言語解析レポート

氏名: 赤松 佑哉 (akamatsu, YUYA)
学生番号: 09B23595
E-mail: p5785hcf@s.okayama-u.ac.jp

提出日: 2025 年 7 月 16 日
締切日: 2025 年 7 月 16 日

1 概要

本課題では、日本の三人の作家（芥川龍之介、江戸川乱歩、森鴎外）による文章を対象に、著者推定を行う分類モデルの構築を試みた。与えられた文を Bag-of-Words（BoW）によってベクトル化し、サポートベクターマシン（SVM）を用いて分類を行った。

2 文の特徴抽出

本課題では、著者識別のために各文を数値ベクトルに変換する必要がある。そこで、文の特徴抽出には Bag-of-Words（BoW）モデルを採用している。BoW は、文中に含まれる単語の出現情報をもとに、固定次元のベクトルとして文を表現する手法である。次に、学習データとテストデータの全体から語彙を抽出し、各単語に一意の ID を割り当てた。この語彙集合に基づいて、各文を語彙数分の次元を持つベクトルに変換した。変換後のファイルを head コマンドで前から 10 行だけ参照すると

```
~# !head train.feature
1 2905:1 1235:1 1151:1 2723:1 2006:1 1132:1 310:1 2197:1 1103:1 2020:1 841:1（以下略
0 2036:1 1450:1 2365:1 683:1 2372:1 2222:1 2473:1 1151:1
2 1554:1 2838:1 2417:1 2001:1 841:1 2099:1 464:1 2372:1 1151:1 2237:1 1103:1（以下略
2 407:1 310:1 1977:1 831:1 570:1 1504:1 1754:1 1504:1 2498:1 841:1 1151:1（以下略
1 673:1 310:1 333:1 2372:1 2903:1 310:1 2276:1 2417:1 757:1 1853:1 1710:1 （以下略
2 2854:1 2324:1
2 2905:1 1235:1 2426:1 58:1 1151:1 1697:1 2372:1 819:1 310:1 1483:1 2006:1（以下略
0 996:1 397:1 1860:1 310:1 2366:1 2372:1 1235:1 1151:1 1242:1 58:1 3001:1 （以下略
0 1574:1 92:1 1235:1 879:1 1217:1 1151:1 1002:1 310:1 89:1 1217:1 1151:1 （以下略
1 2036:1 2303:1 1151:1 2382:1 58:1 872:1 2538:1 2380:1 753:1 310:1 1294:1（以下略
```

今回の出力結果を見てわかる通り”: 1”である結果しか出てきていない．これは、Bag-of-Words（BoW）モデルで「単語の出現有無のみ」を特徴量として使っているからだ。

3　SVM による学習

サポートベクターマシン（Support Vector Machine, SVM）を用いた分類モデルの学習を行った。SVM は、高次元空間においてクラス間の分離境界を最大化する分類器であり、文の語彙分布に基づく特徴量との相性が良いとされる。学習には、Bag-of-Words（BoW）によりベクトル化された文データを使用した。各文は、語彙集合に基づくバイナリベクトル（単語の出現有無を表す）として表現されている。学習データは ‘train.feature‘、正解ラベルは ‘train.csv‘ から取得した。分類器のプログラムファイルをみると、scikit-learn ライブラリの ‘SVC‘ クラスを用いており，学習は Google Colab 環境上で実施した.
実行結果は次のようになった.

```
num_axis= 3010
shape (175, 3010)
correct, estimated
0   0   a, 婆さんはどこからとり出したか、眼をつぶった妙子の顔の先へ、一挺のナイフを突きつけました。
1   1   e, 私は仕方がないので母親に貰ったお小遣いをふんぱつして、人力車に乗りました。
2   2   m, 小川君は好奇心が起って溜まらなくなった。
0   0   a, 一そ警察へ訴えようか？

(途中略)

0   0   a, イツモダト私ハ知ラズ知ラズ、気ガ遠クナッテシマウノデスガ、今夜ハソウナラナイ内ニ、ワザト魔法ニカカッタ真似ヲシマス。
0   0   a,「折角御嬢さんの在りかをつきとめながら、とり戻すことが出来ないのは残念だな。
2   2   m, それに肌が好くって。」
0   0   a,「この近所にいらっしゃりはしないか？
acccuracy= 0.903
```

num_axis は単語の次元数であり shape は行数，最後に精度の結果が出力されている．90% 以上の精度となっている.

4　考察

著者識別を目的として SVM による分類モデルを構築し、テストデータ 175 文に対して約 90% の分類精度を達成した。これは一見すると非常に高い精度であるが、いくつかの要因がこの結果に影響していると考えられる。

まず、テストデータにおける文のサンプル分布には偏りが見られた。芥川龍之介が 61 文、江戸川乱歩が 60 文、森鴎外が 54 文と、文数に若干の差がある。特に芥川の文が最も多く、分類器がこのクラスに対して学習しやすい状況が生まれていた可能性がある。

サンプル数が 175 文と少ないにもかかわらず高精度が得られた理由として、著者ごとの文体や語彙の違いが明確であった点が挙げられる。三人の著者はそれぞれ，違う世界観で物語を構成しており，それぞれの世界観に合った単語がたくさん出てくる傾向があるだろう．例えば．芥川龍之介は少しファンタジーな世界観で物語を書いているので，幻想的な言葉がよく出てくる．これらの語彙は Bag-of-Words による特

徴抽出でも十分に識別可能であり、SVM による線形分離に適していた。

さらに、著者ごとの口調の違いも分類に寄与したと考えられる。芥川の文は文語的で硬質な表現が多く、江戸川は会話体が多く、森鴎外は地の文が長い。これらの口調の違いは、語彙だけでなく文の構造にも影響を与え、分類器が文体の傾向を捉える手がかりとなったのだろう。

以上のように、分類精度の高さは、文の分布や文体の違い、語彙の偏りなど複数の要因が重なった結果であり、必ずしもモデルの汎化性能が高いとは限らない。

5　工夫した点（AI 使用）

本課題では、著者識別の分類精度を向上させるために、いくつかの工夫を施した。特に、学習データにおけるクラス間のバランスを意識し、分類器が特定の著者に偏らないよう配慮した。学習を行う physon ファイルを copilot に聞きながら各文に重みをつけて学習させる方法で学び適応した．まず、SVM の学習において‘`class_weight='balanced'`‘ を指定することで、各クラスの文数に応じて自動的に重みを調整した。これにより、文数が少ない著者（森鴎外など）に対しても分類器が適切な境界を学習できるようにした。実際、テストデータにおいては芥川龍之介が 61 文、江戸川乱歩が 60 文、森鴎外が 54 文と、若干の偏りが見られたが、重み調整によって分類性能の公平性を保つことができた。また、特徴量の設計においては、Bag-of-Words モデルを採用し、文を語彙の出現有無に基づくバイナリベクトルとして表現した。この設計により、文の長さや頻度のばらつきによる影響を抑え、著者ごとの語彙傾向を安定的に捉えることが可能となった。結果として、分類精度に大きな変化は見られなかったが、90.1% に少し減少した．この理由としては，少数派にも注意を向けるため、多数派の分類が少し犠牲になってしまうことが要因の一つだろう．また，balanced にすると少数派の誤分類を減らす代わりに、多数派の誤分類が増えることがある．以上のことから，すこし工夫して学習させてみたが一長一短の改良で大きな精度向上は見られなかった．