אוניברסיטת אריאל בשומרון
המחלקה למדעי המחשב

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.

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

ב ה צ ל ח ה!

03.9066640
03.9066692
אריאל, 40700
אוניברסיטת אריאל בשומרון ע"ר

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 ש dayAsInt 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 dayAsInt, count (*) as total_women
FROM Patients p JOIN TreatmentForPatient tp
ON p.patientId = tp.patientId
WHERE gender = 1
GROUP BY dayAsInt
```

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

```
SELECT dayAsInt, sum(price) as total_revenue
FROM  TreatmentForPatient tp JOIN TreatmentTypes tt
ON  tp.treatmentTypeId = tt.treatmentTypeId
GROUP BY dayAsInt
HAVING total_revenue >= 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?

{E, F, B}, {E, F, A} are the two candidate-keys.

The “algorithm” to find them is the following:

First verify which attribute(s) can’t be determined by other attribute(s): In our case E, and F.

Second, verify which attribute(s) can we determined with the previous attribute(s) (EF): In our case D.

Finally, since it’s left only 3 attributes verify all the possibilities ({E, F, A}, {E, F, B}, {E, F, C})

and check if the possibilities are candidate-key or not.

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?

The non-prime attributes are C D (To answer to the question, we HAVE TO search which attributes are non-prime).

Assuming 1Nf, we first check 2NF:

It’s enough to show that one of the non-prime attribute depends of a PROPER subset of one of the CANDIDATE-KEYS. In our case we have : D depends on {E, F}. Obviously, D is a non-

prime attribute and {E, F} is a proper subset of the candidate-key {E, F, A}, which close the claim.

The, the level of the normal form is 1NF.

NO-SQL

3 Redis (5 pt)

Which of the following statements is FALSE:

איזה מבין המשפטים הבאים **אינו** נכון:

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

ברדיס אפשר ליצור כניסה שהיא עצמה טבלת גיבוב.

b. 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).

ברדיס יש פקודות לאחזור חלקים ממסמך.

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

רדיס תומך ברשימות שאפשר להכניס ערכים גם בהתחלה וגם בסוף.

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

ברדיס יש פקודה שמאפשרת למחוק איברים אוטומטית לאחר מספר שניות.

Explanation: The best way to answer to that question is by elimination, it's reduce the possibility of a wrong answer.

First, the "a" assumption is not false since in REDIS it's possible to use the command HSET as HSET user:302 first_name "Tamar" (from the slides).

The "c" assumption is also true, as we can show by providing the next legal command:

RPUSH user:571:items "water" (RPUSH ~ Right push, then we push the new item at the end).

RPUSH do the same but with the beginning of the list.

The "d" assumption isn't false since it's possible in REDIS to use the EXPIRE command which emit to delete an item after a given time.

The the good answer (which is actually the false assumption) is "b".

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 ל CQL.

a. SELECT B FROM T1 WHERE A=345 illegal
b. SELECT C FROM T2 WHERE A=354 legal
c. SELECT D FROM T1 WHERE A=345 AND B=756 legal
d. SELECT * FROM T2 WHERE A=345 INNER JOIN T1 ON T1.C=T2.C syntax
e. SELECT * FROM T1 WHERE C<100 AND A=765 AND B=192 legal

Explanation:

a) In T1, the partition key are A and B and THEY MUST APPEAR in each query, with an equal “=” statement. Trivially, in the query the B key is not mentioned.

b) In T2, only A is include on the partition key, and A appear with an equal statement. Since nothing else is doing here, the query is perfectly legal.

c) Since both of attribute in the partition key are mentioned with an equal statement, the query is legal.

d) The is no INNER JOIN in CASSANDRA as well as UNION.

e) Since both of attribute in the partition key are mentioned with an equal statement, adding the FIRST clustering key (it’s possible either an = or an <,> statement) the query is 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 (a:Person)-[:friend1..3]→(b:Person {name = “Gal Gazit”} COLLECT(a) as friend_group WHERE ALL (x in friend_group WHERE (friend_group)- [:watched]→(d:Movie {name : “Harry Potter”})) return friend_group

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: $tfidf(d) = \sum_{k=0}^{|Q|} \frac{\#k\ in\ d}{|d|} \log(\frac{|D|}{\#D\ with\ k})$

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
Explanation:
You can provide two ways for the answer.
First way:
By |sentence| we mean the number of word in the sentence.

By |word| we mean the number of time the the word (in the question Q) appear in the sentences (D1, … ,D5).

By D1 → word (s)we mean that in D1 there are the word(s).

We next count the data for each sentence:

|D1| = 5 |D2| = 10 |D3| = 5 |D4| = 5 |D5| = 8

|apples| = 3 |apes| = 2 |and| = 4

(about the punctuation, I have no idea how this is work, but I think there is nothing to worried about cause the test question will not ask about).

Now sort the results as : (ascendant sorting)

|D1| = |D3| = |D4| < |D5| < |D2|

|apes| < |apples| < |and|

Then we already know which sentence is the forth (D5) and last (D2).

We need to class D1, D3 AND D4.

D1→ |apes| + |and| = 6 D3 → |apples| + |and| = 7 D4 → |apes| + |apples| = 5

sort the result as : (ascendant sorting)

D4 < D1 < D3

We're done.

For the second solution, please use the formula.

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>
```

1. bath is opened after kitchen, but he is closed before kitchen.

2. cleaning must be closed and the end of the document cause XML authorize only one root element.

3. Same for elephant, which is opened but never closed.

ATTENTION, closing apples which wasn't open previously is not a problem, but only some useless code.

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: $\prod_{i=0}^{N} \frac{1}{x_i}$

(in words: you need to multiply the inverse of all numbers in the stream. E.g. {1, 2, 2, 3} -> $\frac{1}{12}$)

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.

Assuming we initialize a list as follow:

```
List<Double> myList = Arrays.asList{1.0, 2.0, 2.0, 3.0};
```

We can now use a stream:

```
myList.stream().reduce(1.0,  (x, y) → x * 1 / 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())
```

a. 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)) \
        .filter(lambda a, a != " ") \
        .map(lambda x: (x, 1)) \
        .reduceByKey(lambda a, b: a + b)
print(count.sortByKey(False).collect())
```

b. 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.

```
count = text_file.flatMap(lambda s: zip(list(s), zip(list(s)[1:], zip(list(s)[2:])) \
        .filter(lambda a, a != " ") \
        .map(lambda x: (x, 1)) \
        .reduceByKey(lambda a, b: a + b)
print(count.sortByKey(False).collect())
```

c. 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.

```
count = text_file.flatMap(lambda s: list(s)) \
        .map(lambda x: (x, 1)) \
        .reduceByKey(lambda a, b: a + b) \
        .map(lambda (x, y) : (y, x)) \
        .sortByKey(False) \
```

```
.map(lambda (x, y) : (y, x)
```

10 Naive Bayes (6 pt):

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

a. 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).

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

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

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

c. 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).

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

d. 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).

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

Explanation:

From the slides we have seen:

To prepare multiple queries you need to:

… count how many documents every CLASS have (so it's already depends on the class)...

Why it's not also depends on the other words of the message?

Because we only use the words of the dictionary, which is different of the other words of the message since in the dictionary are included words from all the messages.

11 Linear regression (8 pt)

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

$J(w,b) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$

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]

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

מה הטעות כאשר w=0,b=0?

The error may be calculate by using the formula of the loss function:

Recall h(xi) = w*xi + b

As w = b = 0, obviously h(xi) = 0

Since m =4, we have (y1^2 + y2^2 + y3^2 + y4^2) / (2 * 4)

(1 + 4 + 1 + 4) / 8 = 10 / 8 = 5/4, which is our answer.

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

מצא w ו b עם טעות קטנה יותר.

We need to find either w or b different that 0 in which the loss will be shorter.

As we can notice, the vector y is perfectly symmetric (-1_1, -2_2), which means that every change of the b parameter will not have any effect.

We have to change the w parameter, let explain how to choose it:

Recall our goal, we want to reduce the loss. We calculate the loss with the function above.

We want to find a w such that (h(xi) – y) will be shorter than previously.

Since we will not apply any change in b, we actually need to find a proportionnal relation between the vector x and y. We can easily see that 0.5 x – y is short than y.

Let's calculate:

(0.5^2 + 0^2 + 0^2 + 0.5^2) / 8 = 0.5/8 = 1/16.

We're good since 1 / 16 < 5 / 4.

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:

a. Predicting house price given house size and location.

ניבוי מחיר בית בהינתן גודל הבית ומיקומו.

b. Predicting house size given house price and location.

ניבוי גודל הבית בהינתן מחיר הבית ומיקומו.

c. 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.

d. Predicting whether a Bachelor's student will apply to a master's program, given transcript grades.

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

Explanation:

The logistic regression is used for binary choice: either 0 or 1.

The only answer which is a binary choice is the d: either the student will apply to a master's program or not.