Regressionを用いた作品タイトルの略語予測

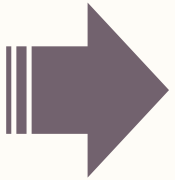
チームG1

発表の流れ

本実験の背景と目的

背景

・言語処理における新たなアプローチを行いたい
・法則性の無さそうな言葉から予測したい



作品タイトルの略称

目的

・略前タイトル→略後タイトルの**法則性**を見つけ出す
Ex)けものフレンズ　→　けもフレ
・導き出した法則性から**新たな略称**を予測する

得られる利益： 初見言葉への対応策

実験の流れ(予定)

データセット生成

データ収集(タイトル群)
・手書き
・webスクレイピング

データの数値化
・母音子音情報を数値化
・品詞の関連性を数値化
・略前の表記体系を保存

機械学習

機械学習
・Clusteringによる略称の分類
・Regressionによる実数値予測
・RNNによる精度向上

モデルを用いて略タイトルを予測
・数値として出力
・ひらがな、カタカナのみで出力
・元表記体系で出力

成果

・予測値(数値)をカタカナやひらがなに変換・出力
・入力、出力を元表記体で表示
・「!?」や「☆」などの記号にも対応する

実験の流れ(成果)

データセット生成

データ収集(タイトル群)
- **手書き**
- **webスクレイピング**

データの数値化
- **母音子音情報を数値化**
- 品詞の関連性を数値化
- 略前の表記体系を保存

機械学習

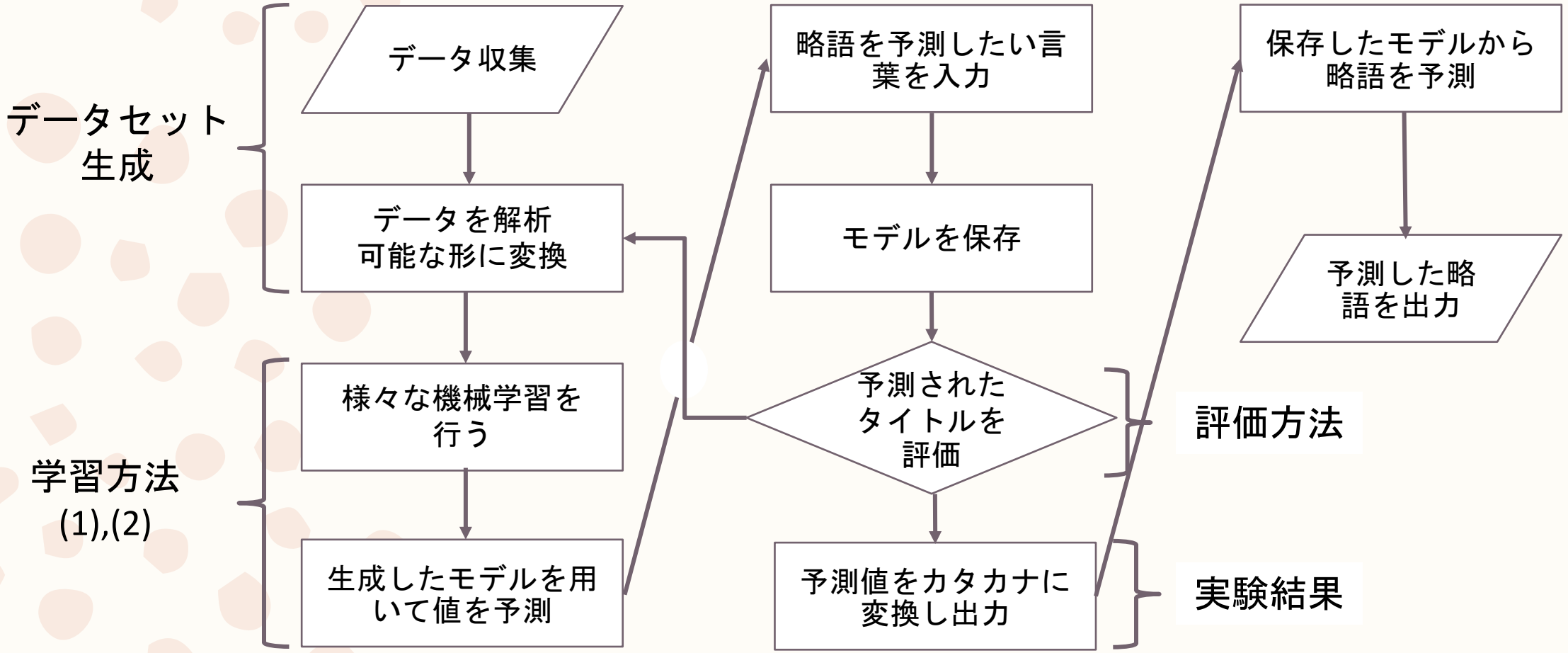
機械学習
- Clusteringによる略称の分類
- **Regressionによる実数値予測**
- RNNによる精度向上

モデルを用いて略タイトルを予測
- **数値として出力**
- ひらがな、**カタカナのみで出力**
- 元表記体系で出力

成果

- **予測値(数値)をカタカナやひらがなに変換・出力**
- 入力、出力を元表記体で表示
- **「!?」や「☆」などの記号にも対応する**

実験設計


データセット
生成
データ収集
データを解析
可能な形に変換
学習方法
(1),(2)
様々な機械学習を
行う
生成したモデルを用
いて値を予測
略語を予測したい言
葉を入力
モデルを保存
予測された
タイトルを
評価
予測値をカタカナに
変換し出力
評価方法
実験結果
保存したモデルから
略語を予測
予測した略
語を出力

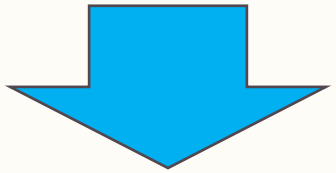
データセット生成

タイトル群を入手

アニメタイトルとその略称
まとめサイトなどから
webスクレイピングを
活用してデータを取得

タイトル群をベクトル変換

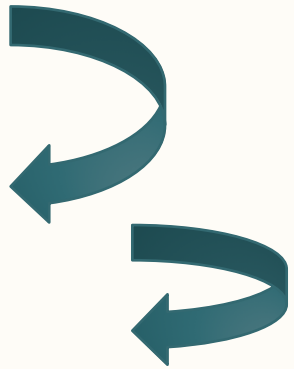
それぞれの情報を配列に
代入・統合する
(詳しくはプログラムを参照)



LinerRegression関数を用いてモデル生成

学習方法(1)

- **LinerRegression関数で学習**(パラメータはデフォルト)
- **課題**：データ数が少ないため例外に引っ張られる



- **解決**：**交差検証**を活用した学習を行う
- **利益**：
  - １つのデータに対してテストを行っていくため予測が難しいタイトルがわかる
  - データを効率よく活用できる

交差検証とは？

交差検証(Cross-Validation)

交差検証(Cross-Validation)とは、

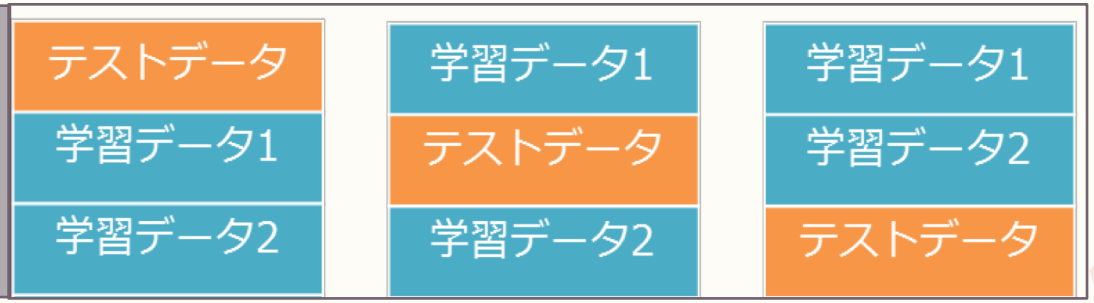
1.データを分割
2.一部データで解析(訓練)
3.残ったデータで評価(テスト)
4.分割部分を変更し2と3を繰り返す
5.複数回行ったテスト結果をもとに評価する

の手順でデータ解析手法の「良さ」を評価できる

LinerRegression
に適用

比較的少ないデータを
効率よく活用できる

機械学習での交差検証例



学習方法(2)

予測された値の調整(0-1の範囲)

欲しい予測値(理想)

```
[[0 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 ...
```

実際の予測値(現実)

```
[[ 0.15361429  0.64284875 -0.13029998 ...  0.          0.
   0.        ]
 [ 0.01245172  0.91328923  0.00617569 ...  0.          0.
   0.        ]
```

文字変換するための調整が必要

評価方法

平均絶対誤差(Mean Absolute Error)：
・教師データのベクトル値
・予測値
の誤差の平均
評価関数mean_absolute_error()

独自の評価関数を用いて検証:
・教師データの略後タイトルの文字と
・予測値からカタカナに変換したタイトルの文字
の差異を点数化し平均
評価関数calc_accuracy()

```
セ ラ ム ン
イ イ エ カ


イ ナ ク ロ
イ ナ イ レ


ジ ン ス イ
キ ン ア ク
```

```
model:Linear
mean_absolute_error:0.11583927380661907
calc_accuracy:0.2
```

```
model:tree
mean_absolute_error:0.07927272727272727
calc_accuracy:0.19101123595505617
```

```
model:Bagging
mean_absolute_error:0.08605818181818183
calc_accuracy:0.2125
```

実験結果

- **交差検証**有無の結果比較
  - 有り：mean_absolute_error：0.15981188924141454
  - 無し：mean_absolute_error：0.18797223803418409

- 予測値調整の結果比較
  - i<=0.45の場合：0.2855...
  - i<=0.475の場合：0.2867...
  - i<=0.5の場合：0.2843...

今後の課題

- RNNの実装
- より多くのデータを使った学習
- クラスタリングを用いたモデルの複数生成とその利用
- タイトルの特徴を母音子音だけでなく、品詞のつながりなど他の特徴で学習する
- アンサンブル学習による精度向上

考察

- 予測対象の**種類や例外が多い**場合に**重線形回帰**を行うと高い精度を出すことが難しいと考える
- 母音子音の発音を特徴としていたが、**データごとの類似点が少ない**ためクラスタリングの範囲指定が困難である
- 言葉の音に**焦点を当ててデータを扱おう**とすると、**構造が複雑**になってしまうため、**複数のカラムに分けることが必要**である

参考文献

- Cross-Validation(交差検証)を行い、最適なアルゴリズム/パラメータを発見する(Vertica9.0新機能), http://vertica-tech.ashisuto.co.jp/cross_validation/
- scikit-learnで回帰モデルの結果を評価する,https://pythondatascience.plavox.info/scikit-learn/回帰モデルの評価