Pythonの今後の展望

～機械学習とスクレイピング～

2019年度　Python部会　Aチーム

部会員紹介

◆スクレイピングチーム

株式会社オーイーシー	村田　美香
株式会社オーガス	松村　哲
大分大学	松田　千輝

◆機械学習チーム

株式会社オルゴ	森　　逸充
株式会社オーイーシー	大石　竜也
株式会社モバイルクリエイト	伊藤　僚祐

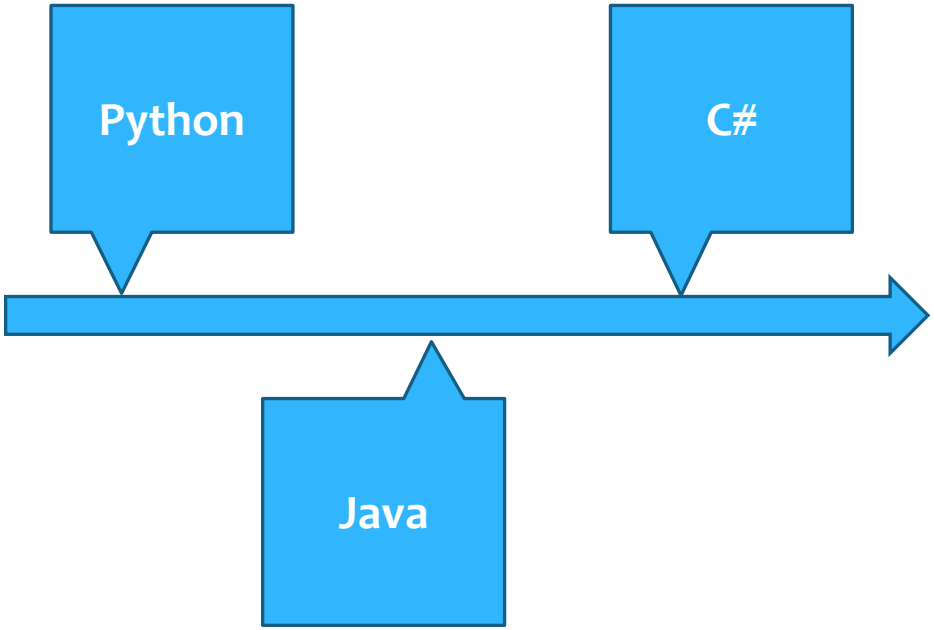
目次

- 1.Pythonとは
- 2.スクレイピングチームの研究結果
- 3.機械学習チームの研究結果
- 4.まとめ

1.Pythonとは？

Pythonの歴史

- Python
  - 1991年に開発
- Java
  - 1995年に開発
- C#
  - 2002年に開発



Pythonのメリット・デメリット

- メリット
  - コードが読みやすい
  - ライブラリが豊富
- デメリット
  - 実行速度が遅い
  - インデントが必須

Pythonのメリット1

- コードが読みやすい
  - Pythonはコードが短く、だれが書いても同じようなソースコードになり、初心者でも理解がしやすい。

- Java

```
public class HelloWorld{
    public static void main(String[] args){
        System.out.println("Hello World!!");
    }
}
```

- Python

```
print("Hello World")
```

Pythonのメリット2

- ライブラリが豊富
  - Pythonは数多くのライブラリが作成されており、目的に合わせて最適なライブラリを使用できる。

- ライブラリ例
  - pip：パッケージマネージャー
  - Python-dateutil：日付計算ライブラリ
  - OpenCV：画像処理ライブラリ
  - Keras：深層学習ライブラリ

Pythonのデメリット1

- 実行速度が遅い
  - 処理実行時にソースコードから機械語に変換するため速度が遅い
- 改善手段
  - ライブラリを用いる
  - 速度の速い他言語で実行した結果を呼び出す

Pythonのデメリット2

- インデント（字下げ）が必須
  - 文のブロックに括弧を使用せずインデントで表現するため、インデントが必須。

- ソース

```
a=3
if a == 5:
    print “AAA” #if文の対象
    print “BBB” # if文の対象
print “CCC” # if文の対象外
```

- 結果

```
CCC

a=5の場合は
AAA
BBB
CCC
```

2.スクレイピングチームの研究結果

2.スクレイピングチームの研究結果

◆スクレイピングとは

- Webスクレイピングのことを指し、Webサイトから任意の情報を取得する技術。
- WebサイトのHTMLをもとに必要なデータを取得する。

◆代表的なライブラリ

- Requests　　：HTTPライブラリでWebページを取得する。
- Beautiful Soup：取得したHTMLから情報を抽出する。
- Selenium　　：ブラウザを操作する。

2.スクレイピングチームの研究結果

◆注意点

- 無許可のスクレイピングは拒否（禁止）されているものもある。
  →罰則を受けたり違法となり法的に処罰される！

各サイトの利用規約やルール、著作権法に違反しないか等をよく確認して、アクセス頻度なども考えて迷惑をかけないように注意が必要。

2.スクレイピングチームの研究結果

◆やってみたこと　その1

- 天気情報サイトから、任意の地域の本日の天気をtxtファイルとして取得する。

■使用した主なライブラリ
requests、Beautiful Soup

2.スクレイピングチームの研究結果

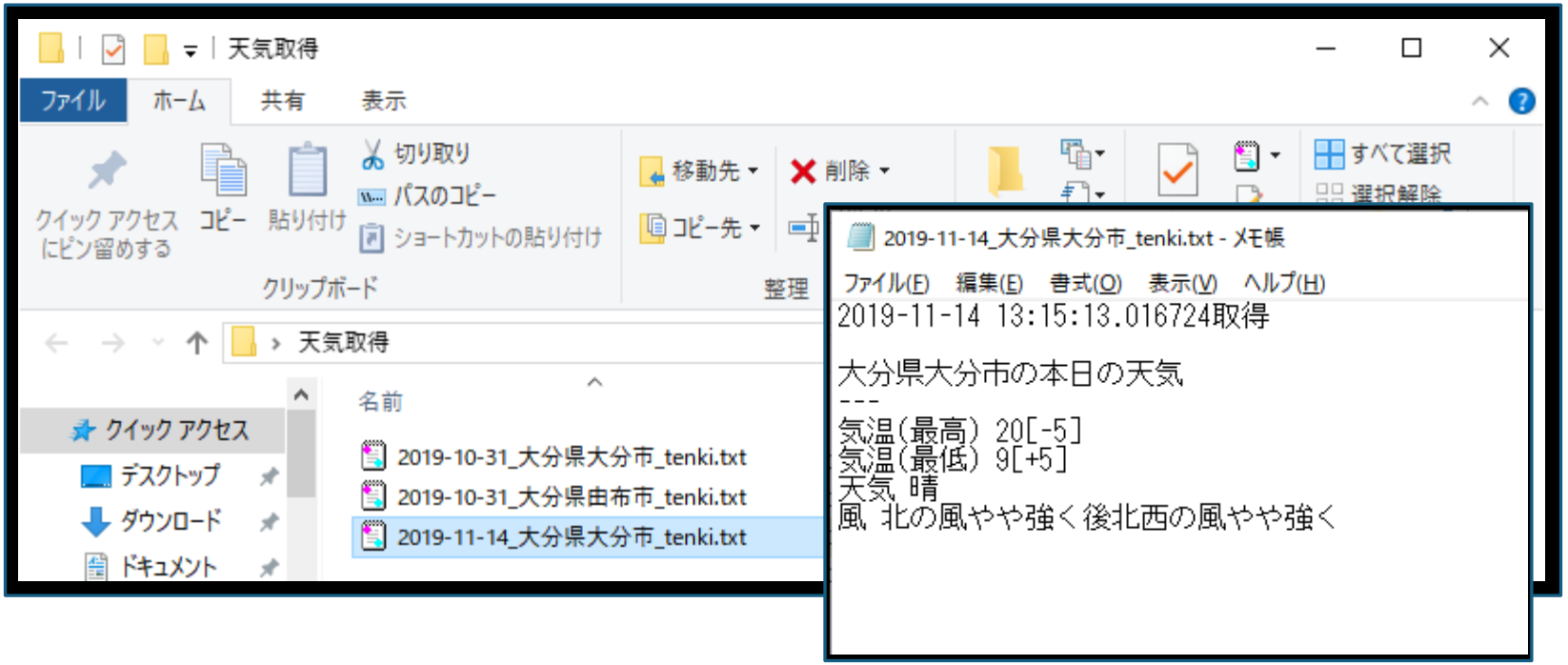
- 天気情報の取得の流れ
  1. コマンドプロンプトでpyファイルと地域を指定し、プログラムを実行。
  2. 特定の天気情報サイトから指定した地域の当日の天気情報を取得し、txtファイルで出力。

2.スクレイピングチームの研究結果

- コマンドプロンプトでpyファイルと地域を指定し、プログラムを実行。

2.スクレイピングチームの研究結果

◆特定の天気情報サイトから指定した地域の当日の天気情報を取得し、txtファイルで出力。



2.スクレイピングチームの研究結果

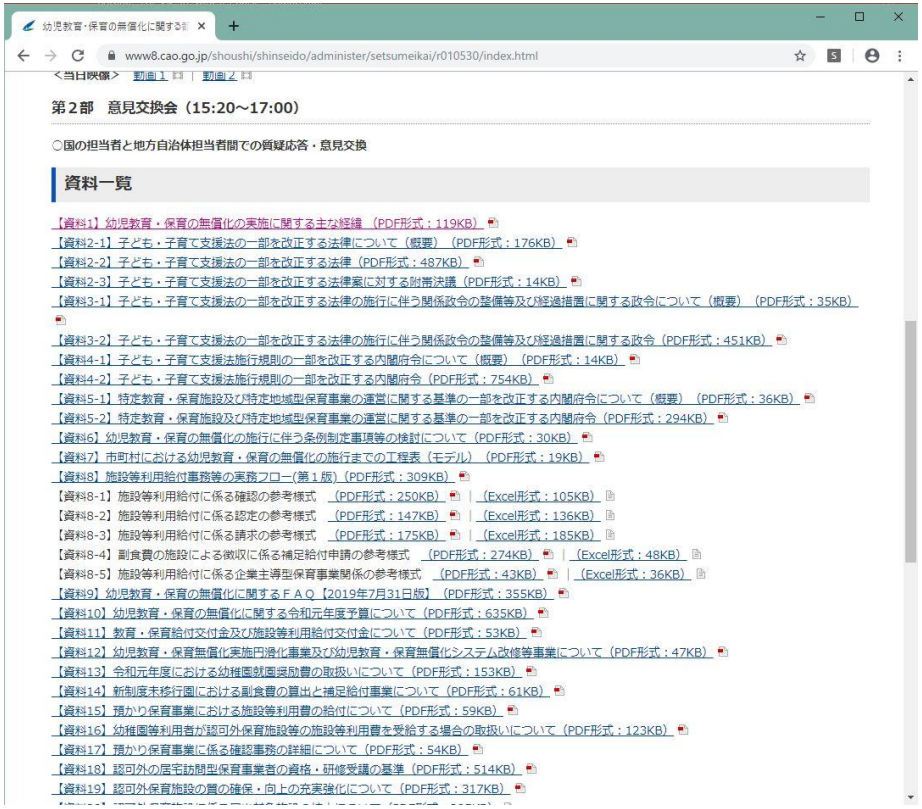
◆やってみたこと　その2

- 任意のサイトにある全てのpdfファイルをスクレイピングし、指定したフォルダにダウンロードする。
  ■使用した主なライブラリ
  requests、Beautiful Soup

- 指定フォルダ内のpdfファイルを全て連結する。
  ■使用した主なライブラリ
  PyPDF2、glob

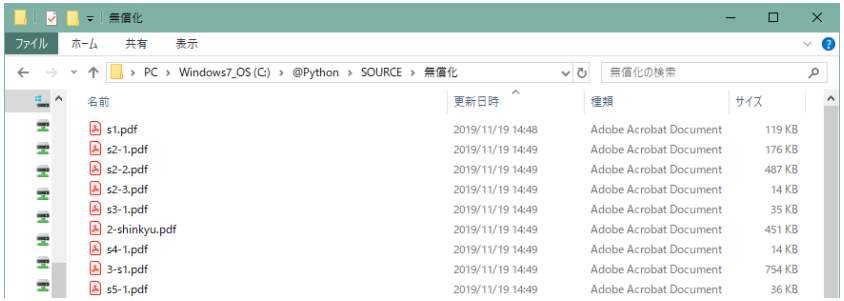
2.スクレイピングチームの研究結果

【任意のサイト】

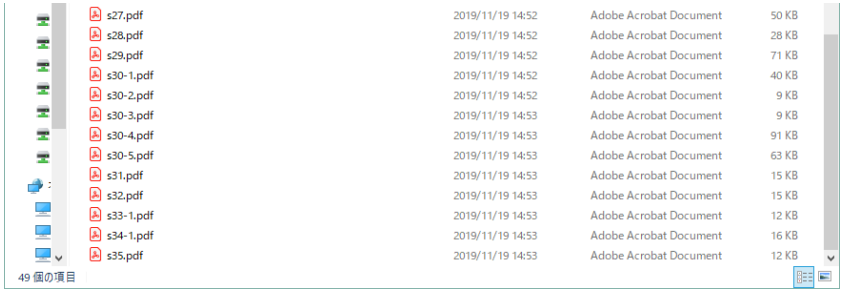


【pdfファイルをスクレイピング＋ダウンロード：pdf_DL.pyを実行】

【実行結果】

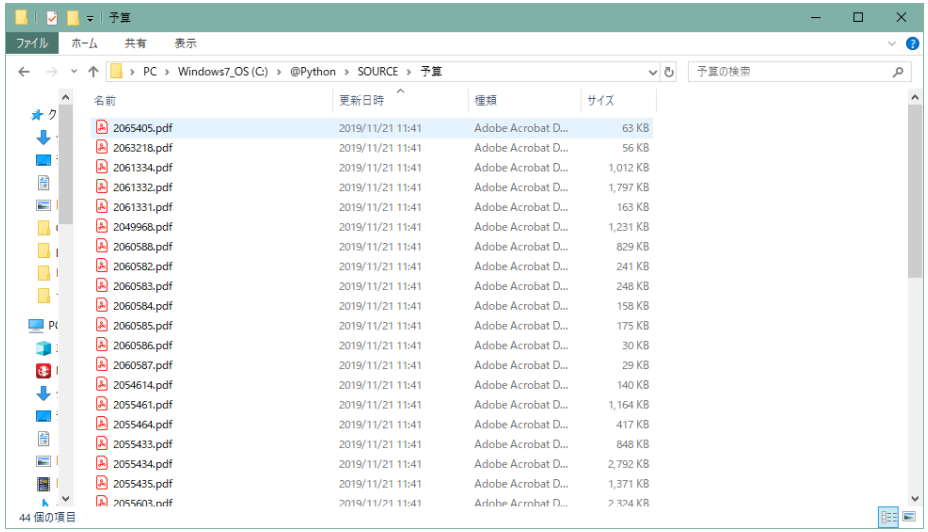


＜中略＞



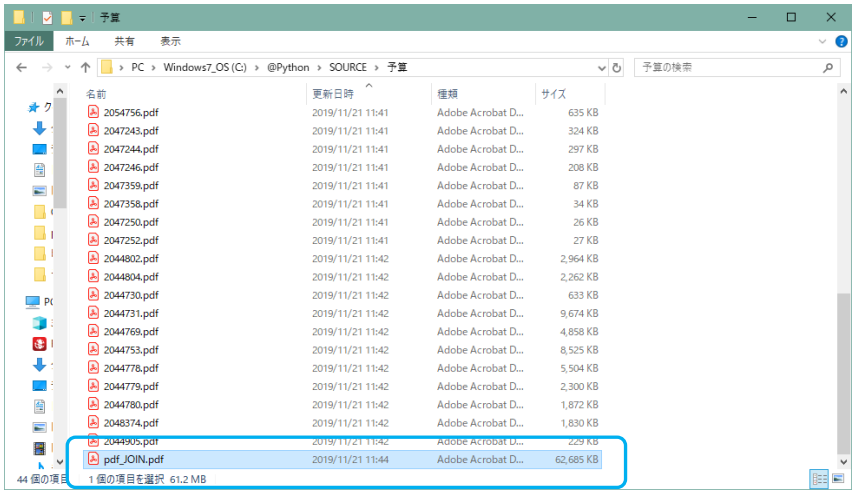
2.スクレイピングチームの研究結果

【指定フォルダ】

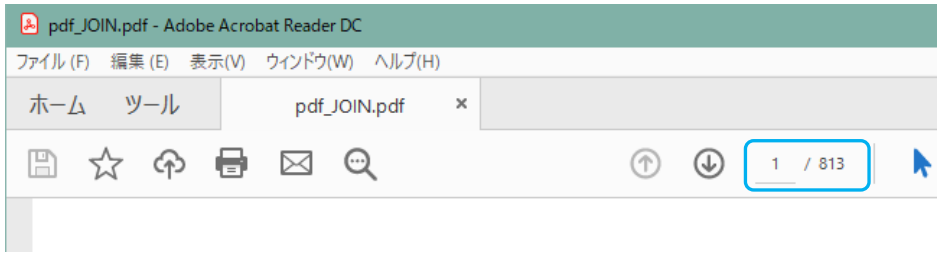


【pdfファイルを結合：pdf_JOIN.pyを実行】

【実行結果】



【実行結果(結合されたpdf：813ページ)】



2.スクレイピングチームの研究結果

◆苦労したこと

- 自宅では実行できたソースコードが会社ではエラーになる。
  →認証プロキシの環境かそうでないかの問題。

- pdfファイルの編集で「PyPdf2」を使ったが、エラーになる。
  →保護されているドキュメントはエラーになる。

- エラーが発生した場合、情報が英語で大量に出る。

2.スクレイピングチームの研究結果

◆感想

- ウェブ上や書籍等に作成の参考になる資料が多くあり、行き詰っても解決しやすかった。
- ライブラリが豊富で、やりたいことを実現させやすかった。
- コードが分かりやすく、難しい記述も少なく、初心者でもある程度扱いやすい言語だと感じた。

⇒ゼロからの独学でもプログラムを作成できる！
（でも、何事も何もわからないところから始めるのはやる気と時間が必要。）

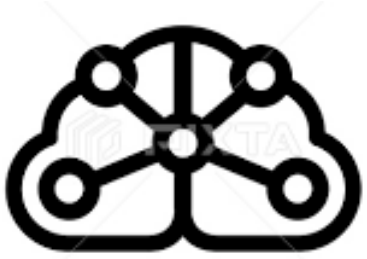
3.機械学習チームの研究結果

3.機械学習チームの研究結果

◆機械学習とは

”特徴”をつかみ
”法則化”する

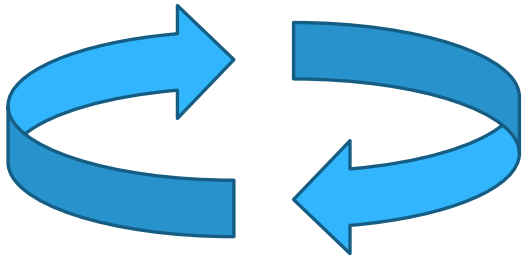
データから反復学習し、
学習結果を法則化をする



機械学習

法則を
”自動化”する

ノンプログラミングで
システム化し、再現性を作る

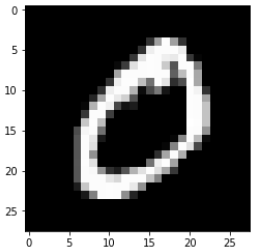


3.機械学習チームの研究結果

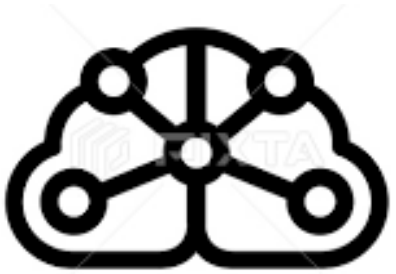
◆機械学習の仕組み

データセット

MNIST



学習モデル

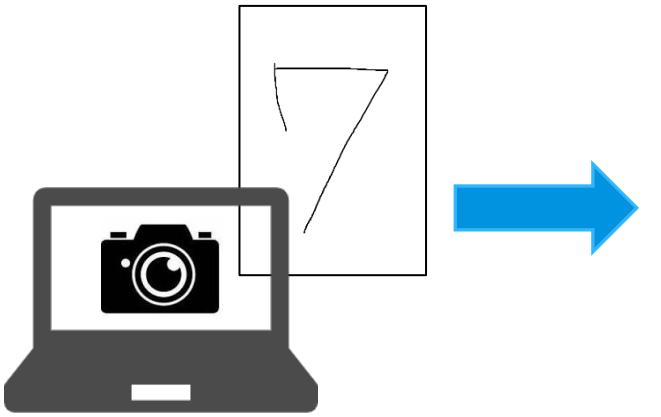


特徴を捉え、学習する

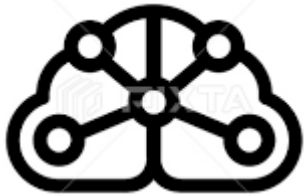
3.機械学習チームの研究結果

◆機械学習の仕組み

手書き数字



学習済みモデル



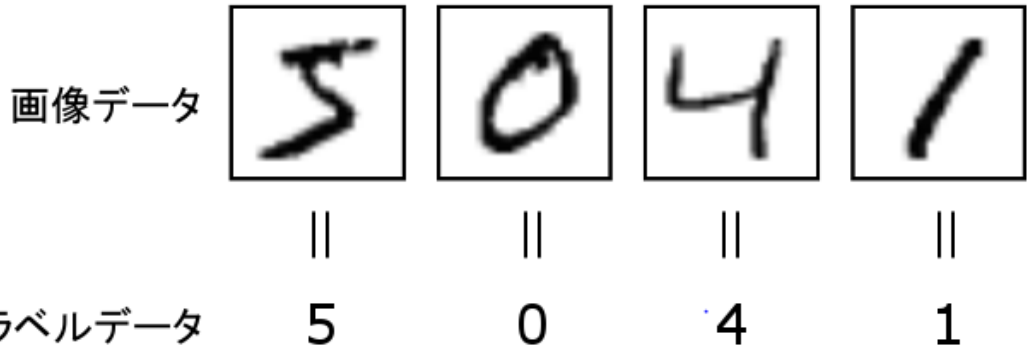
学習済みモデル
で予測

予測結果

3.機械学習チームの研究結果

◆使用したライブラリ、データセット

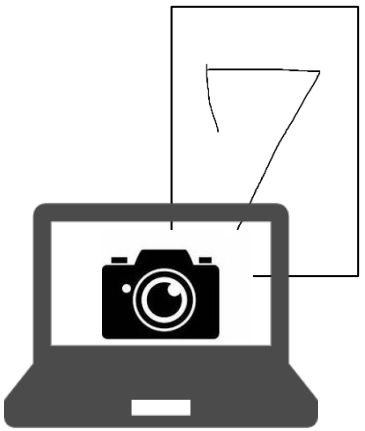
- Keras
  - ・Pythonで書かれたオープンソースのライブラリ
  - ・他のライブラリと比べコードを書くのが簡単
- MNIST
  - ・28ピクセル×28ピクセルの手書き数字画像60,000枚と、テスト画像10,000枚を集めた、画像データセット



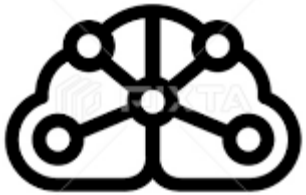
3.機械学習チームの研究結果

◆デモンストレーション

手書き数字



学習済みモデル



学習済みモデル
で予測

予測結果

3.機械学習チームの研究結果

◆苦労したこと

- 機械学習の理論を理解することに苦労した。
  →ソースコードはかけても理解が難しい

- 理想の予測結果が得られない
  →MINISTの画像をかけると99%ぐらいの正答率
  写真で撮った物だと、60%ぐらいの正答率
  ※光が当たったり、数字が斜めになる等が原因

4.まとめ

4.まとめ

◆Pythonの人気度　引用：IEEE Spectrum「The Top Programming Languages 2019」より

1	Python		100.0
2	Java		96.3
3	C		94.4
4	C++		87.5
5	R		81.5
6	JavaScript		79.4
7	C#		74.5

4.まとめ

◆Pythonの今後

- 近年注目されているAI（人工知能）開発でさらに活躍
- 蓄積されたビッグデータの解析を機械学習で行う、Webスクレイピングで欲しい情報を効率的に収集する等、マーケティングやビジネス戦略で有効（もちろん個人でも）
- RPAのように、業務の効率化を行うことも可能
- 日本では、海外に比べるとまだエンジニア不足

4.まとめ

◆活動を通して

- 当初のテーマ
  スクレイピングして集めた画像で機械学習を行い、
  画像内の英単語を翻訳するアプリを作成

◆困難だった事

- 画像内の英単語を探索し、認識させること
- スクレイピングで英単語が入った大量の画像のみを取得すること
- 利用する複数ライブラリの仕組みの理解

ご清聴ありがとうございます