



به نام خداوند رنگین کمان

گزارش کار تمرین دوم داده های حجیم

امیرحسین نجفی ۴۰۱۱۲۸۹۴

محمد محمدی ۴۰۱۱۴۷۲۴

استاد مربوطه : دکتر خواسته

دانشگاه صنعتی خواجه نصیرالدین طوسی

قبل از بیان شرح ارائه و مراحل انجام کار لازم به ذکر می باشد که جای دانشجوی در زندان نیست و تمام دانشجوی های زندانی باید هرچه سریعتر به عرصه دانشگاه بازگردند.

برای آزادی دانشجویان : **آراد رستمزاد ، علیرضا پورصفر ، امیررضا محمدی ، محمدامین افتخاری ، محمدحسین رحمتی ، عرفان زارعی ، آرین حیدری ، امیرحسین گرشاسبی ، ریحانه هادی پور ، کیمیا کریمی و علیرضا میرعلی**

ایرانی آزاده ، جهان چشم به راه است	ایران کهن در خطر افتاده خبر شو
مُشتی خس و خارند ، به یک شعله بسوزان	بر ظلمت این شام سیه فام، سحر شو

در ابتدا به مراحل نصب هدوپ می پردازیم:

۱ - ابتدا لازم است جاوا را با این دستور نصب کنیم :

```
sudo apt install default-jdk
```

۲ - سپس دانلود هدوپ با دستور زیر انجام میدهیم :

```
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
```

۳ - حال محتویات دانلود شده اکسترکت کرده و به دایرکتوری مورد نظر انتقال میدهیم :

```
tar -xzvf hadoop-3.3.1.tar.gz
```

```
sudo mv hadoop-3.3.1 /usr/local/hadoop
```

۴ - در این مرحله نیاز است که کانفیگ های لازم را جهت معرفی کردن جاوا به هدوپ را انجام دهیم که برای این کار لازم است این

فایل `hadoop/etc/hadoop/hadoop-env.sh` را باز کرده و این خط را پیدا میکنیم

و آدرس فایل جاوا در مقابل آن قرار میدهیم :

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

۵ - با اجرای دستور زیر امتحان میکنیم که آیا هدوپ به درستی نصب شده است یا خیر:

```
usr/local/hadoop/bin/hadoop/
```

حال به به مرحله انجام کار و ساخت **mapper** و **reducer** در پایتون میرسیم:

۱ - ابتدا در یک دایرکتوری دو فایل `map.py` , `reduce.py` میسازیم و اجرای دو دستور زیر در آن دایرکتوری فایل های پایتون را به حالت قابل اجرا در بیاوریم :

```
chmod +x ./mapper.py
```

```
chmod +x ./reducer.py
```

=====

۲- حال با اجرای فایل `convert_data_to_right_format.py` فایل های excel را به فرمت txt در میاوریم تا بتوانیم عملیات `map` و `reduce` را بر روی آن فایل انجام دهیم.

=====

۳ - سپس در این دو فایل برنامه های `map` و `reduce` را می نویسیم که در ادامه برای مثال تمرین سوال شماره یک از بخش شماره ۱ را تشریح می کنیم :

I. در این تمرین نوع (name) و سایز (size) پیتزا را به عنوان کلید در نظر میگیریم و تعداد سفارشات (quantity) را به

عنوان مقدار در نظر میگیریم. (خط ۱۲ فایل `part1/E1/mapper.py`)

II. در `reducer` را آنهایی که دارای کلید یکسان هستند در یک جا جمع شده و مقدار آنها در نهایت با هم جمع میشوند. (خط ۳۱

تا ۳۹ فایل `part1/E1/reducer.py`)

III. در نهایت به دستور `sort` این لیست مرتب میشود و اطلاعات به صورت نزولی مرتب شده و نمایش داده میشوند. (خط ۴۶

تا ۴۷ فایل `part1/E1/reducer.py`)

=====

۴ - بعد از اتمام کار نوشتن دستورات `map` , `reduce` باید این دو فایل و فایل دیتا را که در مرحله ۲ به فرمت txt تبدیل کردیم را به ارائه به هدوپ عملیات را شروع کنیم که این کار با دستور زیر انجام میشود :

```
usr/local/hadoop/bin/hadoop jar share/hadoop/tools/lib/hadoop-streaming-3.2.4.jar \
```

```
-file /home/amirhosein/assginment2/E4/mapper.py \
```

```
-mapper /home/amirhosein/assginment2/E4/mapper.py \
```

```
-file /home/amirhosein/assginment2/E4/reducer.py \
```

```
-reducer /home/amirhosein/assginment2/E4/reducer.py \
```

```
-input /home/amirhosein/assignment2/pizza.txt \
```

```
-output /home/amirhossein/assignment/E4/result
```

=====

۵ - بعد از اجرای این دستور یک پوشه به نام **result** در دایرکتوری E1 ساخته میشود که در این دایرکتوری یک فایل به نام

```
[['Vegan', 'medium'] 2890  
['Cheese', 'medium'] 2920  
['Pepperoni', 'large'] 3020  
['Pepperoni', 'medium'] 3120  
['Pepperoni', 'small'] 3305  
['Cheese', 'large'] 3420  
['Vegan', 'large'] 3620  
['Vegan', 'small'] 3755  
['Cheese', 'small'] 4015
```

part-00000 وجود دارد که خروجی مورد نظر ما را نشان میدهد:

منابع :

<https://www.edureka.co/blog/hadoop-streaming-mapreduce-program>

<https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python>

<https://www.digitalocean.com/community/tutorials/how-to-install-hadoop-in-stand-alone-mode-on-ubuntu-20-04>