1. a) Z can be rewritten as $Z = |X - Y|^2 = (\sqrt{(X-Y)^2})^2 = (X-Y)^2$.

A property of expectation that will be used is that if X and Y are absolutely continuous random variables with joint density $f_{X,Y}(x,y)$ and $h : \mathbb{R}^2 \to \mathbb{R}$ is a function, then

$$\mathbb{E}[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f(x,y)\, dx\, dy \tag{1}$$

Using Z in place $h(X,Y)$ and $(x-y)^2$ in place of $h(x,y)$ in (1) gives

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-y)^2 f(x,y)\, dx\, dy \tag{2}$$

Since X and Y are uniform random variables on $[0,1]$, it follows that $f_X(x) = f_Y(y) = 1$ for all $x, y \in [0,1]$ and $f_X(x) = f_Y(x)$ for all $x, y \notin [0,1]$. Since X and Y are independent, it follows that for all x and y,

$$f_{X,Y}(x,y) = f_X(x) f_Y(x) \tag{3}$$

so $f_{X,Y}(x,y) = (1)(1) = 1$ for $(x,y) \in [0,1]^2$ and $f_{X,Y} = 0$ for $(x,y) \notin [0,1]^2$. Substituting this into (2) gives

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^1 \int_0^1 (x-y)^2 (1)\, dx\, dy \\
&= \int_0^1 \int_0^1 (x^2 - 2xy + y^2)\, dx\, dy \\
&= \int_0^1 \left(\frac{x^3}{3} - \frac{2x^2 y}{2} + \frac{xy^2}{1} \bigg|_0^1 \right) dx\, dy \\
&= \int_0^1 \left[\left(\frac{1}{3} - \frac{2y}{2} + \frac{y^2}{1} \right) - \left(\frac{0}{3} - \frac{2(0)(y)}{2} + \frac{(0)(y^2)}{1} \right) \right] dx\, dy \\
&= \int_0^1 \left(\frac{1}{3} - y + y^2 \right) dy \\
&= \left(\frac{y}{3(1)} - \frac{y^2}{2} + \frac{y^3}{3} \right) \bigg|_0^1 \\
&= \left(\frac{1}{3} - \frac{1}{2} + \frac{1}{3} \right) - \left(\frac{0}{3} - \frac{0}{2} + \frac{0}{3} \right) \\
&= \frac{1}{6}
\end{aligned}$$

Another property that will be used is

$$\operatorname{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \tag{4}$$

We have $Z^2 = [(X-Y)^2]^2 = (X-Y)^4$. Using Z^2 in place of $h(X,Y)$, $(x-y)^4$ in place of $h(x,y)$, and (3) in (1) gives

$$\mathbb{E}[Z^2] = \int_0^1 \int_0^1 (x-y)^4 (1)\, dx\, dy \tag{5}$$

Let $u = x - y$ so that $du = dx$. The limits of integration of the inner integral in (5) become $1 - y$ and $0 - y = -y$ so (5) becomes

$$\begin{aligned}
\mathbb{E}[Z^2] &= \int_0^1 \int_{1-y}^{-y} u^4 \, du \, dy \\
&= \int_0^1 \left(\frac{u^5}{5} \right) \Bigg|_{-y}^{1-y} dy \\
&= \int_0^1 \left(\frac{(1-y)^5 - (-y)^5}{5} \right) dy \\
&= \frac{1}{5} \left(\int_0^1 (1-y)^5 \, dy + \int_0^1 y^5 \, dx \right)
\end{aligned}$$

Let $v = 1 - y$ so that $dy = -dv$. The limits of integration of the first integral in (6) become $1 - 1 = 0$ and $1 - 0 = 1$ so (6) becomes

$$\begin{aligned}
\mathbb{E}[Z^2] &= \frac{1}{5} \left(-\int_1^0 v^5 \, dv + \int_0^1 y^5 \, dy \right) \\
&= \frac{1}{5} \left(\int_0^1 v^5 \, dv + \int_0^1 y^5 \, dy \right) \\
&= \frac{2}{5} \left(\int_0^1 v^5 \, dv \right) \\
&= \frac{2}{5} \left(\frac{v^6}{6} \Bigg|_0^1 \right) \\
&= \frac{2}{5} \left(\frac{1}{6} - \frac{0}{6} \right) \\
&= \frac{1}{15}
\end{aligned}$$

Substituting $\mathbb{E}[Z] = \frac{1}{6}$ and $\mathbb{E}[Z^2] = \frac{1}{15}$ into (4) gives

$$\begin{aligned}
\mathrm{Var}[Z] &= \frac{1}{15} - \left(\frac{1}{6} \right)^2 \\
&= \frac{12}{180} - \frac{5}{180} \\
&= \frac{7}{180}
\end{aligned}$$

Thus, $\mathbb{E}[Z] = \frac{1}{6}$ and $\mathrm{Var}[Z] = \frac{7}{180}$.

b) For all $1 \leq i \leq d$, $Z_i \sim Z$ where Z is defined in a). Thus, $\mathbb{E}[Z_i] = \frac{1}{6}$ and $\text{Var}[X_i] = \frac{7}{180}$ for all $1 \leq i \leq d$. Since expectation is linear,

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}\left[\sum_{i=1}^{d} Z_i\right] \\
&= \sum_{i=1}^{d} \mathbb{E}[X_i] \\
&= \frac{d}{6}
\end{aligned}$$

For any $1 \leq i < j \leq d$, X_i, Y_i, X_j, and Y_j are pairwise independent. Thus, $Z_i = g(X_i, Y_i)$ and $Z_j = g(X_i, Y_i)$ with $g(A, B) = A - B$ are independent so the Z_i are pairwise independent. Thus,

$$\begin{aligned}
\text{Var}[R] &= \text{Var}\left[\sum_{i=1}^{d} Z_i\right] \\
&= \sum_{i=1}^{d} \text{Var}[Z_i] \\
&= \frac{7d}{180}
\end{aligned}$$

Thus, $\mathbb{E}[R] = \frac{d}{6}$ and $\text{Var}[R] = \frac{7d}{180}$.

c) Let $\mathbf{0}_d$ and $\mathbf{1}_d$ respectively denote the d-dimensional vectors with all entries equal to 0 and 1. These points are opposite corners of a d-dimensional unit cube so the squared distance between them is the maximum squared distance between two points in a d-dimensional unit cube, which is

$$\begin{aligned}
||\mathbf{1}_d - \mathbf{0}_d||_2^2 &= \left(\sqrt{\sum_{i=1}^{d}(1-0)^2}\right)^2 \\
&= d
\end{aligned}$$

so $\frac{\mathbb{E}[R]}{d} = \frac{d/6}{d} = \frac{1}{6}$ is the ratio of the average squared distance to the largest square distance between two points in a d-dimensional unit cube. This shows that the squared distance grows with the maximum square distance, which grows as the dimension increases so most points are far away in high dimensional space.

The standard deviation of R is $\sqrt{\frac{7d}{180}} = \sqrt{d}\sqrt{\frac{7}{180}}$ so $\frac{\sqrt{d}\sqrt{7/180}}{d} = \sqrt{\frac{7}{180}}\frac{1}{\sqrt{d}}$ is the ratio of the standard deviation of the square distance to the largest square distance between two points in a d-dimensional unit cube. This ratio decreases at d increases so distances between points tend to be close to the average distance between two points, which shows these distances are approximately the same at a large d.

2. a) For any $x \in \mathcal{X}$, we have $0 < p(x) \leq 1$ so $\frac{1}{p(x)} \geq 1 \Rightarrow \log_2\left(\frac{1}{p(x)}\right) \geq 0 \Rightarrow p(x)\log_2\left(\frac{1}{p(x)}\right) \geq 0$. Thus, each term of $H(X)$ is non-negative so $H(X)$ is non-negative.

b) Let $p(x,y)$ be the joint probability function of X and Y. By definition of joint entropy, we have

$$H(X,Y) = \sum_x \sum_y p(x,y)\log_2\left(\frac{1}{p(x,y)}\right) \tag{6}$$

Since X and Y are independent, it follows that $p(x,y) = p(x)p(y)$ for all x and y so (6) becomes

$$\begin{aligned}
H(X,Y) &= \sum_x \sum_y p(x)p(y)\left[\log_2\left(\frac{1}{p(x)}\right) + \log_2\left(\frac{1}{p(y)}\right)\right] \\
&= \left[\sum_x \sum_y p(x)p(y)\log_2\left(\frac{1}{p(x)}\right)\right] + \left[\sum_x \sum_y p(x)p(y)\log_2\left(\frac{1}{p(y)}\right)\right] \\
&= \left[\left(\sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)\right)\left(\sum_y p(y)\right)\right] + \left[\left(\sum_x p(x)\right)\left(\sum_y p(y)\log_2\left(\frac{1}{p(y)}\right)\right)\right] \\
&= \left(\sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)\right)(1) + (1)\left(\sum_y \log_2 p(y)\log_2\left(\frac{1}{p(y)}\right)\right) \\
&= H(X) + H(Y)
\end{aligned}$$

c) By definition of joint entropy, we have

$$\begin{aligned}
H(X,Y) &= \sum_x \sum_y p(x,y)\log_2\left(\frac{1}{p(x,y)}\right) \\
&= \sum_x \sum_y p(x)\frac{p(x,y)}{p(x)}\log_2\left(\frac{1}{p(x)}\frac{p(x)}{p(x,y)}\right) \\
&= \sum_x \sum_y p(x)\frac{p(x,y)}{p(x)}\left[\log_2\left(\frac{1}{p(x)}\right) + \log_2\left(\frac{p(x)}{p(x,y)}\right)\right] \\
&= \sum_x \sum_y p(x)\frac{p(x,y)}{p(x)}\log_2\left(\frac{1}{p(x)}\right) + \sum_x \sum_y p(x)\frac{p(x,y)}{p(x)}\log_2\left(\frac{p(x)}{p(x,y)}\right) \\
&= \sum_x \sum_y p(x)p(y|x)\log_2\left(\frac{1}{p(x)}\right) + \sum_y \sum_x p(x)p(y|x)\log_2\left(\frac{1}{p(y|x)}\right) \\
&= \sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)\sum_y p(y|x) + \sum_y p(y|x)\log_2\left(\frac{1}{p(y|x)}\right)\sum_x p(x) \\
&= \sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)(1) + \sum_y p(y|x)\log_2\left(\frac{1}{p(y|x)}\right)(1) \\
&= H(X) + H(Y|X)
\end{aligned}$$

d) Let X be the random variable that satisfies $p_X\left(\frac{q(x)}{p(x)}\right) = p(x)$ for all $p(x) \neq 0$, and $p_X(x) = 0$ for all other x. Since p is a distribution, $\sum_x p_X(x) = \sum_x p(x) = 1$ so X is a valid random variable. From the appendix, $f(x) = \log_2(x)$ is concave so $-f(x) = -\log_2(x)$ is convex by definition of concave. Using Jensen's inequality with X as the random variable and $-f(x)$ as the function gives

$$
\begin{aligned}
-f(\mathbb{E}(X)) &\leq \mathbb{E}(-f(X)) \\
\Rightarrow -\log_2(\mathbb{E}(X)) &\leq \mathbb{E}(-\log_2(X)) \\
\Rightarrow -\log_2\left(\sum_x p_X\left(\frac{q(x)}{p(x)}\right)\frac{q(x)}{p(x)}\right) &\leq \sum_x p_X\left(\frac{q(x)}{p(x)}\right)\left(-\log_2\left(\frac{q(x)}{p(x)}\right)\right) \\
\Rightarrow -\log_2\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) &\leq \sum_x p(x)\left(-\log_2\left(\frac{q(x)}{p(x)}\right)\right) \\
\Rightarrow -\log_2\left(\sum_x q(x)\right) &\leq \sum_x p(x)\log_2\left(\frac{p(x)}{q(x)}\right)
\end{aligned}
$$

Since q is a distribution, $\sum_x q(x) = 1$ so

$$
\begin{aligned}
-\log_2 1 &\leq \sum_x p(x)\log_2\left(\frac{p(x)}{q(x)}\right) \\
\Rightarrow 0 &\leq \sum_x p(x)\log_2\left(\frac{p(x)}{q(x)}\right)
\end{aligned}
$$

so $\mathrm{KL}(p||q)$ is non-negative.

e) By definition of $\mathrm{KL}(p(x,y)||p(x)p(y))$, we have

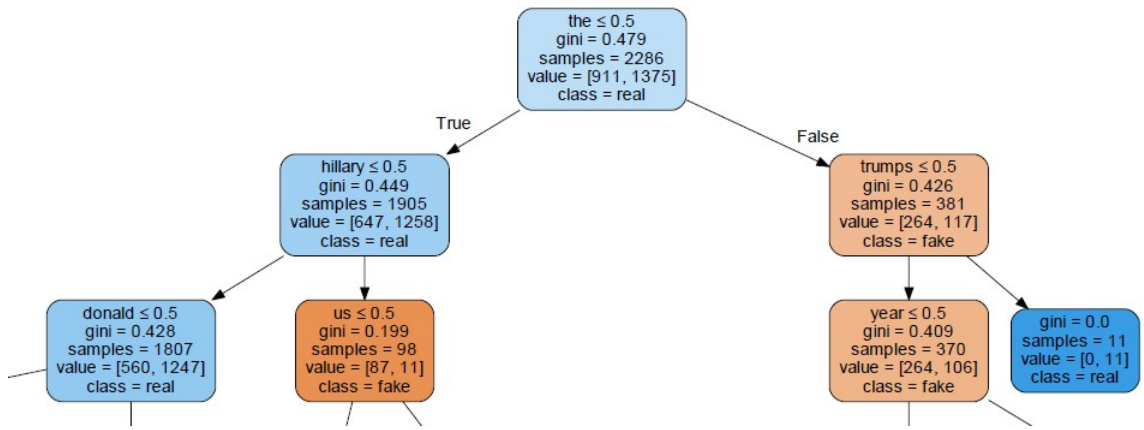
$$
\begin{aligned}
\mathrm{KL}(p(x,y)||p(x)p(y)) &= \sum_x \sum_y p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right) \\
&= \sum_x \sum_y \frac{p(x,y)}{p(x)} p(x) \left[\log_2 \left(\frac{1}{p(y)} \right) + \log_2 \left(\frac{p(x,y)}{p(x)} \right) \right] \\
&= \sum_x \sum_y p(y|x) p(x) \left[\log_2 \left(\frac{1}{p(y)} \right) + \log_2 p(y|x) \right] \\
&= \sum_y \sum_x p(y|x) p(x) \left[\log_2 \left(\frac{1}{p(y)} \right) - \log_2 \left(\frac{1}{p(y|x)} \right) \right] \\
&= \sum_y \sum_x p(y|x) p(x) \log_2 \left(\frac{1}{p(y)} \right) - \sum_y \sum_x p(y|x) p(x) \log_2 \left(\frac{1}{p(y|x)} \right) \\
&= \sum_y \log_2 \left(\frac{1}{p(y)} \right) \sum_x p(y|x) p(x) - \sum_y p(y|x) \log_2 \left(\frac{1}{p(y|x)} \right) \sum_x p(x) \\
&= \sum_y \log_2 \left(\frac{1}{p(y)} \right) \sum_x p(x,y) - \sum_y p(y|x) \log_2 \left(\frac{1}{p(y|x)} \right) (1) \\
&= \sum_y \log_2 \left(\frac{1}{p(y)} \right) p(y) - \sum_y p(y|x) \log_2 \left(\frac{1}{p(y|x)} \right) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

3b)

```
Accuracy for Gini classifier of depth 1 : 0.5959
Accuracy for IG classifier of depth 1 : 0.5959


Accuracy for Gini classifier of depth 3 : 0.7184
Accuracy for IG classifier of depth 3 : 0.6653


Accuracy for Gini classifier of depth 7 : 0.7102
Accuracy for IG classifier of depth 7 : 0.7041


Accuracy for Gini classifier of depth 15 : 0.7408
Accuracy for IG classifier of depth 15 : 0.7265


Accuracy for Gini classifier of depth 30 : 0.7653
Accuracy for IG classifier of depth 30 : 0.7388


Accuracy for Gini classifier of depth 60 : 0.7653
Accuracy for IG classifier of depth 60 : 0.751


Accuracy for Gini classifier of depth 100 : 0.7673
Accuracy for IG classifier of depth 100 : 0.751


Accuracy for Gini classifier of depth 300 : 0.751
Accuracy for IG classifier of depth 300 : 0.7367


Accuracy for Gini classifier of depth 1000 : 0.751
Accuracy for IG classifier of depth 1000 : 0.7367
```

c)

```
Accuracy of optimal classifier on test data: 0.7776
```



d)

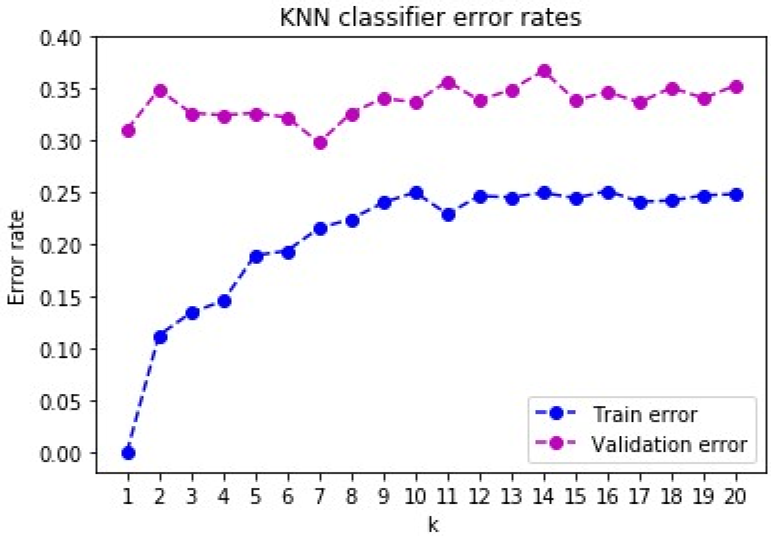
```
Information gain on keyword the : 0.05080161270165395
Information gain on keyword hillary : 0.05175690958227086
Information gain on keyword donald : 0.04971770431093536
Information gain on keyword trump : 0.023846116804407735
```

e)

```
Training error rate for KNN classifier with 1 neighbors: 0.0
Validation error rate for KNN classifier with 1 neighbors: 0.31020000000000003


Training error rate for KNN classifier with 2 neighbors: 0.1129
Validation error rate for KNN classifier with 2 neighbors: 0.349


Training error rate for KNN classifier with 3 neighbors: 0.13470000000000004
Validation error rate for KNN classifier with 3 neighbors: 0.3265


Training error rate for KNN classifier with 4 neighbors: 0.1461
Validation error rate for KNN classifier with 4 neighbors: 0.3245


Training error rate for KNN classifier with 5 neighbors: 0.1894
Validation error rate for KNN classifier with 5 neighbors: 0.3265


Training error rate for KNN classifier with 6 neighbors: 0.19379999999999997
Validation error rate for KNN classifier with 6 neighbors: 0.3224


Training error rate for KNN classifier with 7 neighbors: 0.2157
Validation error rate for KNN classifier with 7 neighbors: 0.29800000000000004


Training error rate for KNN classifier with 8 neighbors: 0.22440000000000004
Validation error rate for KNN classifier with 8 neighbors: 0.3265
```

```
Training error rate for KNN classifier with 9 neighbors: 0.24019999999999997
Validation error rate for KNN classifier with 9 neighbors: 0.3408


Training error rate for KNN classifier with 10 neighbors: 0.2502
Validation error rate for KNN classifier with 10 neighbors: 0.3367


Training error rate for KNN classifier with 11 neighbors: 0.22919999999999996
Validation error rate for KNN classifier with 11 neighbors: 0.3571


Training error rate for KNN classifier with 12 neighbors: 0.24670000000000003
Validation error rate for KNN classifier with 12 neighbors: 0.3388


Training error rate for KNN classifier with 13 neighbors: 0.24539999999999995
Validation error rate for KNN classifier with 13 neighbors: 0.349


Training error rate for KNN classifier with 14 neighbors: 0.24980000000000002
Validation error rate for KNN classifier with 14 neighbors: 0.36729999999999996


Training error rate for KNN classifier with 15 neighbors: 0.24450000000000005
Validation error rate for KNN classifier with 15 neighbors: 0.3388


Training error rate for KNN classifier with 16 neighbors: 0.25149999999999995
Validation error rate for KNN classifier with 16 neighbors: 0.3469


Training error rate for KNN classifier with 17 neighbors: 0.241
Validation error rate for KNN classifier with 17 neighbors: 0.3367


Training error rate for KNN classifier with 18 neighbors: 0.24229999999999996
Validation error rate for KNN classifier with 18 neighbors: 0.351


Training error rate for KNN classifier with 19 neighbors: 0.24719999999999998
Validation error rate for KNN classifier with 19 neighbors: 0.3408


Training error rate for KNN classifier with 20 neighbors: 0.24850000000000005
Validation error rate for KNN classifier with 20 neighbors: 0.35309999999999997
```

```
Accuracy of KNN with 7 neighbours on test data: 0.6551
```


KNN classifier error rates
Error rate
k
Train error
Validation error
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20