

1 日本で上映中の最新映画の情報収集

概要

Filmarksサイトの「上映中の最新映画おすすめ人気ランキング」をスクレイピングし、映画の各種情報を収集し、Excelにまとめた。フィルターで見たい映画を絞り込んだ。また、ポスターをダウンロードし、JPGファイルに対応する映画名をつけた。

使用技術

Windows / Python3, selenium, pillow, urllibなど

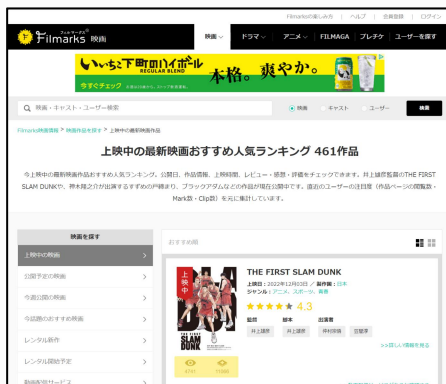


図1 Filmarksサイトの「上映中の最新映画おすすめ人気ランキング」画面
(2022年12月04日)

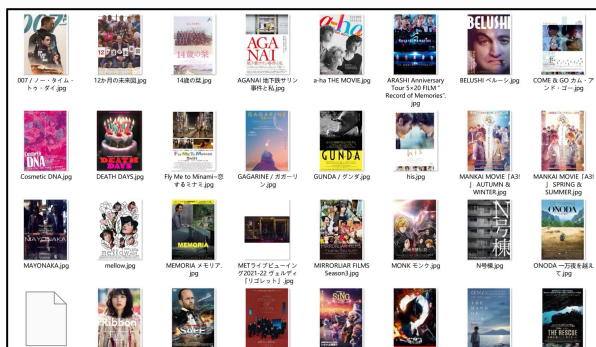


図3 名付けたポスターのJPGファイル

| | A | B | C | D |
|-----|--|-------------|-----|------|
| | 映画名 | 上映日 | 製作国 | 評価点数 |
| 1 | | | | |
| 3 | 劇場版 Free!-the Final Stroke- 後編 | 2022年04月22日 | 日本 | 4.3 |
| 7 | ファーストミッション | 2022年05月06日 | 日本 | 4.3 |
| 10 | 映画 五等分の花嫁 | 2022年05月20日 | 日本 | 4.2 |
| 14 | 劇場版 おいしい給食 卒業 | 2022年05月13日 | 日本 | 4.2 |
| 17 | 銀河英雄伝説 Die Neue These 激突 第三章 | 2022年05月13日 | 日本 | 4.2 |
| 56 | 機界戦隊ゼンカイジャーVSキラメイジャーVSセンパイジャー | 2022年04月29日 | 日本 | 4.2 |
| 65 | シネマ歌舞伎 桜姫東文章 下の巻 | 2022年04月29日 | 日本 | 4.2 |
| 122 | 流浪の月 | 2022年05月13日 | 日本 | 4.1 |
| 134 | RE:cycle of the PENGUINDRUM 前編 君の列車は生存戦略 | 2022年04月29日 | 日本 | 4.1 |
| 233 | 杜人（もりびと） 環境再生医 矢野智徳の挑戦 | 2022年04月15日 | 日本 | 4.1 |
| 235 | シネマ歌舞伎 桜姫東文章 上の巻 | 2022年04月08日 | 日本 | 4.1 |
| 316 | ハケンアニメ! | 2022年05月20日 | 日本 | 4.0 |
| 408 | 名探偵コナン ハロウィンの花嫁 | 2022年04月15日 | 日本 | 4.0 |
| 442 | マイスマーランド | 2022年05月06日 | 日本 | 4.0 |
| 452 | 銀河英雄伝説 Die Neue These 激突 第二章 | 2022年04月01日 | 日本 | 4.0 |

図2 2022年5月22日22:30時点で、日本で上映中の462本の映画の名・上映日・製作国・評価点数をExcelにまとめた。
フィルターをかけ、「上映日」の期間は「2022年04月～05月」、「製作国」は「日本」、「評価点数」は「4.0以上」の映画を絞り込んだ。

成果

今回の結果を判断基準の一つとして、「ハケンアニメ!」と「名探偵コナン ハロウィンの花嫁」のチケットを購入した。

2 タイタニック号沈没事故の生存者の傾向

概要

タイタニック号沈没事故の当事者の個人情報データを加工し、性別・年齢・船室等級などの情報を組み合わせて可視化することによって、生存者の傾向を分析した。

使用技術

Windows / Python3, seaborn, numpy, pandas, matplotlib, scipyなど

仮説：1. 女性・子供の生存率が高い。若年男性より、優先的に救命ボートに乗船されたからだ。
2. 一等船室（3. 乗船料金が高い）乗客の生存率が高い。人数が少ないので、客室係より直接的な支援を受けたからだ。
「4. 乗船港」または「5. 親戚の数」は生存率と関連性がある。

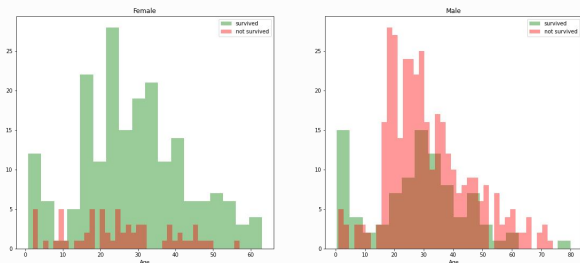


図1 性別・年齢別の生存数のグラフ。左側は女性、右側は男性。緑色は生存者、赤色は死亡者。

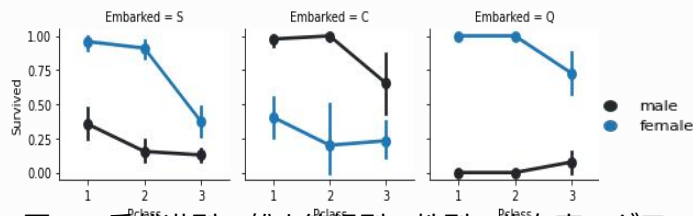


図4 乗船港別・船室等級別・性別の生存率のグラフ。黒色は男性、ブルーは女性。

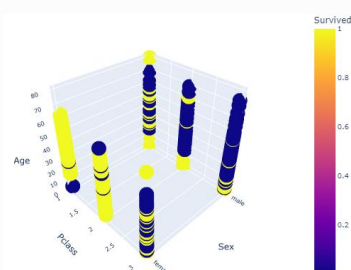


図2 性別・船室等級別・年齢別の生存数のグラフ。黄色は生存者、ブルーは死亡者。

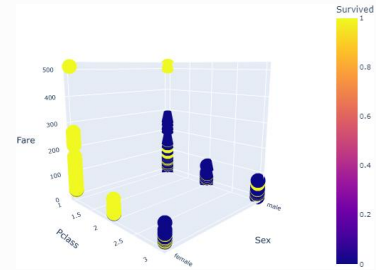


図3 性別・船室等級別・乗船料金別の生存数のグラフ。黄色は生存者、ブルーは死亡者。

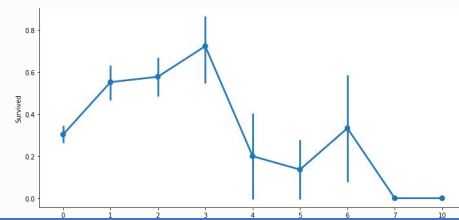


図5 親戚（兄弟や配偶者・親や子供）と同乗する人の生存率のグラフ。

成果

生存者の傾向は判明できる。1. 女性・子供の生存率は若年男性より高い。2. 一等船室乗客の生存率は二等と三等船室乗客より高い。3. 乗船料金が高ければ、乗客の生存率が高くなる。4. 港Qと港Sで乗船する女性の生存率は男性より高いが、港Cでは逆になる。5. 同乗する親戚の数は3人まで生存率は高くなるが、4人以上になると、低くなる。

3 お弁当の需要予測

概要

お弁当の販売情報から曜日やメニュー等の複数の変数を利用し、線形回帰モデルやアンサンブル学習を適用し、販売されているお弁当の販売数を予測するモデルを作成した。

使用技術

Windows / Python3, seaborn, numpy, pandas, sklearnなど

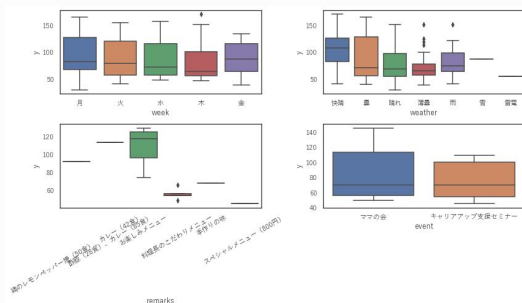


図1 目的変数（販売数）と説明変数（曜日、天気、メインメニューの名前、イベント）の関係を箱ひげ図で確認

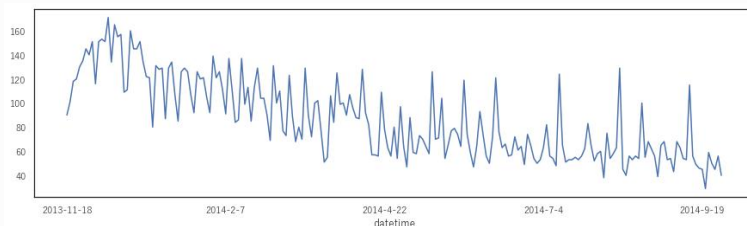


図4 testデータの予測のグラフ

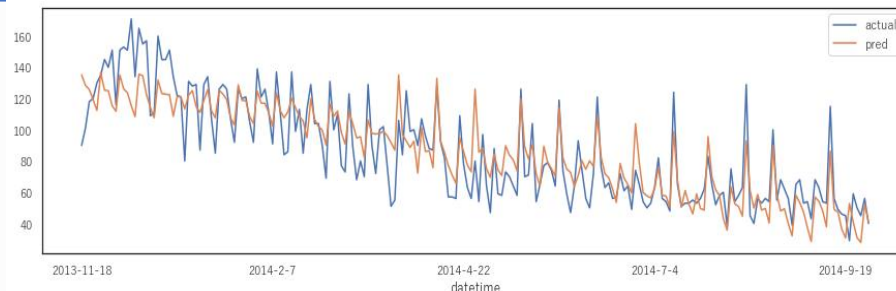


図2 線形回帰の予測値と実数値のグラフ

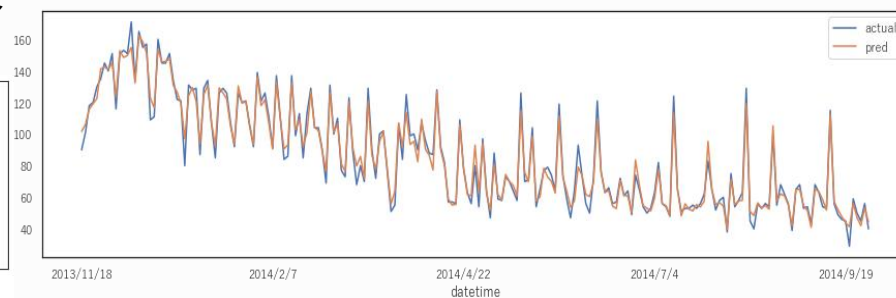


図3 アンサンブル学習後の予測値と実数値のグラフ

成果

お弁当の販売数を予測するモデルを作成した。コンペに参加して、予測値のRMSEは9.3794178となった。

4 毒キノコの分類

概要

キノコの特徴データから重要である特徴値だけを利用し、決定木を適用し、毒キノコの分類を行った。

使用技術

Windows / Python3, numpy, pandas, sklearn, matplotlibなど

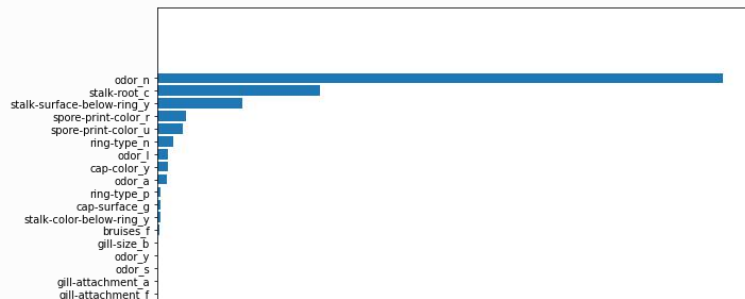


図1 第一回の決定木の学習後、特徴値重要度の高い順にプロットする

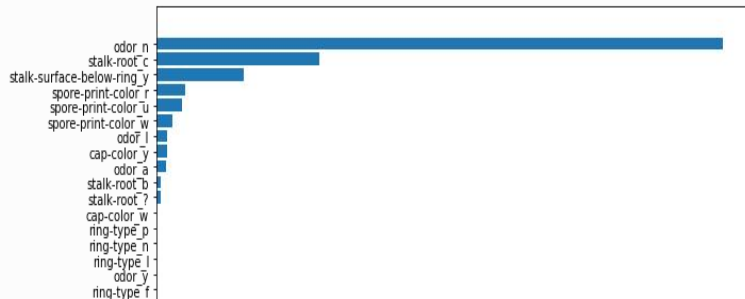


図2 前回の重要度が高い特徴値だけを残して、第二回の決定木の学習後、特徴値重要度の高い順にプロット

| | A | B |
|----|----|---|
| 1 | 1 | p |
| 2 | 4 | e |
| 3 | 6 | e |
| 4 | 8 | p |
| 5 | 9 | p |
| 6 | 12 | e |
| 7 | 14 | p |
| 8 | 15 | e |
| 9 | 17 | e |
| 10 | 18 | e |
| 11 | 20 | p |
| 12 | 22 | p |
| 13 | 23 | e |
| 14 | 29 | p |
| 15 | 30 | p |
| 16 | 31 | p |
| 17 | 32 | e |
| 18 | 33 | e |
| 19 | 34 | e |
| 20 | 36 | e |
| 21 | 37 | p |
| 22 | 38 | e |
| 23 | 39 | e |
| 24 | 42 | p |
| 25 | 44 | e |
| 26 | 45 | p |
| 27 | 48 | e |
| 28 | 49 | e |
| 29 | 50 | p |
| 30 | 51 | p |
| 31 | 57 | e |
| 32 | 59 | p |
| 33 | 64 | e |

図3 テストデータの予測を行う

成果

毒キノコを分類するモデルを作成した。コンペに参加して、予測値の暫定評価は1.0000000となった。

5 エンジニア職年収データの収集

概要

Webサイト「indeed」でスクレイピングし、5つのエンジニア職（データエンジニア、データアナリスト、データサイエンティスト、機械学習エンジニア、データアーキテクト）の年収データを収集。得られたデータのヒストグラムを作成した。

使用技術

Windows / Python, pandas, numpy, streamlit, seleniumなど



図1 indeedの求人検索画面

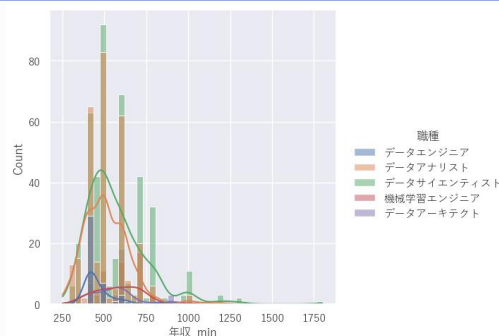


図3 職種別最低年収

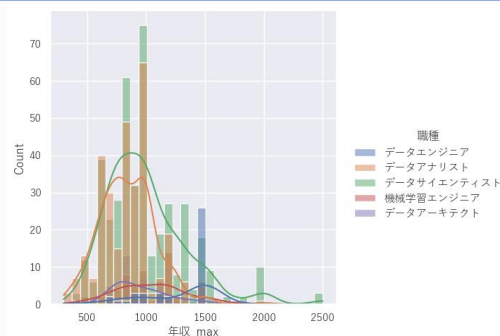


図4 職種別最高年収



図2 indeedの求人検索結果画面
(2022年12月4日)

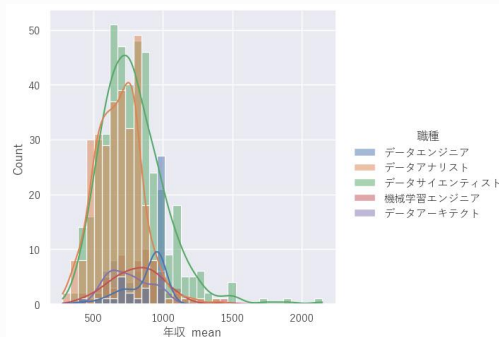


図4 職種別平均年収

| 職種 | 年収_mean | | | | | | | |
|-------------|---------|------------|------------|-------|-------|--------|-------|--------|
| | count | mean | std | min | 25% | 50% | 75% | max |
| データアナリスト | 304.0 | 681.268092 | 166.347452 | 276.0 | 550.0 | 700.00 | 800.0 | 1400.0 |
| データアーキテクト | 45.0 | 753.511111 | 143.833177 | 525.0 | 625.0 | 750.00 | 850.0 | 1000.0 |
| データエンジニア | 42.0 | 870.940476 | 141.950232 | 415.0 | 762.5 | 960.00 | 960.0 | 975.0 |
| データサイエンティスト | 423.0 | 783.401891 | 231.086819 | 350.0 | 625.0 | 750.00 | 900.0 | 2149.5 |
| 機械学習エンジニア | 60.0 | 813.558333 | 189.756648 | 415.0 | 700.0 | 812.25 | 950.0 | 1424.5 |

図4 職種別平均年収概要

成果

全体的に見ると、データサイエンティストの求人数は最も多い。機械学習エンジニアの最低年収の最頻値は最も高い。データエンジニアの最高年収と平均年収の最頻値が一位となる。職種別平均年収の平均値を求めると、データエンジニア求人別の平均年収の平均値は871万円であることは分かる。

6 製品の在庫管理API

概要

製品の在庫管理の FastAPI を作成した。製品名、製品価格といった情報を登録・参照・更新・削除できた。操作に不備があった場合、エラーコードを表示した。この FastAPI を AWS にデプロイした。

使用技術

fastAPI : Windows / Python3, fastapi, uvicorn, sqlalchemy など
 検証 : insomnia, AWS

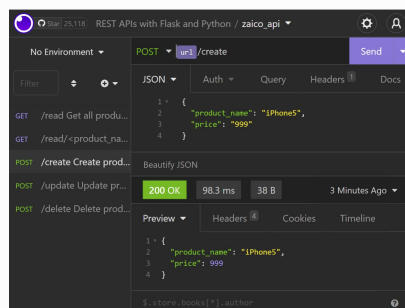


図1 データを登録する。

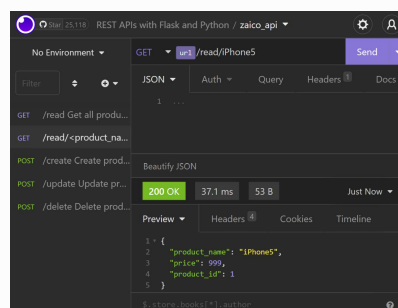


図2 登録したデータを読み込む。

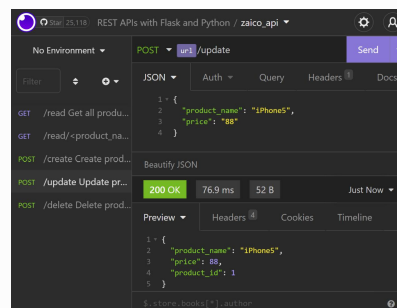


図3 登録したデータを更新する。

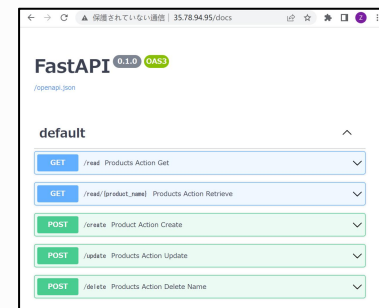


図6 FastAPIをAWSにデプロイした。
<http://35.78.94.95/docs>

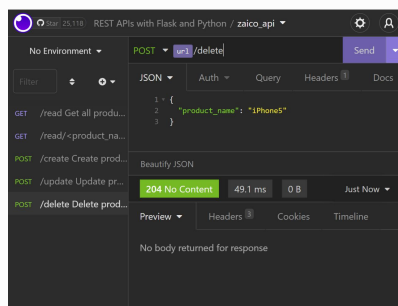


図4 登録したデータを削除する。

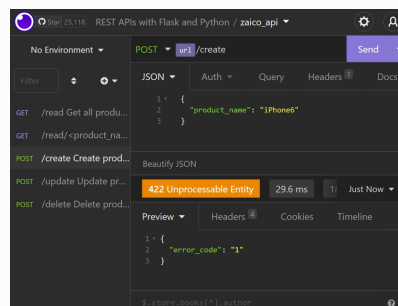


図5

データを登録・更新する場合

- ・受け取った JSON の項目が、product_name, price が揃っていない場合、{"error_code": "1"}を返す
- ・price の値が数字ではなかった場合、{"error_code": "2"}を返す
- ・price の値がマイナスだった場合、{"error_code": "3"}を返す

データを削除する場合。

- ・JSON の product_name のデータがデータベースに無い場合、{"error_code": "4"}を返す

成果

製品の在庫管理APIを完成した。製品情報の登録・参照・更新・削除やエラーコードの表示などができた。さらに、AWSにデプロイした。

7 LSGANによるウサギ画像の生成

概要

LSGAN(Least Square Generative Adversarial Networks)をPyTorchで実装し、ImageNetから収集したウサギ画像を訓練データとして学習し、それらの画像データと似たような新しい画像データを生成した。

使用技術

Windows / Python (Google Colab), numpy, matplotlib, PyTorchなど

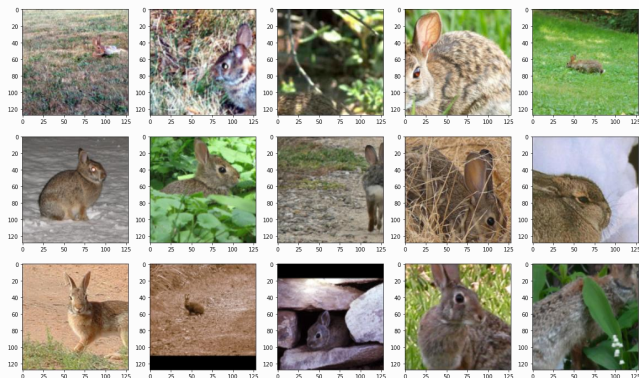


図1 ウサギ画像データ

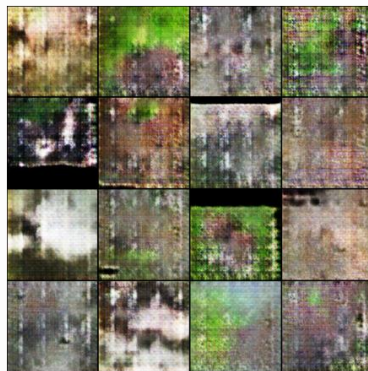


図2 Epochs: 400まで学習して、生成した画像。



図4 Epochs: 1180まで学習して、生成した画像。

成果

GPUの処理速度やコンピューティングユニットの制限によって、Epochs: 1180までしか学習していなかった。まだはっきり認識できないけど、ウサギらしい画像が出できた。

8 鲁迅チャットボット

概要

魯迅の小説を学習データに使い、文章を予測できるようにSeq2Seqのモデルを訓練し、魯迅風の返事や対話を生成した。魯迅チャットボットを作った。

使用技術

Windows / Python (Google Colab), numpy, keras, pykakasi, pickleなど



図1 漢字をひらがなに変換する。
テキストデータをまとめる。

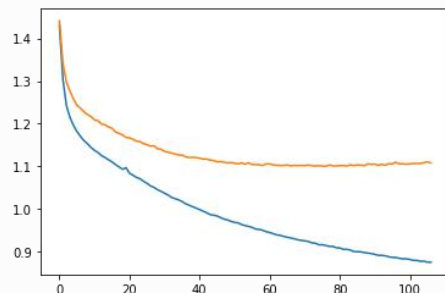


図2 構築したSeq2Seqのモデルを使って、学習を行って。誤差の推移を確認。

Input: ジョバンニさん。
Response: あなたはこんどはおおきなこえでした。

Input: あなたはわかつているのでしょうか。
Response: ジョバンニは、 なにかおもいました。

Input: ジョバンニはせいよくたちあがりましたが、 たつてみるともうは
Response: そしてそのときは、 おあいひかりをして、 そのときあたまをま

Input: ザネリがまえのせきからふりかえって、 ジョバンニをみてくすつ
Response: ジョバンニは、 もうそのときあるいていきました。

Input: ジョバンニはもうどぎまぎしてまっあかになつてしまいました。
Response: そしてそのまえにはつていっているのだした。

図3 訓練データの文を使って、返ってくる返答を確認。

魯迅A: かれはこういっていると、かれはこのときこんなにくついていたが、それはいったいのかんじょうのなかになをしをしていた。

魯迅B: そこでかれはこのときこういうにくをかんじた。

魯迅A: かれらはかれにちかえてきた。

魯迅B: そうしてもういちどたっているのです、それはいったいにくからあQのまえにとびだしていた。

図4 モデル同士を会話させる。

魯迅bot: こんにちは、屈さん。
屈: こんにちは。
魯迅bot: やまのなかにはおおぜいのにんがあった。
屈: なにたべましたか。
魯迅bot: おまえはほんとうにかんじて、あQのまえにとりやくはない。
屈: こんばんは。
魯迅bot: やくにんのたいからあQのおもいでした。
屈: さようなら。
魯迅bot: おまえはおまゑとおるはねえ。

図4 魯迅チャットボットと会話する。

成果

モデル同士の会話において、返答が似たような文章になる傾向があった。魯迅ボットと会話して、正確な返答が来ていないけど、“会話をしている”という感じがかった。