Example

October 9, 2018

1 LFD Final Project - Example code

```
In [1]: import pandas as pd
        from sklearn.pipeline import Pipeline
        from sklearn.feature_extraction.text import CountVectorizer
        from sklearn.naive_bayes import MultinomialNB
        from sklearn.metrics import accuracy_score
```

```
In [2]: train = pd.read_csv('../data/hyperp-training-grouped.csv.xz',
                            compression='xz',
                            sep='\t',
                            encoding='utf-8',
                            index_col=0).dropna()
```

```
In [3]: train.sample(3)
```

```
Out[3]:              id  hyperp         bias  \
        85886   1298911   False  left-center
        172919   732442    True         left
        179773   994298    True        right

                                                              url  labeledby  \
        85886   https://calwatchdog.com/2015/01/12/look-for-th...  publisher
        172919  https://dissentmagazine.org/article/when-g-m-w...  publisher
        179773  http://foxbusiness.com/politics/2014/08/29/mid...  publisher

                                   publisher        date  \
        85886       https://calwatchdog.com/  2018-01-20
        172919  https://dissentmagazine.org/  2018-07-16
        179773       http://foxbusiness.com/  2016-03-05

                                                            title  \
        85886            Look for the budget trailer-bill details
        172919                            When G.M. Wrecked Flint
        179773  Midwestern Manufacturing Activity Highest Sinc...

                                                             text  \
        85886   Gov. Jerry Brown?s budget proposal, released F...
        172919  Roger and Me, a radical documentary ? marketed...
        179773   \nThe pace of business activity in the U.S. M...

                                                         raw_text
        85886   b'<article id="1298911" published-at="2018-01-...
        172919  b'<article id="0732442" published-at="2018-07-...
        179773  b'<article id="0994298" published-at="2016-03-...
```

```
In [4]: pipeline = Pipeline([('vec', CountVectorizer()),
                             ('clf', MultinomialNB())])
```

```
In [5]: model = pipeline.fit(train.text, train.hyperp)
```

```
In [6]: y_pred = model.predict(train.text)
```

```
In [7]: accuracy_score(y_pred, train.hyperp)
```

```
Out[7]: 0.8464934904161594
```