1. [30 points] **Feature Selection**

   Consider the following table, in which there are four terms' frequencies for the class Coffee in the first 100,000 documents of Reuters-RCV1:

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98012	102	1835	51
council	96322	133	3525	20
producers	98524	119	1118	34
roasted	99824	143	23	10

   Please select two of these four terms in terms of the following criteria:

   - [10/30 points] Chi-square test
   - [10/30 points] Mutual information
   - [10/30 points] TF-IDF

(a) Chi-square test formula:

$$\chi^2 = \frac{(N_{11}N_{00} - N_{10}N_{01})^2 \cdot N}{(N_{11}+N_{10})(N_{11}+N_{01})(N_{10}+N_{00})(N_{01}+N_{00})}$$

1. brazil:

   $N = 98012 + 102 + 1835 + 51 = 100000$

   $$\chi^2_{brazil} = \frac{(51 \cdot 98012 - 1835 \cdot 102)^2 \cdot 100000}{(51+1835)(51+102)(1835+98012)(102+98012)} = 818.938$$

2. council:

   $N = 96322 + 133 + 3525 + 20 = 100000$

   $$\chi^2_{council} = \frac{(20 \cdot 96322 - 3525 \cdot 133)^2 \cdot 100000}{(20+3525)(20+133)(3525+96322)(133+96322)} = 40.67$$

3. producers:

   $N = 98524 + 119 + 1118 + 34 = 100000$

   $$\chi^2_{producer} = \frac{(34 \cdot 98524 - 1118 \cdot 119)^2 \cdot 100000}{(34+1118)(34+119)(1118+98524)(119+98524) = 597.3}$$

4. roasted:

   $N = 99824 + 143 + 23 + 10 = 100000$

   $$\chi^2_{roasted} = \frac{(10 \cdot 99824 - 23 \cdot 143)^2 \cdot 100000}{(10+23)()(51+102)(1.02+98012)} = 1950.94$$

Ans: roasted, brazil

(b) Mutual Information formula:
$I(t,c) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_1N_1} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_0N_1} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_1N_0} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_0N_0}$

$N = 100000$

1. brazil: $1.55 * 10^{-3}$
2. council: $1.8 * 10^{-4}$
3. producers: $4 * 10^{-3}$
4. roastedd: $6.5 * 10^{-4}$

Ans: producers, brazil

(c) TF-IDF formula:
$TF - IDF = t_f * \log\frac{N}{df}$

1. brazil: 143.58
2. council: 56.31
3. producers: 95.72
4. roastedd: 28.5

Ans: brazil, producers

2. [20 points] **Evaluations of text classification**
   Considering the following two tables of text classification results:

class 1	**truth: yes**	**truth: no**
call: yes	80	10
call: no	20	890

class 2	**truth: yes**	**truth: no**
call: yes	20	40
call: no	60	880

   Please compute the following classification performance:
   - [10/20 points] Macro-averaged precision
   - [10/20 points] Micro-averaged precision

1. Macro precision formula:
   $P_{i,macro} = \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}$
   $P_{i,macro} = (\frac{80}{90} + \frac{20}{20+40}) * \frac{1}{2} \approx 0.611$
2. Micro precision formula
   $P_{i,micro} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i}$
   $P_{i,micro} = \frac{80+20}{80+10+20+40} \approx 0.667$

1. [20 points] **Vector Space Classification**
   - [15/20 points] Show the decision boundary/surface of Rocchio classification with two centroids (c1, c2). Please specify the derivation in detail.
   - [5/20 points] Describe what would happen when Rocchio classification method encounters the situation of multimodal classes.

1. decision boundary/surface
   $c_1 = \frac{1}{C_1} \sum_x x_i, c_2 = \frac{2}{n_2} \sum_{j=1}^{n_2} x_j$
   Assume: $c_1 = (a1, b1), c_2 = (a2, b2)$
   the perpendicular bisector
   $(x - c_1) \cdot (c_2 - c_1) = 0$
   $(x - (a1, b1)) \cdot ((a2, b2) - (a1, b1)) = 0$
   $(x_1 - a1, x_2 - b1) \cdot (a2 - b2, a1 - b1) = 0$
2. What would happended?
   When classes are distributed multimodally, Rocchio may misclassify samples that are closer to the centroid of another class.
   Also, boundaries that cut through dense regions of a class, leading to poor classification performance.

1. [30 points] **Vector Space Classification**
   In the following figure, there are three points: a = (0.5, 1.5), b = (4, 4), c = (8, 6), each of which representing a specific class, and there is a testing point x = (2, 2):
   Please answer the following questions:
   - [10/30 points] The most similar class to x according to the inner product similarity
   - [10/30 points] The most similar class to x according to the cosine similarity
   - [10/30 points] The most closest class to x according to Euclidean distance

1. Inner Product: $\vec{x} \cdot \mu(c) = x_1\mu_1 + x_2\mu_2$
   $inner_a = (2, 2) * (0.5, 1.5) = 4$

$$inner_b = (2,2) * (4,4) = 16$$
$$inner_c = (2,2) * (8,6) = 28$$

Ans: "c" is the most similar

2. Cosine Similarity: $\frac{\vec{x} \cdot \mu(c)}{||\vec{x}| \cdot |\mu(c)||} = \frac{x_1\mu_1 + x_2\mu_2}{||\vec{x}| \cdot |\mu(c)||}$

$$cos_a = \frac{(2,2) * (0.5,1.5)}{\sqrt{2^2+2^2}\sqrt{0.5^2+1.5^2}} \approx 0.89$$
$$cos_b = \frac{(2,2) * (4,4)}{\sqrt{2^2+2^2}\sqrt{4^2+4^2}} \approx 1$$
$$cos_c = \frac{(2,2) * (8,6)}{\sqrt{2^2+2^2}\sqrt{8^2+6^2}} \approx 0.99$$

Ans: "b" is the most similar

3. Euclidean Distance: $\sqrt{(x_1-\mu_1)^2 + (x_1-\mu_2)^2}$

$$d_a = \sqrt{2.25 + 0.25} \approx 1.581$$
$$d_b = \sqrt{4+4} \approx 2.828$$
$$d_c = \sqrt{36+16} \approx 7.211$$

a is the nearest

Ans: "a" is the most similar