7.5 tf-idfを用いたデータのスケール変換

15t4034s 荘司響之介

tf-idf(term frequency-inverse document frequency)とは

　特徴量がどの程度情報を持っていそうかに応じて、特徴量のスケールを変換する手法の一つ。

　特定の文書にだけ頻出する単語には大きな重みを与え、コーパス上の多数の文書に現れる単語にはあまり重みを与えない。

　つまり、特定の文書にだけ頻出し、他の文書にはあまり現れない単語はその文書の内容をよく示しているのではないか、という発想に基づいている。

scikit-learnはtf-idfを、TfidfTransformerとTfidfVectorizerという２つのクラスで実装しているが、どちらのクラスにおいても、文書dにおける、単語wのtf-idfスコアは以下のように与えられる。

$$tfidf(w,d) = tf\left(\log\left(\frac{N+1}{N_w+1}\right)+1\right)$$

tf：文書d中に単語wが現れる回数

N：訓練セット中の文書の数

N_w：訓練セット中の単語wが現れる文書の数

２つのくらすはいづれも、tf-idf表現を計算した後でL2正規化を行い、文書の長さがベクトル表現に影響を与えなくする。

tf-idfは訓練データの統計的性質を利用するので、6章で述べたようにパイプラインを用いてグリッドサーチの結果が有効になるようにする。コードと出力は次のようになる。

In[23]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import make_pipeline
pipe = make_pipeline(TfidfVectorizer(min_df=5, norm=None),
                     LogisticRegression())
param_grid = {'logisticregression__C': [0.001, 0.01, 0.1, 1, 10]}

grid = GridSearchCV(pipe, param_grid, cv=5)
grid.fit(text_train, y_train)
print("Best cross-validation score: {:.2f}".format(grid.best_score_))
```

Out[23]:

```
Best cross-validation score: 0.89
```

最良の交差検証スコア

tf-idfの低い特徴量と高い特徴量を取り出すと次のようになる。

```
Features with lowest tfidf:   tfidfの低い特徴量
['poignant' 'disagree' 'instantly' 'importantly' 'lacked' 'occurred'
 'currently' 'altogether' 'nearby' 'undoubtedly' 'directs' 'fond' 'stinker'
 'avoided' 'emphasis' 'commented' 'disappoint' 'realizing' 'downhill'
 'inane']
```

多くの文書に共通して出現するか、あまり出現しないか、もしくは非常に長い文書にしか出現しない

```
Features with highest tfidf:  tfidfの高い特徴量
['coop' 'homer' 'dillinger' 'hackenstein' 'gadget' 'taker' 'macarthur'
 'vargas' 'jesse' 'basket' 'dominick' 'the' 'victor' 'bridget' 'victoria'
 'khouri' 'zizek' 'rob' 'timon' 'titanic']
```

‘dillinger’（人名）、’titanic’（映画の名前）

Idfが小さい単語（高い頻度で出現する単語）を取り出すと次のようになる。

```
Features with lowest idf:
['the' 'and' 'of' 'to' 'this' 'is' 'it' 'in' 'that' 'but' 'for' 'with'
 'was' 'as' 'on' 'movie' 'not' 'have' 'one' 'be' 'film' 'are' 'you' 'all'
 'at' 'an' 'by' 'so' 'from' 'like' 'who' 'they' 'there' 'if' 'his' 'out'
 'just' 'about' 'he' 'or' 'has' 'what' 'some' 'good' 'can' 'more' 'when'
 'time' 'up' 'very' 'even' 'only' 'no' 'would' 'my' 'see' 'really' 'story'
 'which' 'well' 'had' 'me' 'than' 'much' 'their' 'get' 'were' 'other'
 'been' 'do' 'most' 'don' 'her' 'also' 'into' 'first' 'made' 'how' 'great'
 'because' 'will' 'people' 'make' 'way' 'could' 'we' 'bad' 'after' 'any'
 'too' 'then' 'them' 'she' 'watch' 'think' 'acting' 'movies' 'seen' 'its'
 'him']
```