Université Libanaise
Faculté d'Information

Br I

الجامعة اللبنانية
كلية الاعلام
الفرع الاول

الامتحانات النهائية للفصل الرابع
من العام الجامعي 2023-2024

المرحلة: الإجازة	المادة: Data Analysis
السنة المنهجية: الثانية	المدة: ساعة ونصف
الاختصاص: علم البيانات - Data Science	الأستاذ: د. حسين هزيمة

Exercise 1: (30 pts) Numerical data analysis:

1. Given the following problem: the number of sold products in a company depends on the number of advertisements they create. The monthly sold products and their monthly number of advertisements are shown in the table below:

#products sold / month (x)	# advertisements/month
13	2
16	5
15	3
11	X
15	8
17	6

   i. Suppose that X is a missing value, find the value of X using linear interpolation rule. **(5pts)**

   ii. After finding the regression equation $\hat{y}=mx+b$, calculate the future value of #products sold when having 7 advertisements arrived. **(25pts)**

Exercise 2: (70 pts) Textual data analysis:

I. Given the following 3 matrices of the terms inside five web pages. Matrix A contains the total number of terms inside each web page including without data pre-processing. Matrix B contains the total number of stop words only inside each web page. Matrix C contains the total number of special characters only inside each web page.

m = 0.44
b = -2.38

Web page \ term	t_1	t_2	t_3	t_4
w_1	8	5	7	3
w_2	10	7	6	5
w_3	4	3	3	12
w_4	15	10	10	2
w_5	8	2	4	1

Matrix A

Web page \ term	t_1	t_2	t_3	t_4
w_1	1	1	1	1
w_2	2	1	1	3
w_3	1	1	1	4
w_4	3	2	1	0
w_5	1	1	0	0

Matrix B

Web page \ term	t_1	t_2	t_3	t_4
w_1	1	1	1	1
w_2	2	1	1	1
w_3	1	1	1	1
w_4	2	2	1	0
w_5	1	1	1	0

Matrix C

1. Construct the new term-document matrix after performing the pre-processing on Matrix A. **(15pts)**

2. List all the vectors of the five web pages. **(5pts)**

3. Calculate the most similar web page to w_1, by using the cosine similarity. **(10pts)**

4. Compare the cosine similarity values, before and after pre-processing, $\cos_b(w_1, w_2)$, $\cos_a(w_1, w_2)$. **(15pts)**

5. Suppose that after pre-processing the terms inside w_1 and w_2 are as follows:

   a. w_1 = "Lebanese computer data science beirut"

   b. w_2 = "Data computer physics faculty beriut"

   c. Compute the Jaccard similarity between w_1 and w_2. **(10pts)**

   Note: construct the matrix and circle the final edit distance value

   d. **II.** Find the corresponding number of edits after comparing the two strings in the table below **(15pts)**

Intention
execution

#	I	N	T	E	N	T	I	O	N
#									
E									
X									
E									
C									
U									
T									
I									
O									
N									

Appendix (permitted document)

$$idf(w) = log(\frac{N}{df_t})$$

$$tf_{t,j} = \frac{n_{t,j}}{\sum_k n_{t,j}}$$

$$w_{t,j} = tf_{t,j} \times \log\left(\frac{N}{df_t}\right)$$

$$\text{CosSim}(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$

$\bar{x} = \Sigma x / n$

This is just the mean of the x values.

$\bar{y} = \Sigma y / n$

This is just the mean of the y values.

$S_{xx} = SS_{xx} = \Sigma(x(i) - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2 / n$

$S_{yy} = SS_{yy} = \Sigma(y(i) - \bar{y})^2 = \Sigma y^2 - (\Sigma y)^2 / n$

$S_{xy} = SS_{xy} = \Sigma(x(i) - \bar{x})(y(i) - \bar{y}) = \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y) / n$

Slope $m = SS_{xy} / SS_{xx}$

Intercept, $b = \bar{y} - m \cdot \bar{x}$

y-predicted = $\text{yhat}(i) = m \cdot x(i) + b$.

Residual(i) = Error(i) = $y - \text{yhat}(i)$.

$SSE = S_{res} = SS_{res} = SS_{errors} = \Sigma[y(i) - \text{yhat}(i)]^2$.

Standard deviation of residuals = $s = S_{res} = S_{errors} = [SS_{res} / (n-2)]^{1/2}$.

Standard error of the slope (m) = $S_{res} / SS_{xx}^{1/2}$.

Standard error of the intercept (b) = $S_{res}[(SS_{xx} + n \cdot \bar{x}^2) / (n \cdot SS_{xx})]^{1/2}$.

Good Work