

Projects of using Spark

Deadlines:

We have two deadlines:

(1) Deadline 1: May 29 (5:00 pm) 2018. you need to submit the results of E1 -- E5, N1 -- N5.

(2) Deadline 2: June 12 (5:00 pm) 2018. You need to submit the results of N6 -- N7, H1 -- H4.

That means you need to write two reports. Your new report can use some sentences from your old report. Send the reports to syb3181@gmail.com, 14210240021@fudan.edu.cn

1 Introduction

1.1 Collaboration Policy

You are not allowed to work in a group. This project should be done by your own. You will be graded on the creativity of your solutions, and the clarity with which you are able to explain them. If your solution does not live up to your expectations, then you should explain why and provide some ideas on how to improve it. You are free to use any third-party ideas or codes that you wish as long as it is publicly available. You must provide references to any work that is not your own in the write-up. BUT, THE WHOLE ALGORITHMS MUST BE DONE ON SPARK PLATFORM.

1.2 Writing Policy

The final report should be written in English. The main components of the report will cover

1. Introduction to background and potential applications (2%);
2. Algorithms and critical codes in a nutshell (10%);
3. Experimental analysis and discussion of proposed methodology (8%).

1.3 Submitting Policy

The paper must be in NIPS format (downloadable from <https://nips.cc/Conferences/2016/PaperInformation/StyleFiles>).

We donot need the double blind review.

Package your code and a copy of the write-up pdf document into a zip or tar.gz file called Project-*your-student-id1.[zip|tar.gz]. Also include functions and scripts that you had used. The codes should be almost directly runnable on our servers. **We will release the instruction about how to use the servers as well as the password.** 我们会代码查重；发现代码抄袭，后果严重。
Any question, feel free to drop an email to TA.

1.4 Evaluation of Final Projects

The paper is reviewed as the following NIPS criteria:

Overview:

you should briefly summarize the main content of this paper, as well as the Pros and Cons (advantages and disadvantage) in general. This part aims at showing that you had read and at least understand this paper.

Quality:

Is the paper technically sound? Are claims well-supported by theoretical analysis or experimental results? Is this a complete piece of work, or merely a position paper? Are the authors careful (and honest) about evaluating both the strengths and weaknesses of the work?

Clarity:

Is the paper clearly written? Is it well-organized? (If not, feel free to make suggestions to improve the manuscript.) Does it adequately inform the reader? (A superbly written paper provides enough information for the expert reader to reproduce its results.)

2 Dataset

Who would have imagined that backwards ideologies, cronyism and rising religious extremism in Turkey would lead to a crumbling and vulnerable technical infrastructure? The leaked database however only includes adults of 18 or older and do not includes deceased citizens as of 2009. Mernis database would have over 120 million records whereas the leaked one only have about 48 million. This leak contains the following information for 49,611,709 Turkish citizens:(IN CLEARTEXT)

1. * National Identifier (TC Kimlik No)
2. * First Name
3. * Last Name
4. * Mother's First Name
5. * Father's First Name

6. * Gender
7. * City of Birth
8. * Date of Birth
9. * ID Registration City and District
10. * Full Address

Data schema

Schema =

(uid, national_identifier, first name, last name, mother_first, father_first, gender, birth_city, date_of_birth, id_registration_city, id_registration_district, address_city, address_district, address_neighborhood, street_address, door_or_entrance_number, misc)

0. uid	1. national_identifier	2. first name	3. last name
4. mother_first	5. father_first	6. gender	7. birth_city
8. date_of_birth	9. id_registration_city	10. id_registration_district	11. address_city
12. address_district	13. address_neighborhood	14. street_address	15. door_or_entrance_number

Data Sample:

297107	55711266610	HUSNE GULEC	FATMA ALI K	KULUNCAK	12/6/1988	MALATYA KULUNCAK	MALATYA KULUNCAK	SULTANLI KOYU	KOYUN KENDISI	13	<NULL>
297108	55726266100	MENDUH GULEC	FATMA MUHUTTIN	E KULUNCAK	15/8/1984	MALATYA KULUNCAK	MALATYA KULUNCAK	SULTANLI KOYU	KOYUN KENDISI	79	<NULL>
297109	55732265982	TEYFIK GULEC	RAZIYE SULEYMAN	E KULUNCAK	1/1/1984	MALATYA KULUNCAK	MALATYA KULUNCAK	SULTANLI KOYU	KOYUN KENDISI	70	<NULL>

Columns are seperated by “\t”

Data Access:

Data is stored on hdfs:

hdfs_url	hdfs://10.190.2.112
data for E and N problems	hdfs://10.190.2.112/data/data_dump.txt
for H problems	
training set	hdfs://10.190.2.112/data/train_set.txt
validation set	hdfs://10.190.2.112/data/val_set.txt
test set	hdfs://10.190.2.112/data/test_set.txt

A Code Snippet For Data Accessing:

```
sc = SparkContext("spark://10.190.2.112:7077", "Task_E0_14210240021")
data = sc.textFile('hdfs://10.190.2.112/data/val_set.txt') \
    .map(lambda x: re.split('\t', x)) \
    .map(lambda record: (record[6], 1)) \
    .reduceByKey(lambda a, b: a + b) \
    .collect()
```

Python version of master and workers is **2.7.X** , please make sure everything goes well before submitting your program to the cluster.

3 Tasks

根据给定数据回答一下问题，并在报告中附上每个问题的程序代码：

- E1. 统计土耳其公民中所有人中年龄最大的男人
- E2. 统计姓名中最常出现的字母
- E3. 统计该国人口的年龄分布，年龄段分（0-18, 19-28, 29-38, 49-55, >60）
- E4. 按月份，统计该国人口生日在每个月上的分布
- E5. 统计一下该国的男女比例，男女人数

- N1. 统计男性，女性最常见的10个姓
- N2. 统计每个城市市民的平均年龄
- N3. 说一下该国平均人口最年轻的5的城市
- N4. 统计一下该国前10大人口城市中，每个城市的前3大姓氏
- N5. 计算一下该国前10大人口城市中，每个城市的人口生日最集中分布的是哪2个月
- N6. 统计一下，该国男、女最为常用的5个名字是什么
- N7. 计算一下该国前10大人口城市中，每个城市的最受欢迎的3个名字是什么

我们提供的数据按照70%，10%，20%的比例分为训练集、验证集和测试集。（数据已经预先分好了，并上传到hdfs上）

H1. 构建人所在城市预测模型：根据给定一个人的所有信息（除了所在城市），预测该人所在的城市。分析该模型Top1到 Top 5的预测准确度。

H2. 性别预测模型：根据给定一个人的信息（除了性别），能否出该人的性别？

H3. 身份ID预测模型：我们知道我们国家的身份证ID是跟城市、出生年月相关的，我们希望建立一个模型，根据给定一个人的所有信息（除了ID信息），能否预测出该人的身份ID的某些位数（预测出得ID位数越多越好）。

H4. 姓名预测模型：假设给定一个人的所有信息（除了姓名），能否预测该人最可能的姓氏？分析该模型Top1到 Top 5的预测准确度。

