Case Study 3: Textual Analysis of Movie Reviews

Team 9

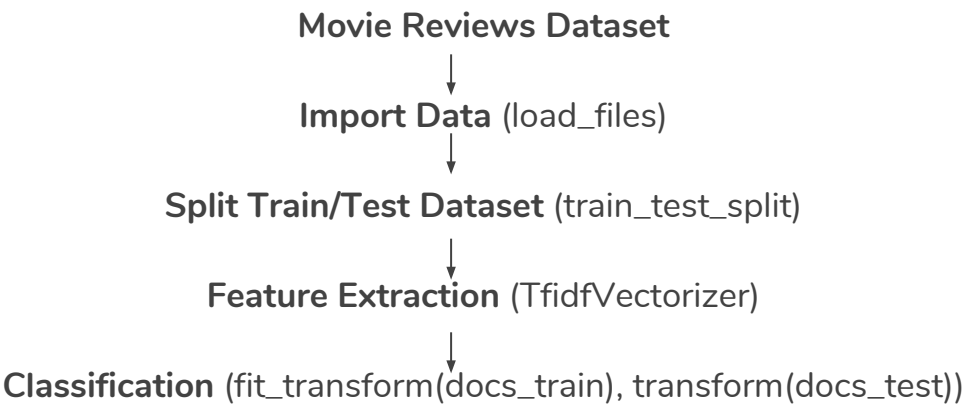
Tingting Ma Jinyan Lyu Jiani Gao
Tianhao Guo Mo Cheng

Motivation and Background

- **Textual analysis and machine learning**

Sentiment Analysis on Movie Reviews

Pipeline:



Sentiment Analysis on Movie Reviews

```
n_samples: 2000
0 params - {'vect__ngram_range': (1, 1)}; mean - 0.83; std - 0.00
1 params - {'vect__ngram_range': (1, 2)}; mean - 0.86; std - 0.01
             precision    recall  f1-score   support

        neg       0.87      0.82      0.85       256
        pos       0.82      0.87      0.85       244

avg / total       0.85      0.85      0.85       500

[[210  46]
 [ 31 213]]
```

Result:

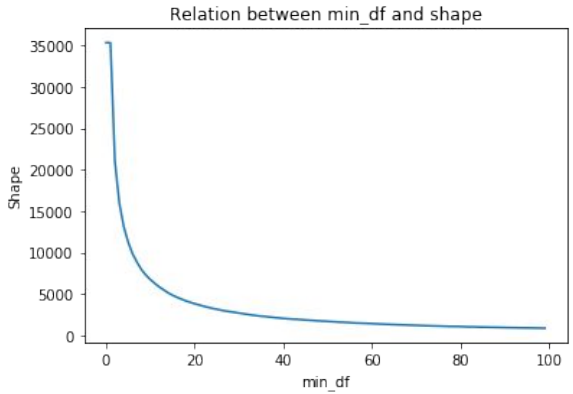
For a total number of 500 comments, 210 out of 256 negative and 213 out of 244 positive comments are correctly predicted.

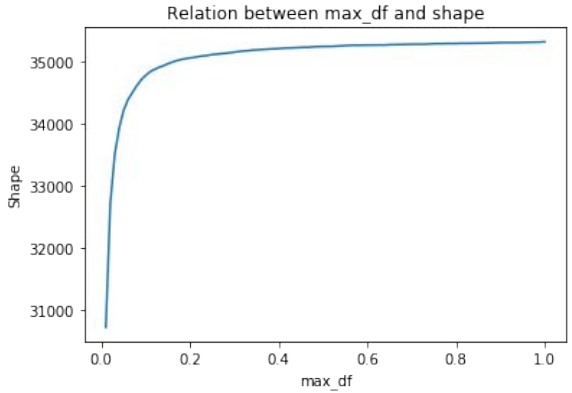
Explore TfidfVectorizer

Terminologies:

- Term Frequency (TF)
- Inverse Document Frequency (IDF)
- TF-IDF
- min_df, max_df
- n-gram, ngram_range

Screening of min_df (upper) and max_df (lower):





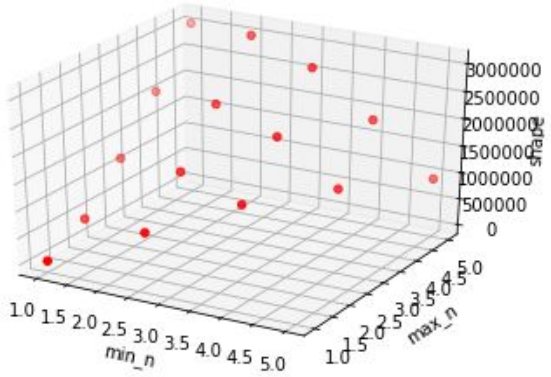
Explore TfidfVectorizer

Exploration of ngram_range:

(1,2) = (1,1) + (2,2)

435,750 = 35,433 + 400

```
(1, 1): shape is 35443;
(2, 2): shape is 400307;
(3, 3): shape is 763139;
(4, 4): shape is 895933;
(5, 5): shape is 925941;
(1, 2): shape is 435750;
(2, 3): shape is 1163446;
(3, 4): shape is 1659072;
(4, 5): shape is 1821874;
(1, 3): shape is 1198889;
(2, 4): shape is 2059379;
(3, 5): shape is 2585013;
(1, 4): shape is 2094822;
(2, 5): shape is 2985320;
(1, 5): shape is 3020763;
```



Machine Learning Algorithms

Classifiers we used: LinearSVC, K-Neighbors Classifier, Random Forest Classifier

For the three classifiers, we both choose the same set of parameters of TfidfVectorizer.

```
'vect__ngram_range': [(1,1),(2,2),(1,2),(1,3)],
'vect__max_df': [0.5, 0.75, 1.0],
'vect__min_df': [1,10,20,30],
```

Machine Learning Algorithms

LinearSVC

penalty parameter C: 1, 500 or 1000

Best score: 0.86

penalty parameter C	vect__min_df	vect__max_df	vect__ngram_range
1	10	0.5	(1, 3)
1	10	0.75	(1, 3)
500	10	0.75	(1, 3)
1000	10	0.75	(1, 3)
1000	10	1.0	(1, 3)

KNeighborsClassifier

n_neighbors parameter: 1, 10, 20

Best Score: 0.73

n_neighbors:	min_df	max_df	ngram_range
20	1	0.5	(1, 2)
20	1	0.5	(1, 3)
20	10	0.5	(1, 2)
20	10	0.5	(1, 3)
20	10	0.75	(1, 3)
20	30	0.75	(1, 3)

Machine Learning Algorithms

Prediction result of LinearSVC and K-NN

LinearSVC

	precision	recall	f1-score	support
neg	0.89	0.89	0.89	266
pos	0.88	0.88	0.88	234
avg / total	0.88	0.88	0.88	500

```
[[237  29]
 [ 29 205]]
```

K-NN

	precision	recall	f1-score	support
neg	0.80	0.64	0.71	266
pos	0.67	0.82	0.73	234
avg / total	0.74	0.72	0.72	500

```
[[170  96]
 [ 43 191]]
```

Machine Learning Algorithms

Random Forest Classifier:

best score: 0.79
parameters:

'clf__n_estimators': 64,
'vect__max_df': 0.25,
'vect__min_df': 30, '
'vect__ngram_range': (1, 3).

Prediction Result:

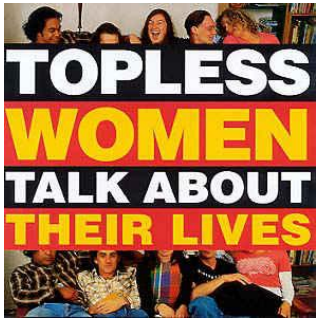
	precision	recall	f1-score	support
neg	0.72	0.88	0.79	251
pos	0.84	0.66	0.74	249
avg / total	0.78	0.77	0.77	500

```
[[220  31]
 [ 84 165]]
```

Better than K-NN, worse than LinearSVC.

Machine Learning Algorithms

Two examples where the prediction was incorrect:



1. Topless Women Talk About Their Lives

Negative but predicted as positive

Reason: Words used in this context are kind of vague.

Not too many negative words to describe.

2. The Jackal

Positive but predicted as negative

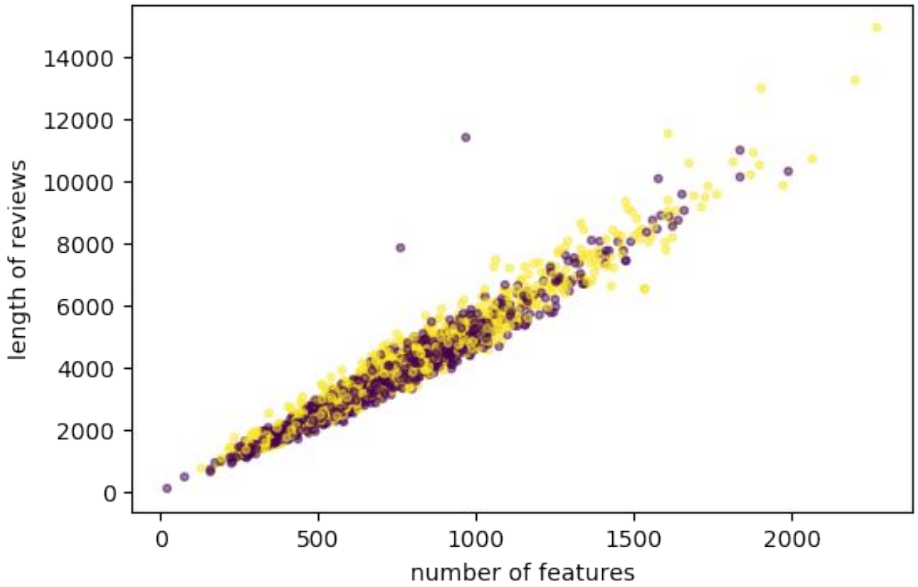
Reason: Too many negative words such as "not", "didn't" and "wasn't" when telling the story of the movie.

Open Ended Question:
Finding the right plot

#1

Professor's method:

The length of the review versus the number of features

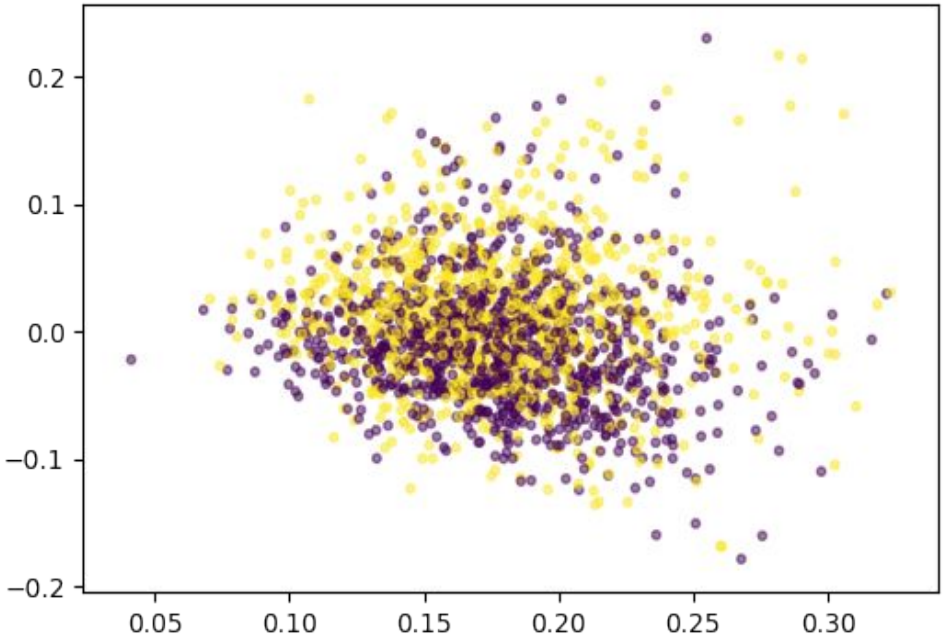


Open Ended Question:
Finding the right plot

#2

Professor's method:

Principle Component Analysis(PCA)



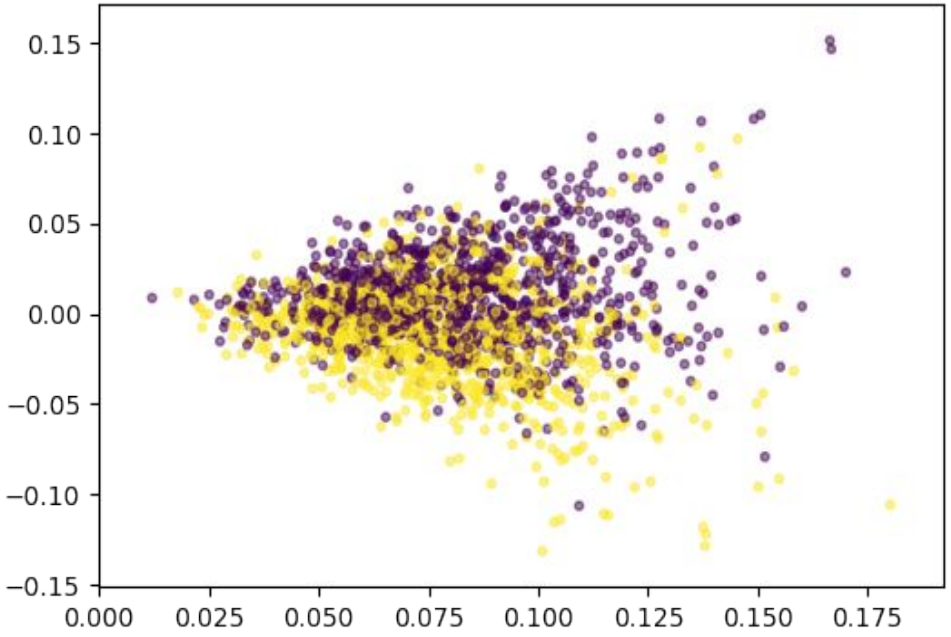
Open Ended Question:
Finding the right plot

#3

Optimized PCA

First use selectKBest, then PCA

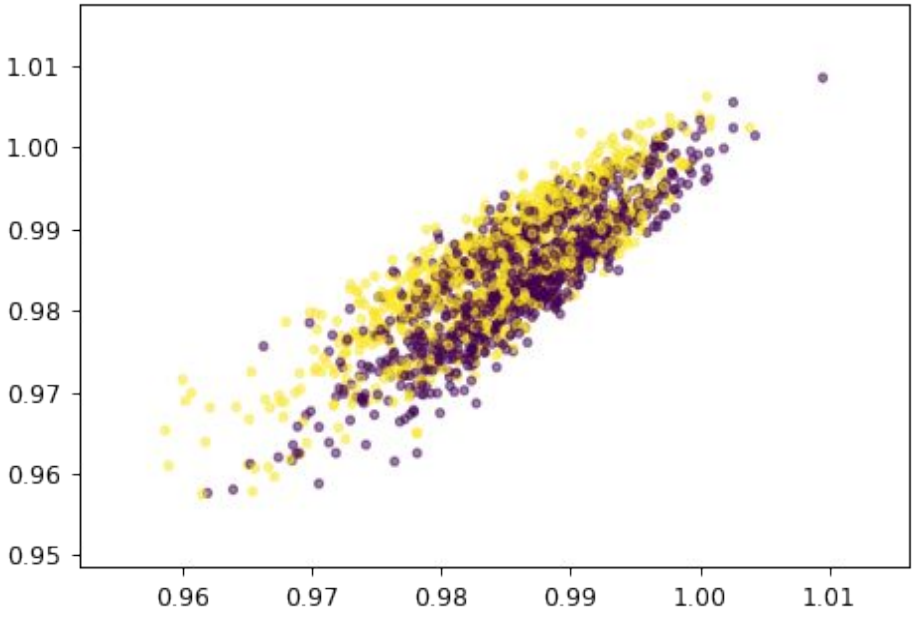
`f_classif`: ANOVA F-value between label/feature for classification tasks.



Open Ended Question:
Finding the right plot

#4

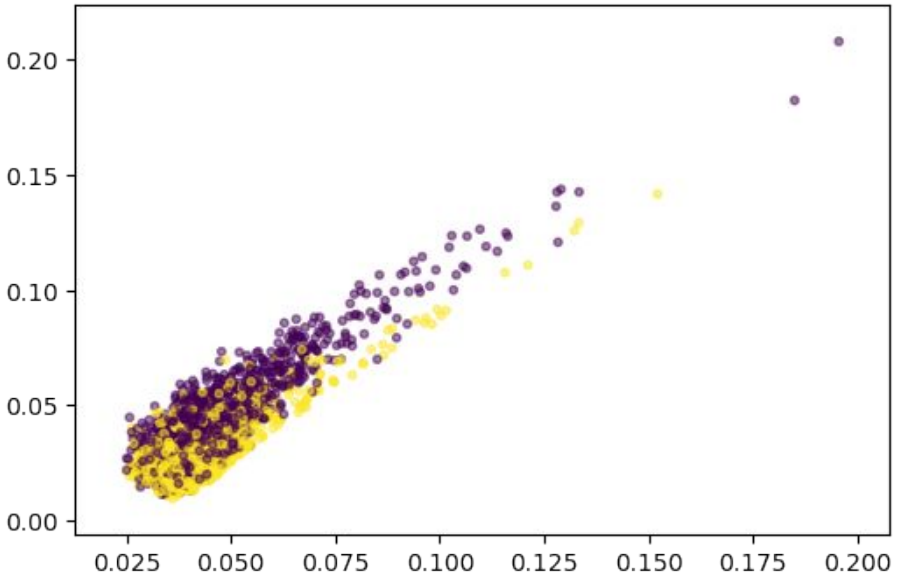
Use k-means to compute centroids, then compute distance



Open Ended Question:
Finding the right plot

#5

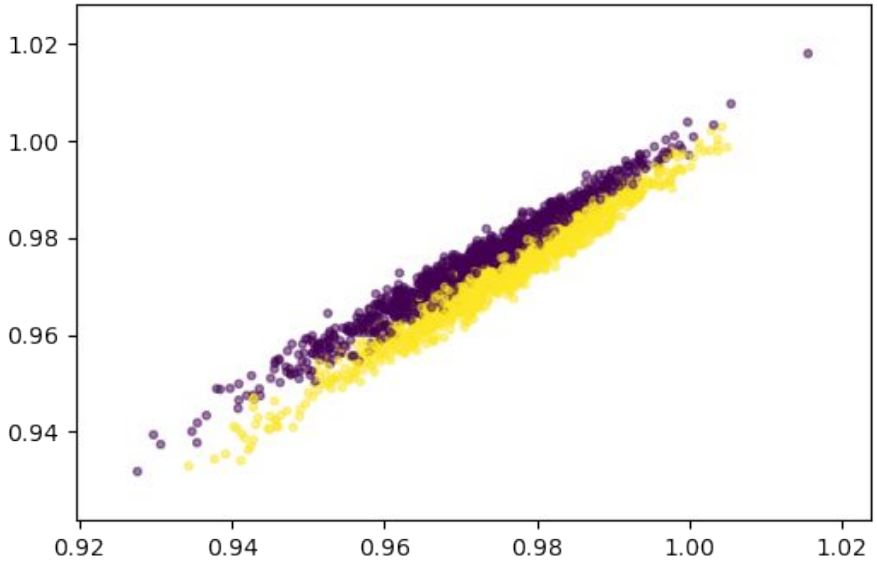
First, use selectKBest, then use k-means to compute centroids, finally compute distance



Open Ended Question:
Finding the right plot

#6

Compute centroids without k-means



Questions?