**3.1.2019**

**Bilgisayar Mühendisliği Bölümü**

**Büyük Veri Analitiği Dersi Final Sınavı**

**Sınav Süresi: 1 gün.**

**Sınav evrakları 3.1.2019 tarihi saat 18.30'da dersin yapıldığı salonda imza karşılığı teslim edilecektir. Geç getirilen sınav evrakları kabul edilmeyecek ve sınava girmemiş olarak kabul görecektir.**

1.a.

Please write out the output for the following codes at the marked locations.

```
val lines = sc.parallelize(List("hello world", "this is a scala program", "to create a pair
RDD", "in spark"))
val pairs = lines.map(x => (x.split(" ")(0), x))
pairs.filter {case (key, value) => key.length <3}.foreach(println)
```

Location A: What is the output here?

b.

```
val pairs = sc.parallelize(List((1, 2), (3, 4), (3, 6)))
val pairs1 = pairs.reduceByKey((x,y) => x*y)
pairs1.foreach(println)
```

Location B: What is the output here?

2.

Suppose you have two files, “one.txt” and “two.txt”. The content of “one.txt” is:

Donald John Trump (born June 14, 1946) is an American real estate developer, television personality, business author and political candidate. He is the chairman and president of The Trump Organization, and the founder of Trump Entertainment Resorts.[1] Trump's career, branding efforts, lifestyle and outspoken manner helped make him a celebrity, a status amplified by the success of his NBC reality show, The Apprentice.[2][2] Trump is a son of Fred Trump, a New York City real estate developer.[9] Donald Trump worked for his father's firm, Elizabeth Trump & Son, while attending the Wharton School of the University of Pennsylvania, and officially joined the company in 1968.[10] In 1971, he was given control of the company, renaming it The Trump Organization.[11][12] Trump remains a major figure in American real estate and a celebrity for his prominent media exposures.[13] On June 16, 2015, Trump formally announced his candidacy for president of the United States in the 2016 election, seeking the nomination of the Republican Party.[14][15] Trump's early campaigning drew intense media coverage and saw him rise to high levels of popular support.[16] Since late July 2015, he has consistently been the front-runner in public opinion polls for the Republican Party nomination.[17][18][19]

The content of “two.txt” is:

Hillary Diane Rodham Clinton (born October 26, 1947) is an American politician who served as the 67th United States Secretary of State under President Barack Obama from 2009 to 2013. The wife of Bill Clinton, the 42nd President of the United States, she was First Lady of the United States during his tenure from 1993 to 2001. She served as a United States Senator from New York from 2001 to 2009. An Illinois native, Hillary Rodham graduated from Wellesley College in 1969, where she became the first student commencement speaker, then earned her J.D. from Yale Law School in 1973. After a stint as a Congressional legal counsel, she moved to Arkansas, marrying Bill Clinton in 1975. She co-founded Arkansas Advocates for Children and Families in 1977, became the first female chair of the Legal Services Corporation in 1978, and was named the first female partner at Rose Law Firm in 1979. The National Law Journal twice listed her as one of the hundred most influential lawyers in America. While First Lady of Arkansas from 1979 to 1981 and 1983 to 1992, she led a task force that reformed Arkansas' education system, while sitting on the board of directors of Wal-Mart, among other corporations. As First Lady of the United States, her major initiative, the Clinton health care plan of 1993, failed to reach a vote in Congress. In 1997 and 1999, she played a leading role in advocating the creation of the State Children's Health Insurance Program, the Adoption and Safe Families Act and the Foster Care Independence Act.

Write a standalone program called ‘DonaldClinton.scala’, which prints out the words that appear in both files and their word counts, with the words sorted by their counts in descending order.

3.

Consider the following query (yes, substr does what you think it does):

```
select * from r,s
where substr(r.lastName,10) = substring(s.lastName,10);
```

Describe two different query plans that might be generated by the compiler to execute this query. In other words, with respect to the relational algebra portion of the query (not the string munging), please describe solutions to the problem using two different data structures or ways of solving the problem generally available within a database, e.g. hashing, indexing, merging, etc.

4.

Consider as input files that look as follows.

**List**

| | |
|---|---|
| parke | ice cream |
| parke | tofu |
| parke | tomatoes |
| jeff | broccoli |
| jeff | hummus |
| jeff | ice cream |
| eric | milk |
| ⋮ | |

**Price**

| | |
|---|---|
| broccoli | 3.99 |
| hummus | 4.50 |
| ice cream | 6.25 |
| milk | 4.99 |
| tofu | 3.59 |
| tomatoes | 5.00 |
| ⋮ | |

Input **List** represents grocery lists for people. Input **Price** yields the price for each grocery item. Assume the format of each is *key-value*.

(a) [4pt] Design a *MapReduce* job that outputs grocery *item* (e.g., "ice cream") as *key* and the *count* of the number of *people* as *value* (e.g., 2) who have that item on their list.

Make clear your *map* and *reduce* procedure pairs:

- the input of each *map*;
- the output key-value of each *map*;
- the output key-value of each *reduce*; *and*
- simple pseudo-code / clear description of what each *map* and *reduce* does.

(b) [4pt] Design a *MapReduce* job that outputs the total cost—sum of prices—for each person's grocery list. E.g.,

| | |
|---|---|
| parke | 14.84 |
| jeff | 14.74 |
| eric | 4.99 |
| ⋮ | |

Do *not* assume that the *values* are sorted by the shuffle; unlike in Project #2, your platform here *cannot* provide this. Note that a *map* can take its input from *more* than directory. Describe your *MapReduce* job as in Question 1a. You may assume that you have the output from the job in Question 1a in a directory *Count*. Be concerned that nearly everyone may have, say, *ice cream* on their list. (*Hint:* Think about *ice cream,1*, *ice cream,2*, . . . , *ice cream,N* as keys.)

(c) [2pt] Can *map* and *reduce* filter keys, to eliminate keys that do not meet some condition?

Briefly explain why or why not.

5.

Consider three users *u1*, *u2*, and *u3*, and four movies *m1*, *m2*, *m3*, and *m4*. The users rated the movies using a 4-point scale: -1: bad, 1: fair, 2: good, and 3: great. A rating of 0 means that the user did not rate the movie.

The three users' ratings for the four movies are:

*u1* = (3, 0, 0, -1), *u2* = (2, -1, 0, 3), *u3* = (3, 0, 3, 1)

a. Which user has more similar taste to *u1* based on cosine similarity, *u2* or *u3*? Show detailed calculation process.
b. User *u1* has not yet watched movies *m2* and *m3*. Which movie(s) are you going to recommend to user *u1*, based on the user-based collaborative filtering approach? Justify your answer.

6. Vector is a data structure used in Spark MLLib to store the features for data.
Assume the following libraries are available in your code:
import org.apache.spark.mllib.linalg.Vectors

a. Write Spark code to create 5 dense vectors (0.0, 1.0), (-1.0, 0.2), (1.0, 2.5), (3.0, 4.0), and (4.0, 5.0).
b. The five vectors represent five data points. The data points can be drawn in a twodimensional plot. Mark the data points using crosses in the following graph.
c. Assume a linear regression model for the data points. Draw a line that fits the best to all data points in the graph.