Université Libanaise
Faculté d'Information
Br I

الجامعة اللبنانية
كلية الاعلام
الفرع الاول

الامتحانات النهائية للفصل الرابع
من العام الجامعي 2022-2023

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

Exercise 1: (10 pts) Answer the following questions:

a. What is the difference between XML and JSON, give a simple example? **(5pts)**

b. How does TF-IDF technique lower the importance of stop words?. **(5pts)**

Exercise 2: (30 pts) Numerical data analysis:

a. Mention if the following time series data has trend, seasonality, or both, and discuss why:

i. Dollar rate change in Lebanon from 1992 until 2020, given that the rate from 1992 until 2019 was approx. 1550 LBP, then from 2019 to 2020 it increased from 1550 LBP to 8000 LBP. **(3pts)**

ii. Coronavirus cases in Lebanon in 2019-2020. **(3pts)**

iii. Resting Heart rate of a normal person within 60 seconds. **(3pts)**

b. Given the following problem: the Coronavirus cases evolution in Lebanon depend on the number of airplanes arriving every day to the Beirut Rafik Hariri International Airport. The daily arrived airplanes and the daily Coronavirus cases are given in the following Matrix:

#of arriving airplanes / day (x)	Coronavirus cases / day (y)
2	8
5	19
3	9
5	13
7	16
8	18

i. After finding the regression equation **yhat=mx+b,** calculate the future value of Coronavirus cases when having 4 airplanes arrived. **(6pts)**

ii. Imagine that Coronavirus disease is historical and the new yearly matrix is given as follows:

Quarter \ year	2017	2018	2019
Winter	95	104	115
Fall	85	90	X 32
Spring	70	72	Y 68
Summer	68	70	72

i. Calculate the values of **X** and **Y using Linear Interpolation. (3pts)**

ii. Find the moving average (2-MA) using SMA method

iii. Chose the correct forecasting model to find Spring value of year 2017 (F_3) given $\alpha=0.2$, $\beta=0.3$, $\gamma=0.25$. **(9pts)**

Exercise 3: (20 pts) Similarity measures:

Given the following two matrices of the terms inside five web pages. Matrix A contains the total number of terms inside each web page including stop words. Matrix B contains the total number of stop words only inside each web page.

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	3	1	1	4
w_4	5	0	1	0
w_5	1	1	0	0

Matrix B

i. Construct the new term-document matrix after removing the stop words from each term. **(3pts)**

ii. List all the vectors of the five web pages. **(2.5pts)**

iii. Calculate the most similar web page to w_1, by using the cosine similarity. **(7.5pts)**

iv. Compare the cosine similarity values, before and after eliminating the stop words, $\cos_b(w_1, w_2)$, $\cos_a(w_1, w_2)$. **(7pts)**

Good Work

Appendix (permitted document)

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

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

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 = yhat(i) = m·x(i) + b.

Residual(i) = Error(i) = y – yhat(i).

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

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) = $S_{res}[(SS_{xx} + n. \bar{x}^2) / (n \cdot SS_{xx}]^{1/2}$.

$$a_{t+1} = \propto \left(\frac{D_{t+1}}{C_{t+1}}\right) + (1 - \propto)(a_t + b_t)$$

$$b_{t+1} = \beta(a_{t+1} - a_t) + (1 - \beta)\, b_t$$

$$C_{t+p+1} = \gamma \left(\frac{D_{t+1}}{a_{t+1}}\right) + (1 + \gamma)\, C_{t+1}$$

$$F_{t+1} = (a_t + b_t)\, C_{t+1}$$