# 例題アプリケーション：
# 映画レビューのセンチメント分析

18SS319U
オウ　ヨウケイコ

本章では、説明のための例題としてスタンフォード大学の研究者Andrew Maasが収集したIMDb Webサイトの映画レビューデータセットを用いる。このデータセットは映画レビューのテキストと、そのレビューが「肯定的」か「否定的」か示すラベルで構成されている。IMDbWebサイトでは1から10点の採点がされている。モデリングを簡単にするために、この点が6点以上の場合に「肯定的」、それ以外の場合を「否定的」として、2クラス分類データセットとしている。

In[2]:

```
!tree -L 2 data/aclImdb
```

Out[2]:

```
data/aclImdb
├── test
│   ├── neg
│   └── pos
└── train
    ├── neg
    └── pos
6 directories, 0 files
```

このデータセットはhttp://ai.stanford.edu/~amaas/data/sentiment/から入力できる。データを解凍すると、このデータセットが二つのフォルダに格納されたテキストファイルとして提供されていることがわかる。それぞれのフォルダに、さらにposとnegというサブフォルダがある。

In[3]:

```
from sklearn.datasets import load_files

reviews_train = load_files("data/aclImdb/train/")
# load_files返回一个Bunch对象，其中包含训练文本和训练标签
text_train, y_train = reviews_train.data, reviews_train.target
print("type of text_train: {}".format(type(text_train)))
print("length of text_train: {}".format(len(text_train)))
print("text_train[1]:\n{}".format(text_train[1]))
```

Out[3]:

```
type of text_train:  <class 'list'>
length of text_train:  25000
text_train[1]:
b'Words can\'t describe how bad this movie is. I can\'t explain it by writing
only. You have too see it for yourself to get at grip of how horrible a movie
really can be. Not that I recommend you to do that. There are so many
clich\xc3\xa9s, mistakes (and all other negative things you can imagine) here
that will just make you cry. To start with the technical first, there are a
LOT of mistakes regarding the airplane. I won\'t list them here, but just
mention the coloring of the plane. They didn\'t even manage to show an
airliner in the colors of a fictional airline, but instead used a 747
painted in the original Boeing livery. Very bad. The plot is stupid and has
been done many times before, only much, much better. There are so many
ridiculous moments here that i lost count of it really early. Also, I was on
the bad guys\' side all the time in the movie, because the good guys were so
stupid. "Executive Decision" should without a doubt be you\'re choice over
this one, even the "Turbulence"-movies are better. In fact, every other
movie in the world is better than this one.'
```

In[4]:

```
text_train = [doc.replace(b"<br />", b" ") for doc in text_train]
```

レビューにはHTMLの改行シーケンス(<br/>)が含まれている場合がある。
これがあっても機械学習モデルに取っては大きな影響はないと思われるが、
先に進む前に取り除いてデータをきれいにした方がよいだろう。

In[5]:

```
print("Samples per class (training): {}".format(np.bincount(y_train)))
```

Out[5]:

```
Samples per class (training): [12500 12500]
```

In[6]:

```
reviews_test = load_files("data/aclImdb/test/")
text_test, y_test = reviews_test.data, reviews_test.target
print("Number of documents in test data: {}".format(len(text_test)))
print("Samples per class (test): {}".format(np.bincount(y_test)))
text_test = [doc.replace(b"<br />", b" ") for doc in text_test]
```

Out[6]:

```
Number of documents in test data: 25000
Samples per class (test): [12500 12500]
```