Question 1

a) Let X, Y be two independent uni-variate random variables sampled uniformly from the unit interval $[0, 1]$. First we note that $Z = |X - Y|^2 = (X - Y)^2 = X^2 - 2XY + Y^2$. Thus it follows that $E[Z] = E[X^2 - 2XY + Y^2]$, and by linearity of expectation we get that $E[Z] = E[X^2] - 2E[XY] + E[Y^2]$. Furthermore, since X, Y are independent it follows that $E[XY] = E[X]E[Y]$. Also note that since X, Y are both samples from identical distributions, we can replace Y with X, thus arriving at $E[Z] = E[X^2] - 2E[XY] + E[Y^2] = E[X^2] - 2E[X]E[Y] + E[Y^2] = E[X^2] - 2E[X]^2 + E[X^2] = 2(E[X^2] - E[X]^2)$. Next, we observe the following result, which will speed up expectation calculations down the road:

$$E[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx = \int_0^1 x^n \cdot 1 dx = (\frac{1}{n+1} x^{n+1})\Big|_0^1 = \frac{1}{n+1} \qquad (1)$$

where $f(x)$ is the probability mass function of X. Thus, we get that $2(E[X^2] - E[X]^2) = 2(\frac{1}{3} - (\frac{1}{2})^2) = \frac{1}{6}$.

We note that $\text{Var}(Z) = E[Z^2] - E[Z]^2$. Then by plugging in $Z = (X - Y)^2$ and $E[Z] = \frac{1}{6}$ from part *a*), we get that $\text{Var}(Z) = E[(X - Y)^4] - (\frac{1}{6})^2 = E[X^4] - 4E[X^3Y] + 6E[X^2Y^2] - 4E[XY^3] + E[Y^4] - \frac{1}{36}$. Moreover, we note that since X, Y are independent it follows that any function of X and Y are also independent. Hence, X^3, Y, and X^2, Y^2, and X, Y^3 are all independent to each other. Using this fact, we get that $\text{Var}(Z) = \frac{1}{5} - 4E[X^3]E[Y] + 6E[X^2]E[Y^2] - 4E[X]E[Y^3] + E[Y^4] - \frac{1}{36} = \frac{1}{5} - 4E[X^3]E[X] + 6E[X^2]E[X^2] - 4E[X]E[X^3] + E[X^4] - \frac{1}{36} = \frac{1}{5} - 4(\frac{1}{4})(\frac{1}{2}) + 6(\frac{1}{3})(\frac{1}{3}) - 4(\frac{1}{2})(\frac{1}{4}) + \frac{1}{5} - \frac{1}{36} = \frac{7}{180}$.

b) We note that $||X - Y||_2^2 = R = Z_1 + \cdots + Z_d$, where each $Z_i = |X_i - Y_i|^2$. Then by the linearity of expectation and calculations made in part a) we get that $E\big[||X - Y||_2^2\big] = E[R] = E[Z_1 + \cdots + Z_d] = E[Z_1] + \cdots + E[Z_d] = \frac{1}{6} + \cdots + \frac{1}{6} = \frac{d}{6}$. Since each X_i, Y_i are independent, it follows that each Z_i are mutually independent. Combining independence with the result from part a), we get that $\text{Var}[R] = \text{Var}[Z_1 + \cdots + Z_d] = \text{Var}[Z_1] + \cdots + \text{Var}[Z_d] = \frac{7}{180} + \cdots + \frac{7}{180} = \frac{7d}{180}$.

c) First we calculate the maximum squared Euclidean distance between two points within the d-dimensional unit cube, D. Note that $D = \left(\sqrt{\sum_{i=1}^{d} 1}\right)^2 = (\sqrt{d})^2 = d$. The claim that most points are far away in higher dimension can be seen by looking at the mean of $||X - Y||_2^2$, $m = \frac{d}{6}$, where d is the dimension. We see that as the dimension approaches infinity, it follows that the mean distance between any two points also approaches infinity, supporting the claim that most points are far away. Another way to look at it is to compare ratio of the mean to the dimension. We see that there is a proportional relationship between them, namely $m \; \alpha \; d$. Hence as we increase in dimensions, we expect the distance between two points to proportionally increase as the mean is proportionally increasing. Now, for the claim that they are approximately the same distance,

it helps to looks at the relative ratio between the standard deviation, SD, which is $SD = \sqrt{\text{Var}} = \sqrt{\frac{7d}{180}}$, and the maximum total distance between any two points, which is d by previous calculation. We see that $\frac{SD}{D}$ approaches zero as the dimension approaches infinity, meaning the relative variation between the distance between any two points are getting relatively smaller and smaller as we increase in dimension. In this way, the mean and standard deviation of $||X - Y||_2^2$ support the claim that most points are far away and approximately the same distance in higher dimensions.

Question 2

a) $H(X)$ is non-negative because $p(x) \geq 0$ for all $x \in \mathcal{X}$ by definition of a probability mass function. Furthermore, since p is a probability mass function, we know that $p(x) \leq 1$ for every $x \in \mathcal{X}$, hence $\frac{1}{p(x)} \geq 1$ for all x, and so $\log_2(\frac{1}{p(x)}) \geq \log_2(1) = 0$ for all x, since log is an increasing function. Thus we get that $p(x)\log_2(\frac{1}{p(x)}) \geq 0$ for all $x \in \mathcal{X}$, and so $\sum_x p(x)\log_2(\frac{1}{p(x)}) \geq 0$, as wanted.

b) Let X, Y be independent discrete random variables. Let p_x, p_y be the corresponding probability mass functions and p_{xy} be the joint probability mass function.

$$
\begin{aligned}
H(X,Y) &= \sum_x \sum_y p_{xy}(x,y)\log_2(\frac{1}{p_{xy}(x,y)}) && \text{by definition} \\
&= \sum_x \sum_y p_x(x)p_y(y)\log_2(\frac{1}{p_x(x)} \cdot \frac{1}{p_y(y)}) && \text{since } X,Y \text{ are independent} \\
&= \sum_x \sum_y p_x(x)p_y(y)(\log_2(\frac{1}{p_x(x)}) + \log_2(\frac{1}{p_y(y)})) && \text{by properties of log} \\
&= \sum_x \sum_y p_x(x)p_y(y)\log_2(\frac{1}{p_x(x)}) \\
&+ p_x(x)p_y(y)\log_2(\frac{1}{p_y(y)}) \\
&= \sum_x \sum_y p_y(y)\log_2((\frac{1}{p_x(x)})^{p_x(x)}) \\
&+ p_x(x)\log_2((\frac{1}{p_y(y)})^{p_y(y)}) && \text{by properties of log} \\
&= \sum_x \sum_y p_y(y)\log_2((\frac{1}{p_x(x)})^{p_x(x)}) \\
&+ \sum_y \sum_x p_x(x)\log_2((\frac{1}{p_y(y)})^{p_y(y)}) && \text{rearranging} \\
&= \sum_x \log_2((\frac{1}{p_x(x)})^{p_x(x)}) \sum_y p_y(y) \\
&+ \sum_y \log_2((\frac{1}{p_y(y)})^{p_y(y)}) \sum_x p_x(x) && \text{by summation properties}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_x \log_2((\frac{1}{p_x(x)})^{p_x(x)})(1) \\
&+ \sum_y \log_2((\frac{1}{p_y(y)})^{p_y(y)})(1) && \text{by properties of p.m.f.} \\
&= \sum_x p_x(x)\log_2(\frac{1}{p_x(x)}) + \sum_y p_y(y)\log_2(\frac{1}{p_y(y)}) && \text{by properties of log} \\
&= H(X) + H(Y) && \text{by defintion}
\end{aligned}
$$

c) Adopting the notation from part b), with the addition that we let $p_{y|x}$ be the conditional probability distribution of $Y|X$, we get that:

$$
\begin{aligned}
H(Y|X) &= \sum_y p_{y|x}(y|x)\log_2(\frac{1}{p_{y|x}(y|x)}) \\
&= \sum_y \frac{p_{xy}(x,y)}{p_x(x)}\log_2(\frac{p_x(x)}{p_{xy}(x,y)}) && \text{by definition of conditional p.m.f.} \\
&= (\sum_x p_x(x))\sum_y \frac{p_{xy}(x,y)}{p_x(x)}\log_2(\frac{p_x(x)}{p_{xy}(x,y)})) && \text{since } \sum_x p_x(x) = 1 \\
&= \sum_x\sum_y p_{xy}(x,y)\log_2(\frac{p_x(x)}{p_{xy}(x,y)})) && \text{cancelling out } p_x(x) \\
&= \sum_x\sum_y p_{xy}(x,y)(\log_2(p_x(x)) + \log_2(\frac{1}{p_{xy}(x,y)})) && \text{by properties of log} \\
&= \sum_x\sum_y p_{xy}(x,y)\log_2(p_x(x)) \\
&+ \sum_x\sum_y p_{xy}(x,y)\log_2(\frac{1}{p_{xy}(x,y)}) \\
&= \sum_x\sum_y p_{xy}(x,y)\log_2(p_x(x)) + H(X,Y) \\
&= \sum_x \log_2(p_x(x))\sum_y p_{xy}(x,y) + H(X,Y) \\
&= \sum_x \log_2(p_x(x))p_x(x) + H(X,Y) && \text{by marginal distribution property} \\
&= -\sum_x p_x(x)\log_2(\frac{1}{p_x(x)}) + H(X,Y) && \text{by properties of log} \\
&= H(X,Y) - H(X)
\end{aligned}
$$

hence, we get that $H(X,Y) = H(X) + H(Y|X)$

d)

$$\begin{aligned}
\sum_x p(x)\log_2\frac{p(x)}{q(x)} &= \sum_x p(x)(-1)\log_2\frac{q(x)}{p(x)} \\
&= E[-\log_2\frac{p(x)}{q(x)}] \\
&\geq -\log_2(E[\frac{q(x)}{p(x)}]) \qquad \text{by Jensen's inequality since } \log \text{ is a concave function} \\
&= -\log_2(\sum_x p(x)\frac{q(x)}{p(x)}) \\
&= -\log_2(\sum_x q(x)) \\
&= -\log_2(1) = 0
\end{aligned}$$

Hence, $KL(p||q)$ is non-negative.

e)

$$\begin{aligned}
KL(p(x,y)||p(x)p(y)) &= \sum_x\sum_y p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_x\sum_y p(x,y)\log_2\frac{p(y|x)p(x)}{p(x)p(y)} \qquad \text{by properties of conditional distribution} \\
&= \sum_x\sum_y p(x,y)\log_2\frac{p(y|x)}{p(y)} \\
&= \sum_x\sum_y p(x,y)(\log_2\frac{1}{p(y)} - \log_2\frac{1}{p(y/x)}) \\
&= \sum_x\sum_y p(x,y)\log_2\frac{1}{p(y)} \\
&\quad - \sum_x\sum_y p(x,y)\log_2\frac{1}{p(y|x)} \\
&= \sum_x\sum_y p(y)p(x|y)\log_2\frac{1}{p(y)} \\
&\quad - \sum_x\sum_y p(y|x)p(x)\log_2\frac{1}{p(y|x)} \qquad \text{by properties of conditional distribution} \\
&= \sum_y p(y)\log_2\frac{1}{p(y)}\sum_x p(x|y) \\
&\quad - \sum_y p(y|x)\log_2\frac{1}{p(y|x)}\sum_x p(x)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_y p(y) \log_2 \frac{1}{p(y)}(1) && \text{since } p(x), p(x|y) \text{ are probability} \\
&- \sum_y p(y|x) \log_2 \frac{1}{p(y|x)}(1) && \text{distributions} \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

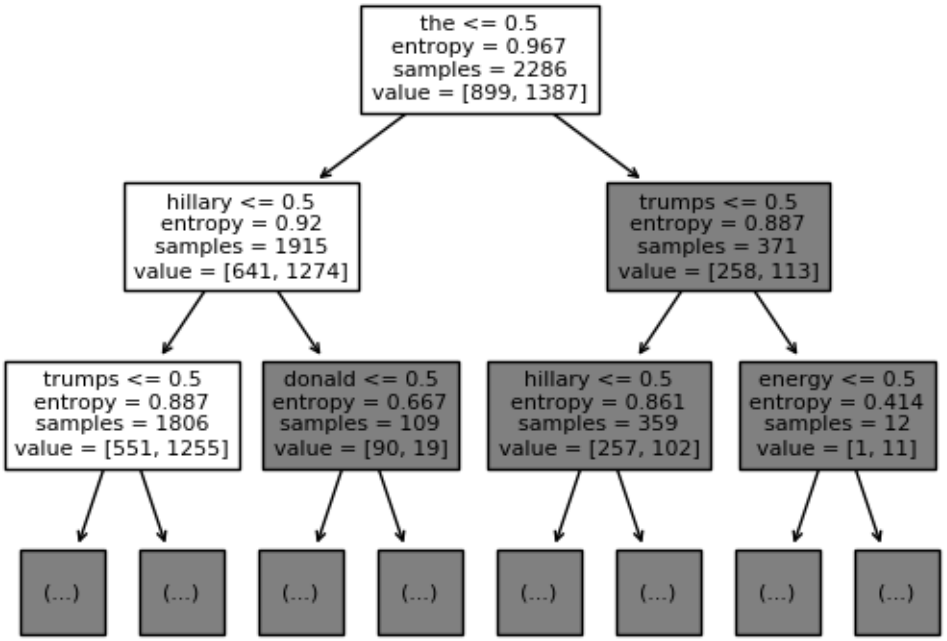
as wanted.

Question 3

a) Look at code hw1_code.py

b) The output of the function select_tree_model is the following:
gini with depth 10 has validation accuracy of 0.710204081632653
gini with depth 25 has validation accuracy of 0.7346938775510204
gini with depth 50 has validation accuracy of 0.7428571428571429
gini with depth 85 has validation accuracy of 0.7551020408163265
gini with depth 110 has validation accuracy of 0.753061224489796
entropy with depth 10 has validation accuracy of 0.7122448979591837
entropy with depth 25 has validation accuracy of 0.7326530612244898
entropy with depth 50 has validation accuracy of 0.7408163265306122
entropy with depth 85 has validation accuracy of 0.763265306122449
entropy with depth 110 has validation accuracy of 0.7510204081632653
final model with criterion entropy and depth 85 has test accuracy of 0.746938775510204

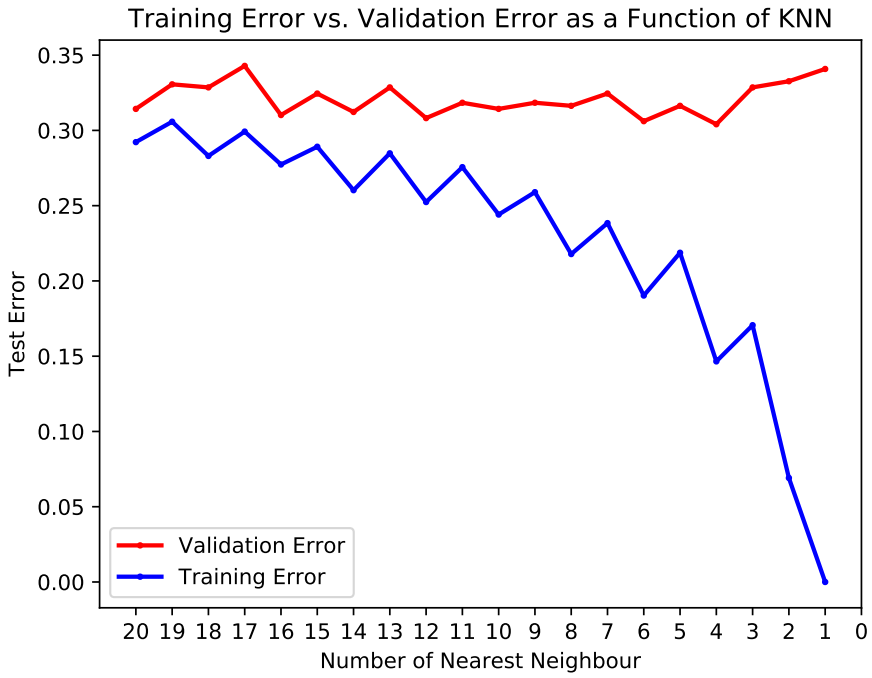
c) The first two layers of the final decision tree model is the following:



d) The information gains for a few of the selected words are the following:
The feature "the" has information gain of 0.05252439711956147
The feature "hillary" has information gain of 0.04053352264739729

The feature "donald" has information gain of 0.04234763323723778
The feature "trumps" has information gain of 0.04542260727760428
The feature "energy" has information gain of 0.001178833829682091
The feature "turnbull" has information gain of 0.0137276684390224
The feature "le" has information gain of 0.003541633864944771
The feature "era" has information gain of 0.0004909268668703559
The feature "black" has information gain of 0.013660799064067808
The feature "2016" has information gain of 0.0010946761454183607
The feature "clinton" has information gain of 0.008605186517812169
The feature "breitbart" has information gain of 5.897703276480648e-05
The feature "election" has information gain of 0.0008873708798791125

e) The graph comparing the validation error to the training error for the different values of k in the KNN model:



A summary of all the validation accuracy for the different models:
KNN with k = 1 gives validation accuracy of 0.6591836734693878
KNN with k = 2 gives validation accuracy of 0.6673469387755102
KNN with k = 3 gives validation accuracy of 0.6714285714285714
KNN with k = 4 gives validation accuracy of 0.6959183673469388
KNN with k = 5 gives validation accuracy of 0.6836734693877551

KNN with k = 6 gives validation accuracy of 0.6938775510204082
KNN with k = 7 gives validation accuracy of 0.6755102040816326
KNN with k = 8 gives validation accuracy of 0.6836734693877551
KNN with k = 9 gives validation accuracy of 0.6816326530612244
KNN with k = 10 gives validation accuracy of 0.6857142857142857
KNN with k = 11 gives validation accuracy of 0.6816326530612244
KNN with k = 12 gives validation accuracy of 0.6918367346938775
KNN with k = 13 gives validation accuracy of 0.6714285714285714
KNN with k = 14 gives validation accuracy of 0.6877551020408164
KNN with k = 15 gives validation accuracy of 0.6755102040816326
KNN with k = 16 gives validation accuracy of 0.689795918367347
KNN with k = 17 gives validation accuracy of 0.6571428571428571
KNN with k = 18 gives validation accuracy of 0.6714285714285714
KNN with k = 19 gives validation accuracy of 0.6693877551020408
KNN with k = 20 gives validation accuracy of 0.6857142857142857

Lastly, the final model with k = 4 has test accuracy of 0.689795918367347.