卒業制作最終発表

IE4A　　村上　駿

センチメント分析

センチメント分析とは？

➡ 感情がポジティブorネガティブ判別する

➡ 教師あり学習のクラス分類に属する

教師あり学習とは？

学習データに“答え”がついているもの

学習手順

Step1.　データセットをクレンジング

Step2.　単語に分割

Step3.　単語を特徴ベクトルに変換

Step4.　分類器をトレーニング

Step1.

・句読点や’？’などの非英字文字を削除する

正規表現ライブラリの　re　を使用して
データセットをクレンジング！

学習手順

Step1.　データセットをクレンジング

Step2.　単語に分割

Step3.　単語を特徴ベクトルに変換

Step4.　分類器をトレーニング

Step2.

頻出単語（ストップワード）を除去　Ex.　is, are, and,　etc…

NLTKライブラリで提供される英語のストップワードを除去する

スペースで区切り単語に分割

学習手順

Step1.　データセットをクレンジング

Step2.　単語に分割

Step3.　単語を特徴ベクトルに変換

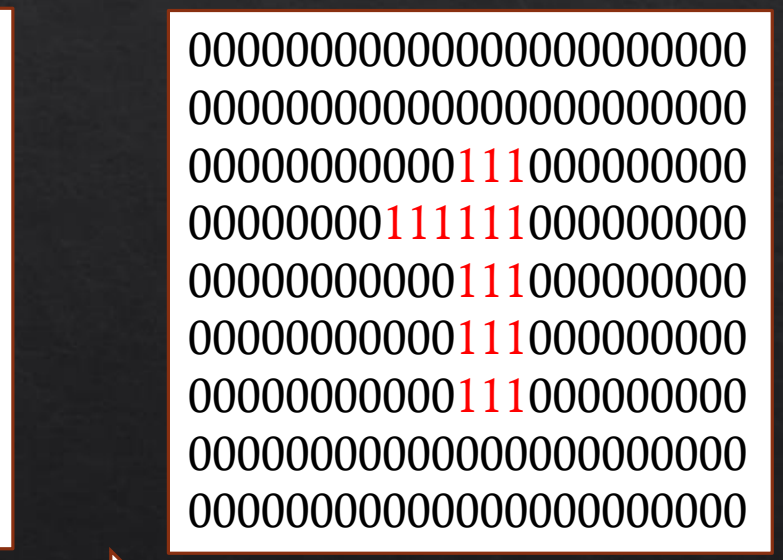
Step4.　分類器をトレーニング

Step3.

特徴ベクトルとは？

コンピュータは人間が使う言葉（自然言語）や画像などのような
数値化されていない情報（非構造化データ）を扱うのには適していない！

画像



数値

役立つ情報をベクトル形式で表すことを特徴ベクトルという

学習手順

Step1.　データセットをクレンジング

Step2.　単語に分割

Step3.　単語を特徴ベクトルに変換

Step4.　分類器をトレーニング

Step4.

45,000個の文章でトレーニング

↓

5,000個の文章でテスト

↓

```
Accuracy: 0.868
```

正解率
86%

まとめ

- 簡易な文章は判別できる
- データセットの入手が難しい
  日本語などは骨が折れそう
- 教師なし学習の回帰分類も挑戦
  したが断念