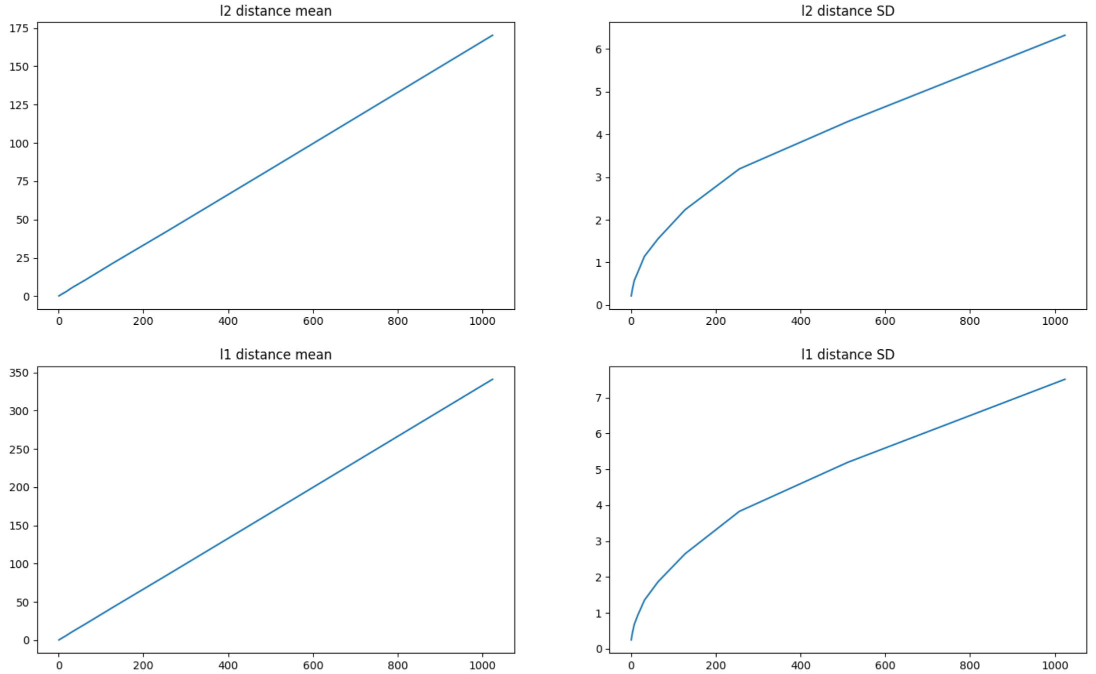
Q1

a)



b) $E[R] = E[z_1 + \dots + z_d]$

$= E[z_1] + \dots + E[z_d]$

$= d \times (\frac{1}{6})$

$= d/6$

Since X_i, Y_i are independently sampled for each i, z_i and z_j are independent for $i \neq j$

$\Rightarrow Var[R] = Var[z_1 + \dots + z_d]$

$= Var[z_1] + \dots + Var[z_d]$

$= d \times \frac{7}{180}$

$= 7d/180$

c) i) Let R be the Euclidian distance

then E is: $R - E[R] \leq d$

ii) $P(R - E[R] \leq d) = 1 - P(R - E[R] > d)$

$= 1 - P(R - E[R] \geq d)$

(since E is a conti. random variable)

$\geq 1 - \frac{Var[R]}{d^2}$

iii) then $P(E) \geq 1 - \frac{7d/180}{d^2}$

$= 1 - \frac{7}{180d}$

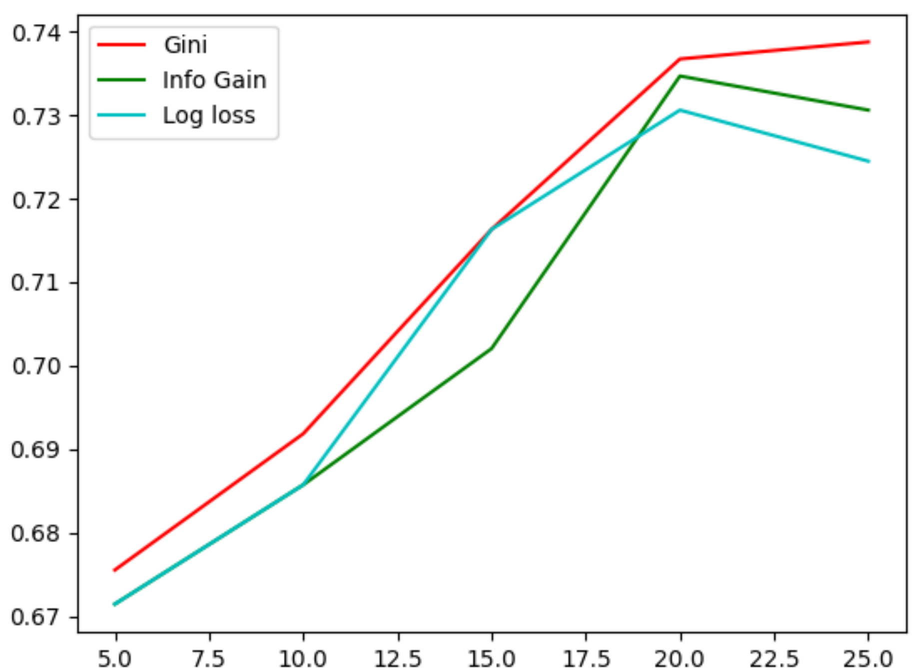
$\Rightarrow \lim_{d \to \infty} P(E) \geq \lim_{d \to \infty} 1 - \frac{7}{180d} = 1$

so as $d \to \infty$, $P(E) = 1$, so any distance is d away from its expectation

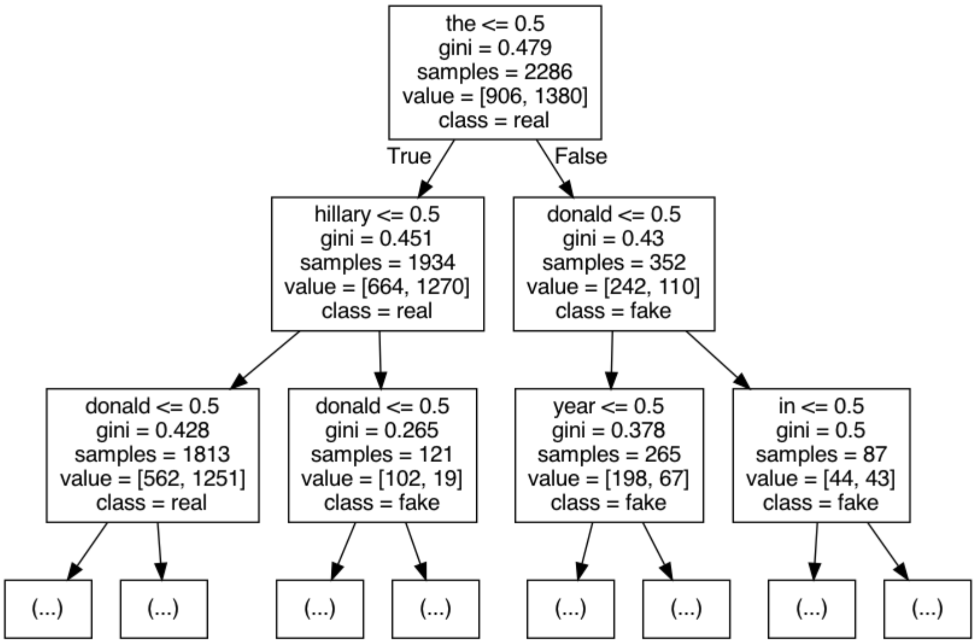
Q2. b) Function output :

```
Gini: score = 0.6755102040816326, depth = 5
Information gain: score = 0.6714285714285714, depth = 5
Log loss: score = 0.6714285714285714, depth = 5
Gini: score = 0.6918367346938775, depth = 10
Information gain: score = 0.6857142857142857, depth = 10
Log loss: score = 0.6857142857142857, depth = 10
Gini: score = 0.7163265306122449, depth = 15
Information gain: score = 0.7020408163265306, depth = 15
Log loss: score = 0.7163265306122449, depth = 15
Gini: score = 0.736734693877551, depth = 20
Information gain: score = 0.7346938775510204, depth = 20
Log loss: score = 0.7306122448979592, depth = 20
Gini: score = 0.7387755102040816, depth = 25
Information gain: score = 0.7306122448979592, depth = 25
Log loss: score = 0.7244897959183674, depth = 25
```

Plot :



c) Gini w/ depth 25 achieved the highest accuracy



d) The keywords are selected from
{"the", "hillary", "trumps", "donald"}

Their IG are as follows:

```
IG(Y|X) is 0.04570772617653496 for the keyword the
IG(Y|X) is 0.04268249633366705 for the keyword hillary
IG(Y|X) is 0.03711785532105771 for the keyword trumps
IG(Y|X) is 0.04197422322376332 for the keyword donald
```

Q3.

a) $$\frac{\partial J}{\partial w_{j'}} = \frac{1}{2N} \cdot \frac{\partial\left(\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2\right)}{\partial w_{j'}}$$

$y^{(i)} = \sum_{j=1}^{D} w_j x_j^{(i)} + b$

$$= \frac{1}{2N} \cdot \frac{\partial\left(\sum_{i=1}^{N}\left(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)}\right)^2\right)}{\partial w_{j'}}$$

$$= \frac{1}{2N} \cdot 2\sum_{i=1}^{N}\left(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)}\right)\left(x_{j'}^{(i)}\right)$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N}\left(\underbrace{\sum_{j=1}^{D} w_j x_j^{(i)} + b}_{= y^{(i)}} - t^{(i)}\right)\left(x_{j'}^{(i)}\right)$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)\left(x_{j'}^{(i)}\right) \qquad (1)$$

$$\frac{\partial R}{\partial w_{j'}} = \frac{\partial\left(\sum_{j=1}^{D} \alpha_j |w_j| + \frac{1}{2}\sum_{j=1}^{D} \beta_j w_j^2\right)}{\partial w_{j'}}$$

$|w_j| := f(w_j) = \begin{cases} w_j, & w_j > 0 \\ 0, & w_j = 0 \\ -w_j, & w_j < 0 \end{cases}$

$$= \begin{cases} \text{if } w_{j'} > 0: & \dfrac{\partial\left(\sum_{j=1}^{D} \alpha_j w_j + \frac{1}{2}\sum_{j=1}^{D} \beta_j w_j^2\right)}{\partial w_{j'}} = \alpha_{j'} + \beta_{j'} w_{j'} \qquad (2) \\ \text{if } w_{j'} = 0: & = 0 \qquad (3) \\ \text{if } w_{j'} < 0: & \partial\left(-\sum_{j=1}^{D} \alpha_j w_j + \frac{1}{2}\sum_{j=1}^{D} \beta_j w_j^2\right) / \partial w_{j'} \end{cases}$$

$$= -\alpha_{j'} + \beta_{j'} w_{j'} \qquad (4)$$

So $\dfrac{\partial J_{reg}^{\alpha\beta}(w)}{\partial w_{j'}}$ is:

if $w_{j'} = 0$: $\frac{1}{N} \cdot \sum_{i=1}^{N} (y^{(i)} - t^{(i)})(x_{j'}^{(i)})$

if $w_{j'} > 0$: $\frac{1}{N} \cdot \sum_{i=1}^{N} (y^{(i)} - t^{(i)})(x_{j'}^{(i)}) + \alpha_{j'} + \beta_{j'} w_{j'}$

if $w_{j'} < 0$: $\frac{1}{N} \cdot \sum_{i=1}^{N} (y^{(i)} - t^{(i)})(x_{j'}^{(i)}) - \alpha_{j'} + \beta_{j'} w_{j'}$

$$\frac{\partial J}{\partial b} = \frac{1}{2N} \cdot \partial \Big[\sum_{i=1}^{N} \big(\underbrace{y^{(i)}}_{\sum_{j=1}^{D} w_j x_j + b} - t^{(i)} \big)^2 \Big] / \partial b$$

$$= \frac{1}{2N} \cdot \sum_{i=1}^{N} 2 \Big(\sum_{j=1}^{D} w_j x_j + b - t^{(i)} \Big)$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} (y^{(i)} - t^{(i)})$$

Let $\alpha > 0$ be the learning rate, so overall:

if $w_{j'} > 0$: $w_{j'} \leftarrow w_{j'} - \alpha \Big(\frac{1}{N} \cdot \sum_{i=1}^{N} (y^{(i)} - t^{(i)})(x_{j'}^{(i)}) + \alpha_{j'} + \beta_{j'} w_{j'} \Big)$

$$\Leftrightarrow w_{j'} \leftarrow w_{j'} - \frac{\alpha}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})(x_{j'}^{(i)}) - \alpha_{j'} \alpha - \alpha \beta_{j'} w_{j'}$$

$$\iff w_{j'} \leftarrow w_{j'}(1-\alpha\beta_{j'}) - \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})(x_{j'}^{(i)}) - \alpha_{j'}\alpha$$

$$b \leftarrow b - \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})$$

if $w_{j'}=0$: $w_{j'} \leftarrow w_{j'} - \alpha\left(\frac{1}{N}\cdot\sum_{i=1}^{N}(y^{(i)}-t^{(i)})(x_{j'}^{(i)})\right)$

$$\iff w_{j'} \leftarrow w_{j'} - \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})(x_{j'}^{(i)})$$

$$b \leftarrow b - \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})$$

if $w_{j'}<0$: $w_{j'} \leftarrow w_{j'} - \alpha\left(\frac{1}{N}\cdot\sum_{i=1}^{N}(y^{(i)}-t^{(i)})(x_{j'}^{(i)}) - \alpha_{j'} + \beta_{j'}w_{j'}\right)$

$$\iff w_{j'} \leftarrow (1-\alpha\beta_{j'})w_{j'} - \frac{\alpha}{N}\cdot\sum_{i=1}^{N}(y^{(i)}-t^{(i)})(x_{j'}^{(i)}) + \alpha\alpha_{j'}$$

$$b \leftarrow b - \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})$$

This is called weight decay possibly because for cases $w_{j'}<0$ and $w_{j'}>0$, the update rule for $w_{j'}$ contains the term $(1-\alpha\beta_{j'})w_{j'}$.

$\alpha>0$ and $\beta_{j'}\geq 0$, so $(1-\alpha\beta_{j'})w_{j'} \leq w_{j'}$

$\Rightarrow$ within the update rule, the weight $w_{j'}$ decays to a lesser term.

b) $\lambda_1 = 0$

$$\Rightarrow J^{\beta}_{reg}(w) = \frac{1}{2N} \cdot \sum_{i=1}^{N} \left(y^{(i)} - t^{(i)}\right)^2 + \frac{1}{2} \sum_{j=1}^{D} \beta_j w_j^2$$

Note: for the sake of consistency$^{\text{to 3a)}}$, in my notations I swapped j and j' defined in the question

$$\Rightarrow \frac{\partial J^{\beta}_{reg}(w)}{\partial w_{j'}} = \frac{1}{2N} \cdot \sum_{i=1}^{N} 2\left(y^{(i)} - t^{(i)}\right)\left(x_{j'}^{(i)}\right) + \beta_{j'} w_{j'}$$

$$= \frac{1}{N} \cdot \underbrace{\sum_{i=1}^{N} \left(\sum_{j=1}^{D} \left(w_j x_j^{(i)}\right) - t^{(i)} \right)\left(x_{j'}^{(i)}\right)}_{= \sum_{i=1}^{N} \sum_{j=1}^{D} \left(w_j x_j^{(i)}\right) x_{j'}^{(i)} - \sum_{i=1}^{N} t^{(i)} x_{j'}^{(i)}} + \beta_{j'} w_{j'}$$

$$= \sum_{j=1}^{D} \sum_{i=1}^{N} \frac{1}{N} w_j x_j^{(i)} x_{j'}^{(i)} - \frac{1}{N} \sum_{i=1}^{N} t^{(i)} x_{j'}^{(i)} + \beta_{j'} w_{j'}$$

Define the indicator function $I(j) = \begin{cases} 0 & \text{if } j \neq j' \\ 1 & \text{if } j = j' \end{cases}$

$$= \sum_{j=1}^{D} \sum_{i=1}^{N} \frac{1}{N} w_j x_j^{(i)} x_{j'}^{(i)} + I(j) \beta_j w_j - \frac{1}{N} \sum_{i=1}^{N} t^{(i)} x_{j'}^{(i)}$$

$$= \sum_{j=1}^{D} \left(\frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} + I(j) \beta_j \right) w_j - \frac{1}{N} \sum_{i=1}^{N} t^{(i)} x_{j'}^{(i)}$$

So $A_{jj'} = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} + I(j) B_j$

$C_{j'} = \frac{1}{N}\sum_{i=1}^{N} t^{(i)} x_{j'}^{(i)}$

c) Note that $X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_D^{(1)} \\ \vdots & \vdots & & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{pmatrix}$

Then $A = \underbrace{\begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} x_1^{(i)} & \cdots & \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} x_D^{(i)} \\ \vdots & \ddots & \\ \frac{1}{N}\sum_{i=1}^{N} x_D^{(i)} x_1^{(i)} & & \frac{1}{N}\sum_{i=1}^{N} x_D^{(i)} x_D^{(i)} \end{pmatrix}} + \begin{pmatrix} B_1 & & 0 \\ & \ddots & \\ 0 & & B_D \end{pmatrix}$

$= \frac{1}{N}\begin{pmatrix} x_1^{(1)}x_1^{(1)} + \dots + x_1^{(N)}x_1^{(N)} & \cdots & x_1^{(1)}x_D^{(1)} + \dots + x_1^{(N)}x_D^{(N)} \\ \vdots & \ddots & \\ x_D^{(1)}x_1^{(1)} + \dots + x_D^{(N)}x_1^{(N)} & \cdots & x_D^{(1)}x_D^{(1)} + \dots + x_D^{(N)}x_D^{(N)} \end{pmatrix}$

let $\vec{x}_i = \begin{pmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(N)} \end{pmatrix}$:

$= \frac{1}{N}\begin{pmatrix} \vec{x}_1 \cdot \vec{x}_1 & \cdots & \vec{x}_1 \cdot \vec{x}_D \\ \vdots & \ddots & \\ \vec{x}_D \cdot \vec{x}_1 & & \vec{x}_D \cdot \vec{x}_D \end{pmatrix}$

$$= \frac{1}{N} \underbrace{\begin{pmatrix} - \vec{x}_1 - \\ \vdots \\ - \vec{x}_D - \end{pmatrix}}_{X^T} \underbrace{\begin{pmatrix} | & & | \\ \vec{x}_1 & \cdots & \vec{x}_D \\ | & & | \end{pmatrix}}_{X}$$

$$= \frac{1}{N} X^T X + \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_D \end{pmatrix}$$

$$C = \frac{1}{N} \begin{pmatrix} x_1^{(1)} & \cdots & x_1^{(N)} \\ \vdots & \ddots & \vdots \\ x_D^{(1)} & \cdots & x_D^{(N)} \end{pmatrix} \begin{pmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{pmatrix}$$

$= \frac{1}{N} X^T \vec{t}$ for target vector $\vec{t} = \begin{pmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{pmatrix}$

then $A\vec{w} - C = 0$

$$\Rightarrow \left[\frac{1}{N} X^T X + \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_D \end{pmatrix} \right] \vec{w} - \frac{1}{N} X^T \vec{t} = 0$$

$$\Rightarrow \left[X^T X + N \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_D \end{pmatrix} \right] \vec{w} = X^T \vec{t}$$

$$\Rightarrow \vec{w} = \left[X^T X + N \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_D \end{pmatrix} \right]^{-1} X^T \vec{t}$$

assuming that $X^T X + N \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_D \end{pmatrix}$ is invertible