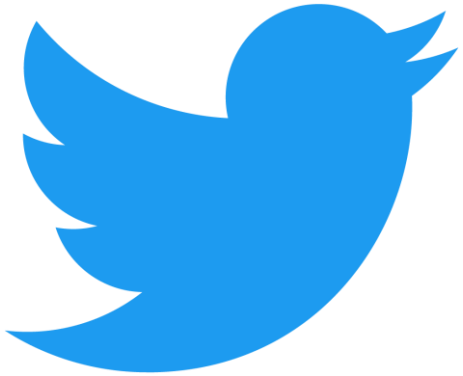
SNSテキスト分類

當麻僚太郎

やりたいこと

- 入力された文章がどのSNS「らしい」かを判定する

- 対象のSNSはTwitterとInstagram



手段

1．　TwitterとInstagramの投稿からテキストを取得

2．　学習しやすいように前処理

3．　分類モデルの学習

手段

１．　TwitterとInstagramの投稿からテキストを取得

２．　学習しやすいように前処理

３．　分類モデルの学習

Twitterのテキスト取得

- tweepy：PythonでTwitter APIを扱えるライブラリ

- Twitterのdeveloperサイトから審査
  → APIトークンを取得

- スクレイピングではない

Instagramのテキスト取得

- seleniumでスクレイピング
- Webブラウザを開いて自動で操作
- ログインの操作やスクロールも書く必要あり
- 面倒

学習データセット

- Twitterのテキスト：240ツイート分
- Instagramのテキスト：181投稿分

手段

１．　TwitterとInstagramの投稿からテキストを取得

２．　学習しやすいように前処理

３．　分類モデルの学習

データの前処理

- テキストそのままを学習データにするのは面倒
- 形態素解析をして、品詞の出現頻度で分類してみよう
- 形態素解析エンジンJUMAN++

JUMAN++

- 京都大学の黒橋・河原研究室で開発
- RNN言語モデルを用いた形態素解析器
- MeCabよりも砕けた表現に対応できる
- 解析速度は遅い

手段

１．　TwitterとInstagramの投稿からテキストを取得

２．　学習しやすいように前処理

３．　分類モデルの学習

分類モデル

- sklearnのSVMを使用

- `sklearn.svm.SVC(C=1, kernel='rbf')`

- 学習データ：テストデータ ＝ 7：3

分類結果

	適合率	再現率	f値
Twitter	0.82	0.81	0.81
Instagram	0.73	0.73	0.73

accuracy 0.78

考察

- 時系列データとしてRNNやLSTMで処理したらもっと良い精度が出そう
- データ量を増やした方が良さそう
- 特徴に関係ない削れる部分がありそう
- そもそも明らかな違いがあるか？