テキストから感情を分析し
ポジ・ネガ判定をするシステムの作成

総合情報学科　情報システム学系
知能情報システム研究室
永井ゼミ ４年　石川　輝星

研究の動機

・１つのトピック（商品、システム、映画など）に対して、SNSやレビュー欄からどんな感情を持ったのかを知りたく、評価値を作成することで分析、理解しやすくなると考えたため。

研究の目的

・星を用いた５段階評価や数字での評価などがない、レビューやSNSなどからトピックに対する（個々や全体）評価が明確になる。

・比較、類似している物がわかることで分析対象の理解が深まる。

システムの作成順序

- レビュー（テキスト）から文を抜き取り、前処理
- 形態素解析を用いて、品詞ごと分解
- 感情分析を行い、単語ごとに数値化する
- 数値の平均でポジティブ、ネガティブ判定（未定）
- 出現頻度の高い固有名詞を表示（同じトピックの文から）
- 数値からポジ、ネガの判定を分類分けし、固有名詞の集計で内容の把握

文章が無いレビューの除外
日本語ではない文章の除外

作成するシステム

- テキストからポジティブ、ネガティブを判定し、数値ではなく基準を設け５段階に分ける。

  １.Very Neg ２.Neg ３.Neutral ４.Pos ５Very Pos

  とても悪い　悪い　　普通　　良い　とてもいい
- テキストの中から出現頻度の高い単語を出す。
- トピック全体の判定
- 上記三つを表示できるシステム

検証

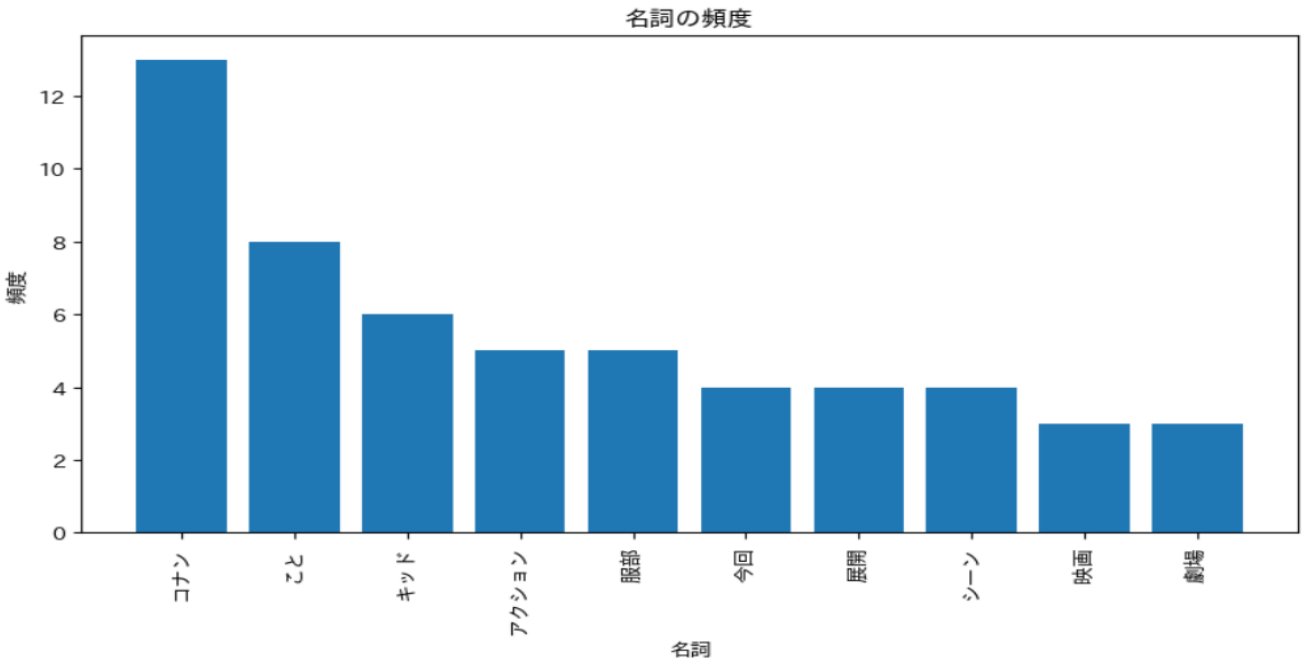
- 実際にレビューやコメントを書いた本人が感じている感情の度合いと類似するかを図るため、評価値がついているテキストを対象にして作成する。
- 映画.comやamazonなどで出力を調べつつ実際の数字や評価と同じ評価値を出力するかで検証し改良していく。

進捗（１）

- 品詞ごとに分解するために用いるMeCabを動かした。
- 文章は２０００字（コナンの映画レビュー５件）
- Matplotlibで名詞だけを抜き取りグラフ化した

```
tagger = MeCab.Tagger()
print(tagger.parse("近年の劇場版シリーズと比較すると、推理パートが多かった。"))
```

```
近年	キンネン	キンネン	近年	名詞-普通名詞-副詞可能
の	ノ	ノ	の	助詞-格助詞
劇場	ゲキジョー	ゲキジョウ	劇場	名詞-普通名詞-一般
版	バン	ハン	版	名詞-普通名詞-助数詞可能
シリーズ	シリーズ	シリーズ	シリーズ-series	名詞-普通名詞--
と	ト	ト	と	助詞-格助詞
比較	ヒカク	ヒカク	比較	名詞-普通名詞-サ変可能	0
する	スル	スル	為る	動詞-非自立可能	サ行変格	終止形-一般
と	ト	ト	と	助詞-接続助詞
、	、	、	補助記号-読点
推理	スイリ	スイリ	推理	名詞-普通名詞-サ変可能	1
```



進捗（２）

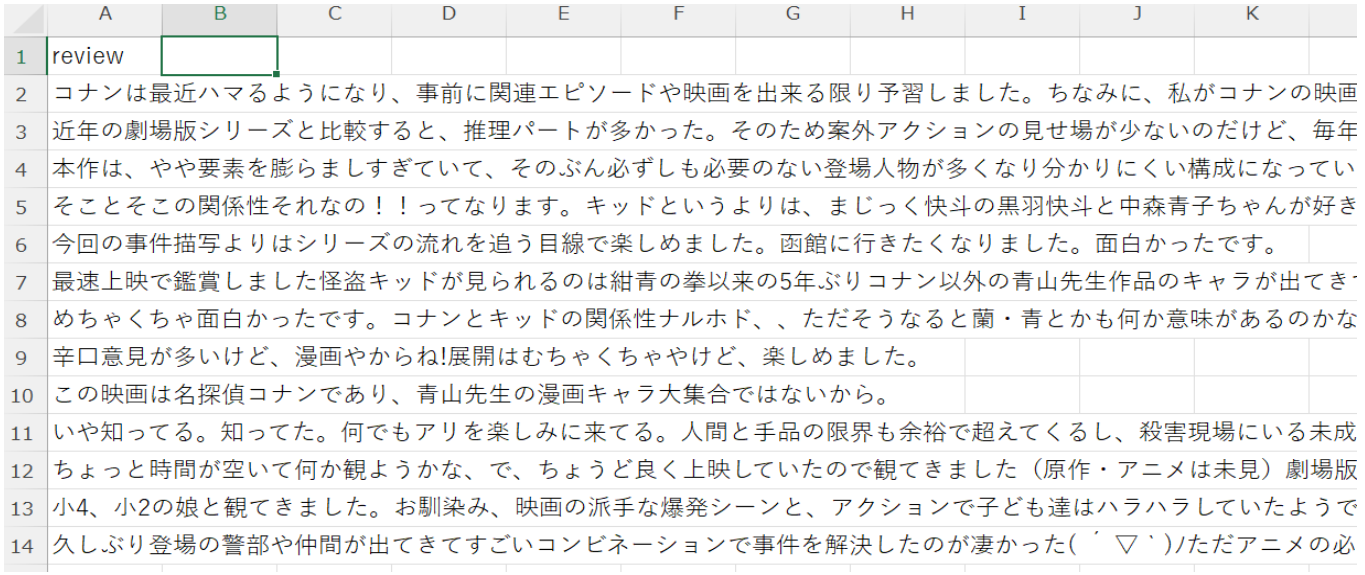
- 映画レビューのウェブサイトのレビュー欄を抽出した。
- HTMLの要素を指定してCSVファイルとして保存。

```
import requests
from bs4 import BeautifulSoup
import pandas as pd

url = 'https://eiga.com/movie/100801/review/'
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

reviews = []
for review in soup.find_all('p',class_='short'):
    reviews.append(review.text)

df = pd.DataFrame(reviews, columns=['review'])
df.to_csv('reviews.csv',index=False,encoding='utf-8-sig')
```



	A	B	C	D	E	F	G	H	I	J	K
1	review										
2	コナンは最近ハマるようになり、事前に関連エピソードや映画を出来る限り予習しました。ちなみに、私がコナンの映画										
3	近年の劇場版シリーズと比較すると、推理パートが多かった。そのため案外アクションの見せ場が少ないのだけど、毎年										
4	本作は、やや要素を膨らましすぎていて、そのぶん必ずしも必要のない登場人物が多くなり分かりにくい構成になってい										
5	そことそこの関係性それなの！！ってなります。キッドというよりは、まじっく快斗の黒羽快斗と中森青子ちゃんが好き										
6	今回の事件描写よりはシリーズの流れを追う目線で楽しめました。函館に行きたくなりました。面白かったです。										
7	最速上映で鑑賞しました怪盗キッドが見られるのは紺青の拳以来の5年ぶりコナン以外の青山先生作品のキャラが出てき										
8	めちゃくちゃ面白かったです。コナンとキッドの関係性ナルホド、、ただそうなると蘭・青とかも何か意味があるのかな										
9	辛口意見が多いけど、漫画やからね!展開はむちゃくちゃやけど、楽しめました。										
10	この映画は名探偵コナンであり、青山先生の漫画キャラ大集合ではないから。										
11	いや知ってる。知ってた。何でもアリを楽しみに来てる。人間と手品の限界も余裕で超えてくるし、殺害現場にいる未成										
12	ちょっと時間が空いて何か観ようかな、で、ちょうど良く上映していたので観てきました（原作・アニメは未見）劇場版										
13	小4、小2の娘と観てきました。お馴染み、映画の派手な爆発シーンと、アクションで子ども達はハラハラしていたようで										
14	久しぶり登場の警部や仲間が出てきてすごいコンビネーションで事件を解決したのが凄かった(´▽`)/ただアニメの必										

進捗（３）

- 現段階では事前学習済みモデルを用いて動作確認、他の感情分析のやり方を調べている段階

```
from transformers import pipeline, AutoModelForSequenceClassification, BertJapaneseTokenizer

model = AutoModelForSequenceClassification.from_pretrained('koheiduck/bert-japanese-finetuned-sentiment')
tokenizer = BertJapaneseTokenizer.from_pretrained('cl-tohoku/bert-base-japanese-whole-word-masking')
classifier = pipeline("sentiment-analysis", model=model, tokenizer=tokenizer)

result = classifier("コナンは最近ハマるようになり、事前に関連エピソードや映画を出来る限り予習しました。ちなみに、私がコナンの映画を劇場で観るのは
print(f"label: {result['label']}, with score: {round(result['score'], 4)}")
```

出力

```
label: POSITIVE, with score: 0.9673
```

- Labelがポジティブ、ネガティブ出力
- scoreが高ければ判定が正確、低ければ曖昧であり中立

現状と今後の予定

現状

・形態素解析から頻出単語を表示

・感情分析を行った

・ウェブサイトからレビューだけを抜き出しcsvファイルに保存

予定

１,感情分析を行う上での方法を考える。（BERTについての理解）

２,csvファイルから読み込む形で形態素解析、感情分析を行いたい

３,類似研究との差別化について考える

（テキストの分野を絞っていないので、精度ではなく結果を視覚化しやすくなる機能追加で考えていく）