Problem 1

Given the following data:

	apple	measure	rest
Class = Physics	5	100	400
Class = -Physics	100	50	9450

$\chi^2(apple, Physics)$:

Observed counts:

	apple	rest
Class = Physics	5	500
Class = -Physics	100	9500

Estimated values:

$$\begin{aligned}
P(apple, Physics) &= \frac{105 * 505}{10105} \\
&= 5.25 \\
P(apple, -Physics) &= \frac{105 * 9600}{10105} \\
&= 99.75 \\
P(rest, Physics) &= \frac{10000 * 505}{10105} \\
&= 499.75 \\
P(rest, -Physics) &= \frac{10000 * 9600}{10105} \\
&= 9500.25 \\
\chi^2(apple, Physics) &= \sum \frac{(O - E)^2)}{E} \\
&= \frac{(5 - 5.25)^2}{5.25} + \frac{(100 - 99.75)^2}{99.75} + \frac{(500 - 499.75)^2}{499.75} + \frac{(9500 - 9500.25)^2}{9500.25} \\
&= 0.0127
\end{aligned}$$

$\chi^2(measure, Physics)$:

Observed counts:

	measure	rest
Class = Physics	100	405
Class = -Physics	50	9550

Estimated values:

$$\begin{aligned}
P(measure, Physics) &= \frac{150 * 505}{10105} \\
&= 7.50 \\
P(measure, -Physics) &= \frac{150 * 9600}{10105} \\
&= 142.50 \\
P(rest, Physics) &= \frac{9955 * 505}{10105} \\
&= 497.50 \\
P(rest, -Physics) &= \frac{9955 * 9600}{10105} \\
&= 9457.50 \\
\chi^2(measure, Physics) &= \sum \frac{(O - E)^2)}{E} \\
&= \frac{(100 - 7.50)^2}{7.50} + \frac{(405 - 497.50)^2}{497.50} + \frac{(142.50 - 50.0)^2}{142.50} + \frac{(9550 - 9457.50)^2}{9457.50} \\
&= 1218.98
\end{aligned}$$

Problem 2

- pmi(t) to discriminate well for a single category calulcated using

  $$pmi_{max}(t) = max_{1..m}pmi(t, c_i) \tag{1}$$

  10 features that is obtained doing feature selection using pmi(t):

  ```
  ['18th', 'reciprocally', 'opake', 'bh', 'endu', 'putty', 'ebook', 'excite', 'thereabouts'
  'fa']
  ```

  Dimension decreased 1337 times, because original dimension was 13377, now it is only 10 features are chosen.

- 10 features that is obtained from feature selection on Mutual Information:

  ```
  ['woody', 'unanswered', 'originality', '4no', 'cordially', 'rusty', '273', '279',
  'inanimate', 'affiliated']
  ```

  The features obtained from MI different from pmi(t). Features for MI are more less present in most of the training data sets, on other hand features of pmi(t) very specific to one class.

- The results of Naive Bayer Classification for test data sets using the features from MI and Pairwise Mutual Information:
  $Test_b$:
  - Pairwise Mutual Information(pmi): Biology
  - Mutual Information(mi): Chemistry

  $Test_b2$:
  - Pairwise Mutual Information(pmi): Biology
  - Mutual Information(mi): Chemistry

  $Test_c$:
  - Pairwise Mutual Information(pmi): Biology
  - Mutual Information(mi): Chemistry

  $Test_c2$:
  - Pairwise Mutual Information(pmi): Biology
  - Mutual Information(mi): Chemistry

  $Test_p$:
  - Pairwise Mutual Information(pmi): Biology
  - Mutual Information(mi): Chemistry

  $Test_p2$:
  - Pairwise Mutual Information(pmi): Biology
  - Mutual Information(mi): Chemistry

Instruction to run the the code (for Linux):

```
$ python prb2.py
```

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Problem 3.

a.

In the design of a text classifier, a word that frequently occurs in one or more documents but not is all of the documents can be used as a good feature.

$$idf = log_e \frac{totalnumofdocs}{numofdocsinwhichthetermoccurs}$$

Hence, this *idf* factor will have a maximum value if the term occurs in only one document, therefore such a term feature is useful in identifying the doc of that class. But, if a term is present in all the documents, the *idf* value will be zero and the term feature cannot be used for classification.

b.

The results for the Naiive Bayes classifier applied for the classification of file: `test_b.txt` are as follows:

```
log of prob.  for class:  Biology is -665827.766826
log of prob.  for class:  Chemistry is -691757.569164
log of prob.  for class:  Physics is -695191.903181
Assigned class by NB is:  Biology
```

c.

The class assigned by the knn classifier with value of k = 1 is `Physics` as it had the smallest euclidean distance to the `Physics` training point.
The results are as follows:

```
Biology.txt 0.0280358530082
Chemistry.txt 0.0267238107569
Physics.txt 0.0229581040736
Assigned class by knn is:  Physics
```

Instruction to run the the code (for Linux):

```
$ python prb3.py
```