Question 1.1

$B = \{1, 2\}\ Box$

$C = \{B, Y, R\}\ Coin\ Color$

$P(two\ heads\ in\ a\ row\ |C = B) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

$P(two\ heads\ in\ a\ row\ |C = Y) = \frac{1}{4} * \frac{1}{4} = \frac{1}{16}$

$P(two\ heads\ in\ a\ row\ |C = R) = \frac{1}{10} * \frac{1}{10} = \frac{1}{100}$

$P(two\ heads\ in\ a\ row\ |\ B = 1) = P(C = B\ |B = 1) * P(two\ heads\ in\ a\ row\ |C = B) + P(C = Y\ |B = 1) * P(two\ heads\ in\ a\ row\ |C = Y)$

$P(two\ heads\ in\ a\ row\ |\ B = 1) = \frac{2}{3} * \frac{1}{4} + \frac{1}{3} * \frac{1}{16} = \frac{9}{48} = \frac{3}{16}$

$P(two\ heads\ in\ a\ row\ |\ B = 2) = P(C = B\ |B = 1) * P(two\ heads\ in\ a\ row\ |C = B) + P(C = R\ |B = 1) * P(two\ heads\ in\ a\ row\ |C = R)$

$P(two\ heads\ in\ a\ row\ |\ B = 2) = \frac{1}{2} * \frac{1}{4} + \frac{1}{2} * \frac{1}{100} = \frac{26}{200} = \frac{13}{100}$

$P(two\ heads\ in\ a\ row\ ) = P(B = 1) * P(two\ heads\ in\ a\ row\ |\ B = 1) + P(B = 2) * P(two\ heads\ in\ a\ row\ |\ B = 2)$

$P(two\ heads\ in\ a\ row) = \frac{1}{2} * \frac{3}{16} + \frac{1}{2} * \frac{13}{100} = 0.15875$

Question 1.2

$P(C = B\ |\ two\ heads\ in\ a\ row) = \frac{P(two\ heads\ in\ a\ row\ |C = B) * P(C = B)}{P(two\ heads\ in\ a\ row)}$

$P(C = B) = \frac{1}{2} * \frac{2}{3} + \frac{1}{2} * \frac{1}{2} = \frac{7}{12}$

$P(C = B\ |\ two\ heads\ in\ a\ row) = \frac{\frac{1}{4} * \frac{7}{12}}{0.15875} = 0.9186$

Question 1.3

$P(C = R\ |\ two\ heads\ in\ a\ row) = \frac{P(two\ heads\ in\ a\ row\ |C = R) * P(C = R)}{P(two\ heads\ in\ a\ row)}$

$P(C = R) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

$P(C = B\ |\ two\ heads\ in\ a\ row) = \frac{\frac{1}{100} * \frac{1}{4}}{0.15875} = 0.01574$

Question 2.1

$$f(x;\ \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE} = argmax\ P(D\ |\mu,\sigma) \qquad P(D\ |\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln P(D\ |\mu,\sigma) = \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] = -N * \ln\sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu}\ln P(D\ |\mu,\sigma) = \frac{\partial}{\partial\mu}[-N * \ln\sigma\sqrt{2\pi} - \textstyle\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}]$$

$$\textstyle\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2} = 0 \ \rightarrow\ -N\mu + \sum_{i=1}^{N} x_i = 0 \rightarrow \mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Question 2.2

$$\lambda_{MLE} = argmax\ P(\mu|D) = argmax\ \frac{P(D\ |\mu) * P(\mu)}{P(D)}$$

$$\ln P(\mu|D) = \ln P(D\ |\mu) + \ln P(\mu)$$

$$\ln P(\lambda|D) = \ln\left[\frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right] + \ln\left[\lambda e^{-\lambda\mu}\right]$$

$$\frac{\partial}{\partial\mu}\ln P(\lambda|D) = \frac{\partial}{\partial\mu}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \ln\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) + \ln\left(\lambda e^{-\lambda\mu}\right)\right)$$

$$0 = -\frac{2(x-\mu)}{2\sigma^2} + (-\lambda)$$

$$\frac{(-x+\mu)}{\sigma^2} = -\lambda$$

$$\mu_{\ MAP} = -\lambda\sigma^2 + x$$

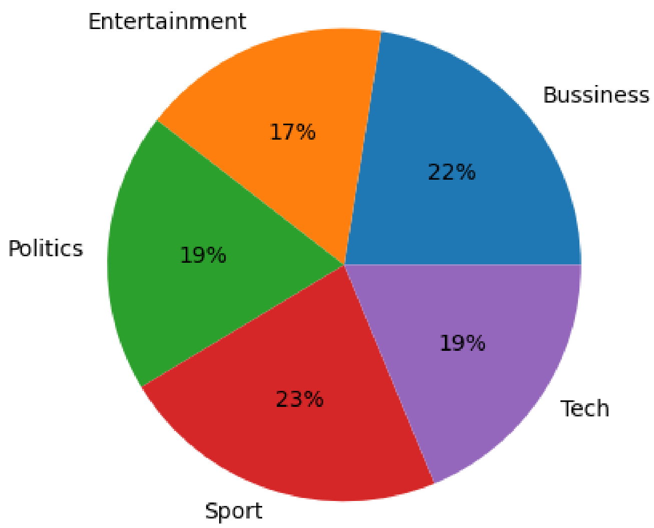
Question 2.3

$$P(x_{n+1} = 1) = N(1|\mu = 1, \sigma = 1) = \frac{1}{1\sqrt{2\pi}} * e^{-\frac{(1-1)^2}{2*1^2}} = 0.39894$$

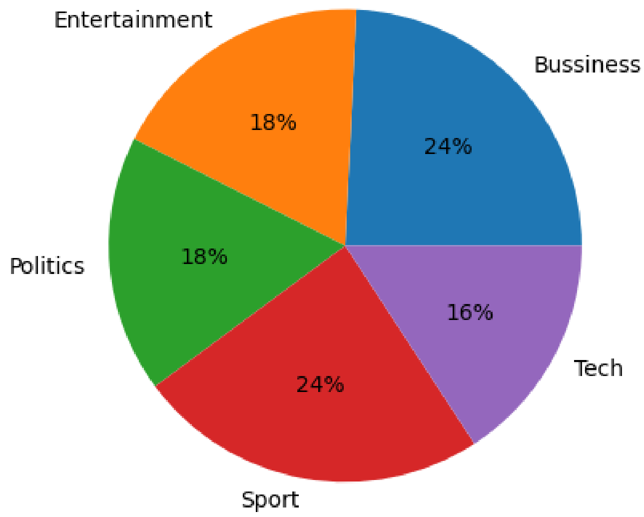
$$L(x_{n+1} = 1) = N(2|\mu = 1, \sigma = 1) = \frac{1}{1\sqrt{2\pi}} * e^{-\frac{(2-1)^2}{2*1^2}} = 0.24197$$

Question 3.1

1) Pie Chart of train data labels



Pie Chart of test data labels



2) $P(Y = Bussiness) = 0.225$
$P(Y = Entertainment) = 0.170$
$P(Y = Politics) = 0.191$
$P(Y = Sport) = 0.226$
$P(Y = Tech) = 0.188$

(rounded to 3 decimal places)

3) This training set is somewhat balanced. Having a balanced training set is important because we use prior probabilities of each class while calculating the maximum likelehood estimatior. Therefore, if the training data is specifically skewed towards one of the classes, that class will have a larger value and impact MLE more than the others.

4) $\ln(P(alien \mid Y = Tech)) = -10.171387747476452$
   $\ln(P(thunder \mid Y = Tech)) = -\infty$

Question 3.2

Results:

```
Question 3.2
Confusion matrix
[[126.   0.   3.   1.   0.]
 [  1.  92.   1.   0.   2.]
 [  5.   2.  90.   0.   0.]
 [  1.   0.   0. 133.   0.]
 [  2.   8.   4.   0.  86.]]
Accuracy: 0.946
```

Question 3.3

Results:

```
Question 3.3
Confusion matrix
[[131.   0.   1.   0.   1.]
 [  0.  97.   0.   0.   0.]
 [  2.   0.  96.   1.   0.]
 [  0.   0.   0. 133.   0.]
 [  2.   5.   1.   0.  87.]]
Accuracy: 0.977
```

Comparing results of Question 3.2 and 3.2 we can clearly see that there are significantly less mislabled documents. By assuming each word appears at least α times we eliminate the posibility of having θ equal to zero, which helps us minimize errors in our MLE. When α = 1, the method assumes that there one of each word in our Bag-of-Words representation. This way, we have the chance to eliminate probability of a word being zero without changing the prior probabilities of the words significantly.

Question 3.4

Results:

```
Question 3.4
Confusion matrix
[[132.   3.   4.   0.   4.]
 [  0.  96.   0.   0.   2.]
 [  2.   1.  94.   0.   0.]
 [  0.   0.   0. 134.   0.]
 [  1.   2.   0.   0.  82.]]
Accuracy: 0.966
```