
※すべての形態素を使わず
名詞や形容詞のみを使うのが一般的

復習) 集合演算

- 積集合(A∩B): AとBの重なる部分



- 和集合(A∪B)



2. Jaccard係数を計算

- 集合A, Bに対するJaccard係数Jの定義

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

分母: 2つの文書両方に含まれる形態素の総数
分子: 共通している形態素の数

→ 両方に含まれる形態素が多ければ
類似度が高い

例) Jaccard係数の計算例(1/3)

- 文書D_1と類似しているのは文書D_2, 文書D_3のどちらであるか判定するには？

D_1

虫歯の原因や治療法、予防法、子供の虫歯、治療費などについて分かりやすく解説します。

D_2

子供の虫歯は早期発見が重要で、最終的な治療費も安くなります。

D_3

うちの子供はいちごパフェが好きでよく食べます。

例) Jaccard係数の計算例(2/3)

- 文書D_1とD_2の類似度計算例(名詞だけを考慮)

D_1

虫歯の原因や治療法、予防法、子供の虫歯、治療費などについて分かりやすく解説します。

D_2

子供の虫歯は早期発見が重要で、最終的な治療費も安くなります。

D1={虫歯, 原因, 治療法, 予防法, 子供, 治療費, 解説}

D2={子供, 虫歯, 早期, 発見, 最終的, 治療費}

分母(すべての形態素の和集合): 10
分子(共通する形態素の総数): 3

Jaccard係数: 3 / 10 = 0.33

例) Jaccard係数の計算例(3/3)

- 文書D_1とD_3の類似度計算例(名詞だけを考慮)

D_1

虫歯の原因や治療法、予防法、子供の虫歯、治療費などについて分かりやすく解説します。

D_3

うちの子供はいちごパフェが好きでよく食べます。

D1={虫歯, 原因, 治療法, 予防法, 子供, 治療費, 解説}

D2={うち, 子供, いちご, パフェ}

分母(すべての形態素の和集合): 11
分子(共通する形態素の積集合): 1(子供)

Jaccard係数: 1 / 11 = 0.09
(D_2よりも類似度が低い！)

演習概要

- 本日の演習ではシェルスクリプトを用いて、MeCabで解析済みの2つの文書の類似度w計算するプログラムを作成します。

- 演習1から演習4までは説明通りに打ち込んでいけばok
- 演習5は演習4までの内容とこれまでに学んだシェルスクリプトの知識を使ってプログラムを記述してみてください。

演習1(ファイルの準備)

- 以下の3つのテキストファイルを作成しなさい。

ファイル名: d1.txt

虫歯の原因や治療法、予防法、子供の虫歯、治療費などについて分かりやすく解説します。

ファイル名: d2.txt

子供の虫歯は早期発見が重要で、最終的な治療費も安くなります。

ファイル名: d3.txt

うちの子供はいちごパフェが好きでよく食べます。リンゴやパイナップルも好きですが、パフェにはやはりイチゴが合うようです。

演習2(MeCabでの解析)

- d1.txt, d2.txt, d3.txtをぞれぞれ以下のようなコマンドで形態素解析し、名詞部分の結果のみ~~.mecabという形式で保存しなさい。

  ~ishigaki/bin/mecab < d1.txt | grep “名詞,” > d1.mecab

演習3: 分母を求めるUNIXコマンド

- 以下のコマンドで分母を求めることが出来ます。

  $ cat d1.txt d2.txt | sort | uniq | wc –l

  catコマンドで2つのファイルを結合
  sort | uniq | wc –lで形態素の数をカウントします。

演習4: 分子を求めるUNIXコマンド

- 以下のコマンドで分子を求めることが出来ます。

  $ cat d1.txt d2.txt | sort | uniq -d | wc –l

  catコマンドで2つのファイルを結合
  sort | uniq -d | wc –lで形態素の数をカウントします。
  uniqコマンドの-dオプションを使うと論理積(∩)を求めることが出来ます。

演習5(小数の計算方法)

- UNIXコマンドで小数演算を行うには以下のようなコマンドを使用します。入力して計算されることを確かめなさい

  。
  $ echo “scale=3; 2.0/3.0” | bc

  1. echoのあとに、scale=3; のように小数点以下の桁数を指定
  2. その後、小数演算の式を記述しbcコマンドにパイプします。

演習6

- 演習４までの内容をもとにd1.mecabとd2.mecabの類似度を計算するシェルスクリプト(jaccard.sh)を記述しなさい

  模範解答は次のページにあるので、
  参考にしながら作っても良いです。

演習5の解答

```
 1 #!/bin/bash
 2 # 変数の初期化
 3 bunbo=1
 4 bunshi=0
 5 # 分母と分子の計算
 6 bunbo=`cat d1.txt d2.txt | sort | uniq | wc -l`
 7 bunshi=`cat d1.txt d2.txt | sort | uniq -d | wc -l`
 8 # Jaccard係数の計算と表示
 9 result=`echo  "scale=3; $bunshi/$bunbo" | bc`
10 echo "Jaccard係数: ${result}"
```