テキストマイニング個人課題

<<必須>>

1. 青空文庫から任意の作家・作品を２つ選んでデータをダウンロードし、それぞれ単語の出現頻度を調べ、どのような内容か推測しなさい

2. 青空文庫以外の無料の小説サイト（星空文庫など）やニュースサイトの記事からテキスト（500文字以上などある程度の量）をコピーしてtextファイルとして保存し、単語の出現頻度を調べなさい。その後、形態素辞書をmecab-ipadic-neologdに変更して単語の出現頻度を調べ、辞書の変更によって、どのような変化があったか確認しなさい。

3. 青空文庫から複数の作品をダウンロードし、類似度を計算せよ。また、クラスタリングしてみよ。
   また、クラスタと筆者やジャンルの関係性について考察を述べよ

<<任意>>

4. wikipediaやニュースサイト,Q&Aサイトなど、カテゴリ情報があるテキストを複数カテゴリ・複数テキストをあつめ、教師あり学習で学習したのち、学習に使用しなかったデータのカテゴリを推定せよ。また、その結果について考察せよ

5. 3または4で使用するクラスタリング/分類手法を他のアルゴリズムに変更し、結果の違いを比較せよ

テキストマイニング個人課題

<<提出物>>

1. 基本的にjupyter notebookにソース及び考察結果等のコメントをまとめ、.jpynbファイル及び使用したデータファイル及びプログラムで生成されたファイル（形態素解析結果など）一式を以下フォルダ構成に纏めてzipファイルに纏めて提出。

```
学籍番号_氏名
+課題_1.jpynb
+課題_2.jpynb
・・・
+data
    +（データファイル）
```

・どのようなデータを利用したか（取得元,著者,作品名,取得元URL等）がわかるように記載すること
・なんらかの理由により、 jupyter notebookを利用しない場合は文書ファイル等によるレポート+ソース+データファイルでもよい。その場合は、課題と利用したソースの対応がわかるように記載すること

<<提出方法>>
メールにて上記zipファイルを提出
宛先　：　exp4.bd@gmail.com

期限　：　10月27日(金) 17:00 JST

テキストマイニング補足

- 青空文庫からデータを取得する場合、「新字新仮名」のものを選ぶ。
  ※「旧字旧仮名」では正しく形態素解析できない箇所が多くなり、実用的な分析が行えない
- 形態素解析の前に行う前処理で、不要部分の削除だけでなく、半角全角の統一などの「正規化処理」を行うと、形態素解析の精度が向上する。
  一般的に行われる正規化処理の内容及びPythonでの正規化処理コードについてはmecab-ipadic-neologdのwiki「解析前に行うことが望ましい文字列の正規化処理」を参照
  https://github.com/neologd/mecab-ipadic-neologd/wiki/Regexp.ja
- 文書特有の語（人名・商品名etc)は形態素解析器の辞書に登録することで正しく形態素解析できるようになる。MeCabでは、システム辞書とは別にユーザ辞書も使用することができる。
  ただし、辞書登録には「連結コスト」等詳しい知識が必要になる。
  natto-pyではparse()メソッドのfeature_constraintsキーワード引数に単語(表層文字)と品詞のタプルのタプルを指定することで簡易的な辞書として利用できる
  参考: feature_constraintsサンプル.ipynb