

**מבחן במסדי נתונים**

**ד"ר עמוס עזריה**

**7028010**

סמסטר ב' מועד א' כ"ג בתמוז התשע"ז, 17.7.2017

**הנחיות כלליות:**

* משך הבחינה: 180 דקות.
* **יש לענות בגוף השאלון!** המחברת תשמש כטיוטא בלבד.
* אין להכניס שום חומר עזר.
* השימוש במחשבון **אסור**.
* בשאלות האמריקאיות רק תשובה אחת נכונה.
* מומלץ לקרוא את ההוראות באנגלית ולפנות לעברית רק במקרה של חוסר הבנה.
* בסיום הבחינה - נא למסור את השאלון ואת המחברת.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
| 103 | 6 | 8 | 6 | 10 | 10 | 5 | 7 | 8 | 10 | 5 | 12 | 16 | Max points |
|  |  |  |  |  |  |  |  |  |  |  |  |  | Grade |

**ב ה צ ל ח ה!**

1 SQL (16 pt)

נתונות הטבלאות הבאות לייצוג מרפאת שיניים:

Patients (patientId, firstName, lastName, YearOfBirth, gender)

TreatmentTypes (treatmentTypeId, treatmentName, price)

TreatmentForPatient (patientId,treatmentTypeId, dayAsInt)

ניתן להניח שdayAsInt הוא מספר שלם המייצג את יום העסקים מאז המרפאה החלה לפעול.

כיתבו את השאילתות הבאות:

1. מספר לקוחות נשים שטופלו בכל יום (gender=1). (8 נק')

\_\_select tp.dayAsInt, count(p.patientId) from

TreatmentForPatient as tp

join Patients as p on(tp.patientId = p.patientId)

where(p.gender = 1) group by(dayAsInt)

1. ההכנסה של המרפאה בכל יום בו הייתה הכנסה לפחות 1000 ש"ח . (8 נק')

\_select tp.dayAsInt, sum(tt.price) as totalPrice from

TreatmentTypes as tt join

TreatmentForPatient as tp

on(tp.treatmentTypeId = tt.treatmentTypeId)

group by(tp.dayAsInt) having (totalPrice > 1000)

2 Normalization (12 pt):

Given the following relation:

נתונה הרלציה הבאה:

R(A, B, C, D, E, F)

And the following dependencies:

והתלויות הבאות:

{A}->B

{A, E, F} ->C

{A, B}->D

{E, F} -> D

{B, D} -> A

a. What are the sets of candidate keys (5 pt):

מה מפתחות הקנדידייט?

1. (A,E,F)
2. (B,E,F)\_

b. What level of NF (normal form) does the relation adhere to? (1NF, 2NF, 3NF, BCNF(3.5NF), 4NF)? Show why the relation does not adhere to any higher NF (7 pt).

באיזו רמה של נורמל פורם?

1NF holds, values are atomic.

2NF fails, since D is non prime and depends on a subset (E,F) of (A,E,F) which is a candidate key.

NO-SQL

3 Redis (5 pt)

Which of the following statements is FALSE:

איזה מבין המשפטים הבאים **אינו** נכון:

1. In Redis it is possible to create an object (named hash) that is itself built from key-value pairs. true

ברדיס אפשר ליצור כניסה שהיא עצמה טבלת גיבוב.

1. Redis has special commands that allow the user to get back only part of the value (e.g. return only a specific node of an XML). False!

ברדיס יש פקודות לאחזור חלקים ממסמך.

1. Redis supports lists, that allow the user to push new items both at the beginning and the end of the list. true

רדיס תומך ברשימות שאפשר להכניס ערכים גם בהתחלה וגם בסוף.

1. Redis has a feature that when used, an item is deleted automatically after a given number of seconds. true

ברדיס יש פקודה שמאפשרת למחוק איברים אוטומטית לאחר מספר שניות.

4 Cassandra (10 pt)

Given the following two tables:

נתונות שתי הטבלאות הבאות:

T1 defined as: CREATE TABLE T1 (A INT, B INT, C INT, D INT, PRIMARY KEY((A, B), C, D));

And:

T2 defined as: CREATE TABLE T2 (A INT, B INT, C INT, D INT, PRIMARY KEY(A, B, C));

For each of the following CQL queries, determine whether it is legal, illegal (e.g. requires ALLOW\_FILTERING) or syntax error (e.g. “EAT C WITH T1”):

לכל אחת מהשאילתות הבאות כיתבו האם היא חוקית, לא חוקית, או בכלל לא בסינטאקס חוקי לCQL.

1. SELECT B FROM T1 WHERE A=345 \_\_illegal\_\_
2. SELECT C FROM T2 WHERE A=354 \_legal\_\_\_
3. SELECT D FROM T1 WHERE A=345 AND B=756 \_\_legal\_\_\_
4. SELECT \* FROM T2 WHERE A=345 INNER JOIN T1 ON T1.C=T2.C \_\_syntax\_\_\_
5. SELECT \* FROM T1 WHERE C<100 AND A=765 AND B=192 \_\_legal\_\_

5 Neo4J (8 pt)

Write a query in Cypher that returns all friends of Gal Gazit, all friends of friends of Gal Gazit and all friends of friends of friends of Gal Gazit, that haved WATCHED the movie Harry Potter.

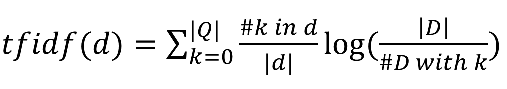
(Assume FRIEND and WATCHED are relations between nodes in the graph.)

כתבו שאילתא שמחזירה את כל החברים של גל גזית, כל החברים שלהם וכל החברים שלהם שצפו בסרט הארי פוטר.

match (n)-[:friend\*..3]->(m:Person{name: “Gal Gazit”}),(n)-[:watched]-(p:Movie{name:”Harry Potter”}) return distinct n

6 TF-IDF (7 pt)

Rank the TF-IDF scores of the following documents given the following query (no calculator is required):

Recall that the TF-IDF formula is:

דרגו את המשפטים הבאים לפי ניקוד הטי אף איי די אף.

Q: apples and apes

D1: **apes and** monkeys eat bananas

D2: a man walked down the street **and** saw an elephant.

D3: monkeys eat **apples and** bananas

D4: **apes** like to eat **apples**

D5: would you like to eat **apples and** bananas?

First: \_\_\_\_D4\_\_\_\_\_\_\_

Second: \_\_\_D1\_\_\_\_\_\_

Third: \_\_\_\_\_\_D3\_\_\_\_

Forth: \_\_\_D5\_\_\_\_\_\_\_\_\_\_  
Fifth: \_\_\_\_D2\_\_\_\_\_\_\_\_

7 XML (5 pt):

The following XML is NOT well formed, state *all* the reasons why:

כתבו את כל הסיבות מדוע האקס אם אל הבא אינו תקין:

<cleaning importance=”high”>

  <kitchen>

  <bath>

  </kitchen>

  </bath>

<cleaning>

<shoppinglist>

  <apples />

  <elephants>

</shoppinglist>

no closing tag for <cleaning> (there is another open tag and no closing tag for both of them), <bath> is declared before </kitchen> and closed after it, <elephants> has no closing tag\_

8 JAVA Streams (10 pt):

In class we have learned the following Java Streams function:

recude([identity], accumulator, [combiner])

Recall that:

.count() is equivalent to: .reduce(0, (x,t)-> x+1, (x,y) -> x+y)

Write a Java streams **reduce** function that given a stream of non-zero numbers will return the following result:

(in words: you need to multiply the inverse of all numbers in the stream. E.g. {1, 2, 2, 3} -> )

You may only use a reduce function, not any other functions (i.e. no map, filter, etc.)

כתבו פונקצית רדיוס בג'אוה סטרימז שמקבלת מספרים ומחזירה את המכפלה של כל המספרים ההופכיים להם.

\_reduce(1, (x,t)->1/x, (x,y)->x\*y)

9 Spark (10 pt):

When applying the list function on a string in python, we get a list composed of all characters of the string, e.g. list(“Hi a”) = [‘H’,’i’,’ ‘,’a’].

The following code will print the count of all ***characters*** in text\_file:

count = text\_file.flatMap(lambda s: list(s)) \

.map(lambda x: (x, 1)) \

.reduceByKey(lambda a, b: a + b)

print(count.sortByKey(False).collect())

1. How would you modify the code so that it ignores all spaces (3 pt)?

איך תשנו את הקוד למעלה כדי שיתעלם מכל הרווחים?

count = text\_file.flatMap(lambda s: list(s))

.map(lambda x: (x, 1)) .filter(lambda v:v!=’ ‘)

.reduceByKey(lambda a, b: a + b)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. How would you modify the code so that it prints counts of tri-chars, that is counts of every 3 consecutive characters. E.g. the following sentence: “hello world”, has the following tri-chars (ignoring the space): [“hel”, “ell”, “llo”, “low “, “owo”, “wor”, “orl”, “rld”)? (5 pt)

Hint: in class we have seen the following function that creates bigrams from a sentence:

def bigram(line):

   words = line.split()

   return zip(words, words[1:])

איך תשנו את הקוד למעלה כדי שיספור את כל התלת-תווים.

from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local[\*]").setAppName("Lab 6")

sc = SparkContext(conf=conf)

rawData = sc.textFile("text\_file.txt")

count = rawData.map(lambda s: s.replace(' ','')).flatMap(lambda s: zip(list(s),list(s)[1:],list(s)[2:])).map(lambda x: (''.join(x),1)).reduceByKey(lambda x,y: x+y)

print(count.sortByKey(False).collect())

1. How would you modify the code so that it prints the tri-chars that are most common first (2 pt)?

איך תשנו את הקוד למעלה כדי שיציג קודם את התלת-תווים הכי שכיחים.

print(count.sortBy(lambda x: x[1], ascending=False).collect())

10 Naive Bayes (6 pt):

Which of the following sentences is **true**:

1. According to the Naive Bayes assumption the probability of any word in a message (xti) **depends** only on other words that appear in that message (xtj). False

לפי ההנחה של נאיב בייס ההסתברות של כל מילה בהודעה תלויה רק במילים אחרות שמופיעות בהודעה.

1. According to the Naive Bayes assumption the probability of any word in a message (xti) **depends** only on the class of the message (yt). True

לפי ההנחה של נאיב בייס ההסתברות של כל מילה בהודעה תלויה רק במחלקה של המשפט.

1. According to the Naive Bayes assumption the probability of any word in a message (xti) **depends** both on the words that appear in that message (xtj) and the class of the message (yt). False

לפי ההנחה של נאיב בייס ההסתברות של כל מילה בהודעה תלויה גם במילים אחרות שמופיעות בהודעה וגם במחלקה.

1. According to the Naive Bayes assumption the probability of any word in a message (xti) is **independent** of both the words that appear in that message (xtj) and the class of the message (yt). False

לפי ההנחה של נאיב בייס ההסתברות של כל מילה בהודעה בלתי תלויה במילים אחרות שמופיעות בהודעה ובלתי תלוי במחלקה.

11 Linear regression (8 pt)

Recall that for linear regression, we have defined the following loss function

https://lh6.googleusercontent.com/3XpcoJ4CjAchf3zlaoBtbvtlgsTaT7XRBnua7zwOoQnEJlPjhdKJxuDh7kGJ8BHP0y9nuD_nRYtasoBXciRjU4c_PvwtADrhhDbnh2kJcwvYrmF9c2giMMan6_-FBZauV6pgHM5E

Recall that the hypothesis in linear regression is: h(x) = xw + b

Given the following data: x = [3, 4, -2, -3] and y = [1, 2, -1, -2]

1. What is the loss when w=0 and b=0 (4 pt)?

מה הטעות כאשר w=0,b=0?

1. Find w and b that will yield a lower loss (not necessarily minimal) (4 pt).

מצא w וb עם טעות קטנה יותר.

12 Logistic regression (6 pt)

Which of the following problems would you use logistic regression to solve (only a single answer is correct):

איזה מאופציות הבאות מתאים ללוג'יסטיק רגרשיין:

1. Predicting house price given house size and location.

ניבוי מחיר בית בהינתן גודל הבית ומיקומו.

1. Predicting house size given house price and location.

ניבוי גודל הבית בהינתן מחיר הבית ומיקומו.

1. Predicting package arrival date and time given source and target locations, shipping date and time, and the package size.

ניבוי זמן ההגעה של חבילה בהינתן המקור ממנו נשלח והמקום אליו נשלח, זמן המשלוח וגודל החבילה.

1. Predicting whether a Bachelor’s student will apply to a master’s program, given transcript grades. True since it’s a classification problem

ניבוי האם סטודנט/ית תואר ראשון יגיש/תגיש מועמדות לתואר שני בהינתן ציוניו/ה.