**CAREER ANALYSIS**

**IN DIFFERENT WORKFIELDS**

Submitted in partial fulfillment of the requirements for the award of degree of

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**PROJECT REPORT**



**Submitted to:**

**DR. ROHIT BAJAJ**

**Submitted By:**

**ACHAL AWASTHI**

**(15BCS2142)**

**SHIKHA MAHAJAN**

**(15BCS2221)**

**ANKITA SINHA**

**(15BCS2240)**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**Chandigarh University, Gharuan**

**Jan-May 2018**

**ACKNOWLEDGEMENT**

I am highly grateful to the HOD CSE, Chandigarh University, Gharuan, for providing this opportunity to carry out the project. I would like to expresses my gratitude to other faculty members of Computer Science & Engineering department CU, Gharuan for providing academic inputs, guidance & encouragement throughout the training period. I would like to express a deep sense of gratitude and thank to Dr. Rohit Bajaj, without whose permission, wise counsel and able guidance, it would have not been possible to pursue my training in this manner. The help rendered by Dr. Rohit Bajaj, Supervisor for experimentation is greatly acknowledged. Finally, I express my indebtedness to all who have directly or indirectly contributed to the successful completion of my industrial training.

**ANKITA SINHA**

**(15BCS2240)**

**ACHAL AWASTHI**

**(15BCS2142)**

**SHIKHA MAHAJAN**

**(15BCS2221)**

## CERTIFICATE

This is to certify that “**ACHAL AWASTHI**, **SHIKHA MAHAJAN**, and **ANKITA SINHA**” have developed the project entitled “**CAREER ANALYSIS”.** They have worked on the project **CAREER ANALYSIS.** They have used IBM Infoshere, VMWare, HADOOP famework, And PIG for the project. Their work is satisfactory.

I wish them all the best for their bright future.

**ABSTRACT**

Since in this very vast world billions of people have there own like and dislikes but among those billions of peoples millions of peoples are also there who have some same qualities Here we make a record regarding that. So This large amount of raw data can be useful to predict the behavior, likes and dislikes of an individual person. Hadoop is one of the best tool options for Social media data analysis and Hadoop works for distributed big data, Streaming data, Time Stamped data, text data etc. This project will analyze the Real Time Data using Hadoop and according which decision-making and strategy planning can be done easily.

Keywords: Hadoop, Pig, career analysis, Hadoop ecosystem.

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| 1. | Waterfall Iterative Model | 6 |
| 2. | Data Flow Diagram Level 0 | 11 |
| 3. | Data Flow Diagram Level 1 | 12 |
| 4. | Data Flow Diagram Level 2 | 13 |
| 5. | Top 10 professions in this data set (OUTPUT) | 14 |
| 6. | Top 5 professions in this data set (OUTPUT) | 15 |
| 7. | Least 5 professions in this data set (OUTPUT) | 16 |

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Topic** | **Page No.** |
|  | **Acknowledgement** |  |
|  | **Certificate** |  |
|  |  |  |
|  | **Abstract** |  |
|  | **List Of Figures** |  |
| **1** | **Introduction** | 1 |
|  |  |  |
| **2** | **Literature Reviews** | 2 |
|  |  |  |
| **3** | **Objective** | 3 |
|  |  |  |
| **4** | **Scope** | 4 |
|  |  |  |
| **5** | **Problem Statement** | 5 |
|  |  |  |
| **6** | **Software Development Methodology** | 6 |
|  |  |  |
| **7** | **System Requirement Specification** | 7 |
| **8** | **Data Flow Diagram** | 10 |
|  |  |  |
| **9** | **Output** | 14 |
|  | **Reference** | 18 |
|  |  |  |

**CHAPTER 1**

**INTRODUCTION**

The amount of data being produced in modern society is growing at an accelerating pace. New problems and possibilities constantly arise from this so-called data explosion. One of the areas where information plays an important role is that of law enforcement. Obviously, the amount of criminal data gives rise to many problems in areas like data storage, data warehousing, data analysis and privacy. Already, numerous technological efforts are underway to gain insights into this information and to extract knowledge from it.

Since in this very vast world billions of people have there own like and dislikes but among those billions of peoples millions of peoples are also there who have some same qualities Here we make a record regarding that. So This large amount of raw data can be useful to predict the behavior, likes and dislikes of an individual person. Hadoop is one of the best tool options for Social media data analysis and Hadoop works for distributed big data, Streaming data, Time Stamped data, text data etc. This project will analyze the Real Time Data using Hadoop and according which decision-making and strategy planning can be done easily.

**CHAPTER 2**

**Literature Review**

In the past years, many works has been released in sentiment analysis. Implementation of sentiment analysis has been carried out for a variety of applications over a wide range of classification algorithms and for varying data size. There exist many possible variants; some of them are discussed in following section

**2.1 Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013.**

Machine learning technologies are widely used in sentiment classification because of their ability to “learn” from the training dataset to predict or support decision making with relatively high accuracy. However, when the dataset is large, some algorithms might not scale up well. In this paper, the authors evaluate the scalability of Naive Bayes classifier (NBC) in large-scale datasets. They have presented a simple and complete system for sentiment mining on large datasets using a Naive Bayes classifier with the Hadoop framework. Instead of using Mahout Library, they implemented NBC to achieve finegrain control of the analysis procedure for a Hadoop implementation.

**2.2 A Content Analysis of Career Development Theory, Research, and Practice 2013** [**Pei‐Chun Hou**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Hou%2C+Pei-Chun)**,** [**Julia F. Kronholz**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Kronholz%2C+Julia+F)**,** [**V. Casey Dozier**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Dozier%2C+V+Casey)**,** [**Mary Catherine McClain**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=McClain%2C+Mary-Catherine)**,** [**Mary Buzzetta**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Buzzetta%2C+Mary)**,** [**Elizabeth K. Pawley**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Pawley%2C+Elizabeth+K)**,** [**Jane T. Finklea**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Finklea%2C+Jane+T)**,** [**Gary W. Peterson**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Peterson%2C+Gary+W)**,** [**Janet G. Lenz**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Lenz%2C+Janet+G)**,** [**Robert C. Reardon**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Reardon%2C+Robert+C)**,** [**Debra S. Osborn**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Osborn%2C+Debra+S)**,** [**Seth C. W. Hayden**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Hayden%2C+Seth+C+W)**,** [**Gloria P. Colvin**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Colvin%2C+Gloria+P)**,** [**Emily L. Kennelly**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Kennelly%2C+Emily+L)

This career counseling and development literature presents a content analysis of refereed journal articles published in 2013. Four research questions guided the analysis: (a) What content topics were included in career development articles published in refereed journals in 2013 (b) To what extent are theory, research, and practice integrated in career development articles published in refereed journals in 2013 (c) What variation exists in the characteristics of career development articles published in refereed journals in 2013 and (d) What variation exists in the content included in theory, research, and practice articles A total of 360 unique topics were identified in 357 articles from 24 journals. Results indicated that topic content in journals evolved slowly with limited integration of theory, research, and practice.

**CHAPTER 3**

**OBJECTIVE**

The main objective of this project is to focus on the career predication of a person based on some simple questions which are asked to him and the answer given by the person are matched with our database where data of billions of peoples behavior ,their likes and dislikes are recorded. Now we apply different types of data analysis on our database and try to predict which career field is best for that person.

People use this project to understand which career field is good for them and they can apply on those fields and go ahead .This project will increase the capabilities of peoples as they already know in which working field they are good and then can take that field as there career building. This project also decrease the pressure and stress on people as there are now work in the environment in which they like and performs task in which they are good.

This project can also help in analyzing new emerging trends and knowing about people's changing behavior with time. In addition, people in different nature have different preferences. So our project will help them to build their career in the field which is best for them.

**CHAPTER 4**

**SCOPE**

Career Analysis attests the effectiveness of career counseling. Our career analysis project can support people with career-related challenges. Through performing analyzing on tab’s of data regarding career development and labor markets, We put a person's qualifications, experience, strengths and weakness in a broad perspective while also considering their desired salary, personal hobbies and interests, location, job market and educational possibilities. Through their answers we can additionally support people in gaining a better understanding of what really matters for them personally, how they can plan their careers autonomously, or help them in making tough decisions and getting through times of crisis. Finally, Our project are often capable of supporting peoples in finding suitable placements/ jobs, in working out conflicts with their employers, or finding the support of other helpful services. It is due to these various benefits of career counseling to without wasting money in huge amount.

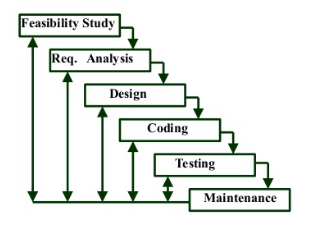
**CHAPTER 5**

**PROBLEM STATEMENT**

In today’s world lot of focus is on the study habits and study processes which improve the knowledge of the students, which are really good. But there are no approaches which mainly concentrate on the mental health of students. Having a specialized consultant at a college level decreases the mental stress by a certain level. In this work student’s choices are taken through a survey from i.e. google form, then the data set is analyzed to make decisions about the problems faced by students. So that steps can be taken to resolve those problems by using ‘*Hadoop and Pig*’. The proposed work is used to demonstrate a workflow of data sense-making for educational purposes, integrating both qualitative analysis and large-scale data analyzing techniques to explore engineering students’ choices in order to understand issues and problems students encounter in their learning experiences and while going for career choices.

**CHAPTER 6**

**SOFTWARE DEVELOPMENT METHODOLOGY**



**Figure: 1**

**WATERFALL ITERATIVE MODEL** contains 6 phases:

* Feasibility study: The feasibility study activity involves the analysis of the problem and collection of the relevant information relating to the product. The main aim of the feasibility study is to determine whether it would be financially and technically feasible to develop the product.
* Requirement analysis and specification: The goal of this phase is to understand the exact requirements of the customer and to document them properly (SRS)
* Design: The goal of this phase is to transform the requirement specification into a structure that is suitable for implementation in some programming language.
* Implementation and unit testing: During this phase the design is implemented. Initially small modules are tested in isolation from rest of the software product. Integration and system testing: In this all the modules are integrated and then tested altogether.
* Operation and maintenance: Release of software inaugurates the operation and life cycle phase of the operation.

**CHAPTER 7**

**SYSTEM REQUIREMENT SPECIFICATIONS**

**HARDWARE SPECIFICATIONS:**

* **PROCESSOR: Intel Core i3**
* **RAM: 4 GB**
* **HARD DISK: 500 GB**
* **MONITOR: LCD Monitor**
* **KEYBOARD: 101 keys or more**
* **MOUSE: Simple Mouse**
* **NETWORK: Internet**

**SOFTWARE SPECIFICATIONS:**

* **PLATFORM: Windows 10**
* **SOFTWARE: VMWARE Work Station**

**IBM Infosphere**

**Hadoop (HDFS and Map Reduce)**

**Pig**

**SOFTWARE SPECIFICATIONS**

1. **VMWARE Work Station:**

**V**Mware Workstation is a hosted [hypervisor](https://en.wikipedia.org/wiki/Hypervisor) that runs on [x64](https://en.wikipedia.org/wiki/X64) versions of Windows and Linux operating systems (an [x86](https://en.wikipedia.org/wiki/X86) version of earlier releases was available); it enables users to set up [virtual machines](https://en.wikipedia.org/wiki/Virtual_machine) (VMs) on a single physical machine, and use them simultaneously along with the actual machine. Each virtual machine can execute its own [operating system](https://en.wikipedia.org/wiki/Operating_system), including versions of [Microsoft Windows](https://en.wikipedia.org/wiki/Microsoft_Windows), [Linux](https://en.wikipedia.org/wiki/Linux), [BSD](https://en.wikipedia.org/wiki/BSD), and [MS-DOS](https://en.wikipedia.org/wiki/MS-DOS). VMware Workstation is developed and sold by [VMware, Inc.](https://en.wikipedia.org/wiki/VMware,_Inc.), a division of [Dell Technologies](https://en.wikipedia.org/wiki/Dell_Technologies). There is a free-of-charge version, VMware Workstation Player, for non-commercial use. An operating systems license is needed to use proprietary ones such as Windows. Ready-made Linux VMs set up for different purposes are available from several sources. VMware Workstation includes the ability to group multiple virtual machines in an inventory folder. The machines in such a folder can then be powered on and powered off as a single object, useful for testing complex client-server environments.

1. **IBM Infosphere:**

IBM Infosphere DataStage is an [ETL](https://en.wikipedia.org/wiki/Extract,_transform,_load) tool and part of the IBM Information Platforms Solutions suite and IBM Info Sphere. It uses a graphical notation to construct data integration solutions and is available in various versions such as the Server Edition, the Enterprise Edition, and the MVS Edition.

1. **Hadoop:**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop **MapReduce** is a software framework for easily writing applications, which process big amounts of data in parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

The **Hadoop Distributed File System (HDFS)** is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data.

1. **PIG:**

**Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in 2007. Pig consists of a language and an execution environment. Pig’s language, called as PigLatin, is a data flow language - this is the kind of language in which you program by connecting things together. Pig can operate on complex data structures, even those that can have levels of nesting. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process the unstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationally complete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requires conditional constructs, an infinite memory model, and looping constructs. PigLatin is not Turing complete on itself, but it can be Turing complete when extended with User-Defined Functions.**

**Execution environment: There are two choices of execution environment: a local environment and distributed environment. A local environment is good for testing when you do not have a full distributed Hadoop environment deployed. You tell Pig to run in the local environment when you start Pig’s command line interpreter by passing it the -x local option. You tell Pig to run in a distributed environment by passing -x mapreduce instead. Alternatively, you can start the Pig command line interpreter without any arguments and it will start it in the distributed environment. There are three different ways to run Pig. You can run your PigLatin code as a script, just by passing the name of your script file to the pig command. You can run it interactively through the grunt command line launched using Pig with no script argument. Finally, you can call into Pig from within Java using Pig’s embedded form.**

**CHAPTER 8**

**DATA FLOW DIAGRAMS**

A graphical tool used to describe and analyze the moment of data through a system manual or automated including the process, stores of data, and delays in the system. Data Flow Diagrams are the central tool and the basis from which other components are developed. The transformation of data from input to output, through processes, may be described logically and independently of the physical components associated with the system. The DFD is also know as a data flow graph or a bubble chart.

DFDs are the model of the proposed system. They clearly should show the requirements on which the new system should be built. Later during design activity this is taken as the basis for drawing the system’s structure charts. The Basic Notation used to create a DFD’s are as follows:

**1. Dataflow:** Data move in a specific direction from an origin to a destination.

**2. Process:** People, procedures, or devices that use or produce (Transform) Data. The physical component is not identified.

**3. Source:** External sources or destination of data, which may be People, programs, organizations or other entities.

**4. Data Store:** Here data are stored or referenced by a process in the System.

* **DATA FLOW DIAGRAM Level 0**

DFD level 0 data flow diagram shows the entire system as a single process, and gives no clues as to its internal organization.

Hadoop

Framework

Classification Result

Career Analysis Data Set

**FIGURE 2**

Data Collection from Form acts as an input where we read the data on the topics from twitter, Facebook, Instagram using hadoop. Hadoop used in designing the system are Open Authentication 1 and Open Authentication 2 along with secret key. Open Authentication 1 is used for registration to access the twitter account, Open Authentication 2 is used to access the account and the secret key is used to support encryption and decryption. The system is responsible for finding the best classification of tweets among five categories, namely Lack of Social Engagement, Career Problem, Heavy Study Load Problems and Negative Emotions. It produces cleaned data, probability, contingency, and enhanced contingency as intermediate results and classification of tweets as final output.

* **DATA FLOW DIAGRAM Level 1**

Data

Probability Computation

Data

Collection

Google Form

Probability

*DATA STORE*

Classification Result

Analyzed

data

Classification Result

**FIGURE 3**

The DFD describes the various modules present in the system and also how the data flows from one module to another to complete the action of understanding tweets description before classification. Career for various cases acts as input to the data collection module.

* **DATA FLOW DIAGRAM Level 2**

Data Collected from Data Set

Comparison

& Compute the Data

Data

Data Set

Compute Data

*Data Store*

Result

Analyzed data

Classification of data into catagories

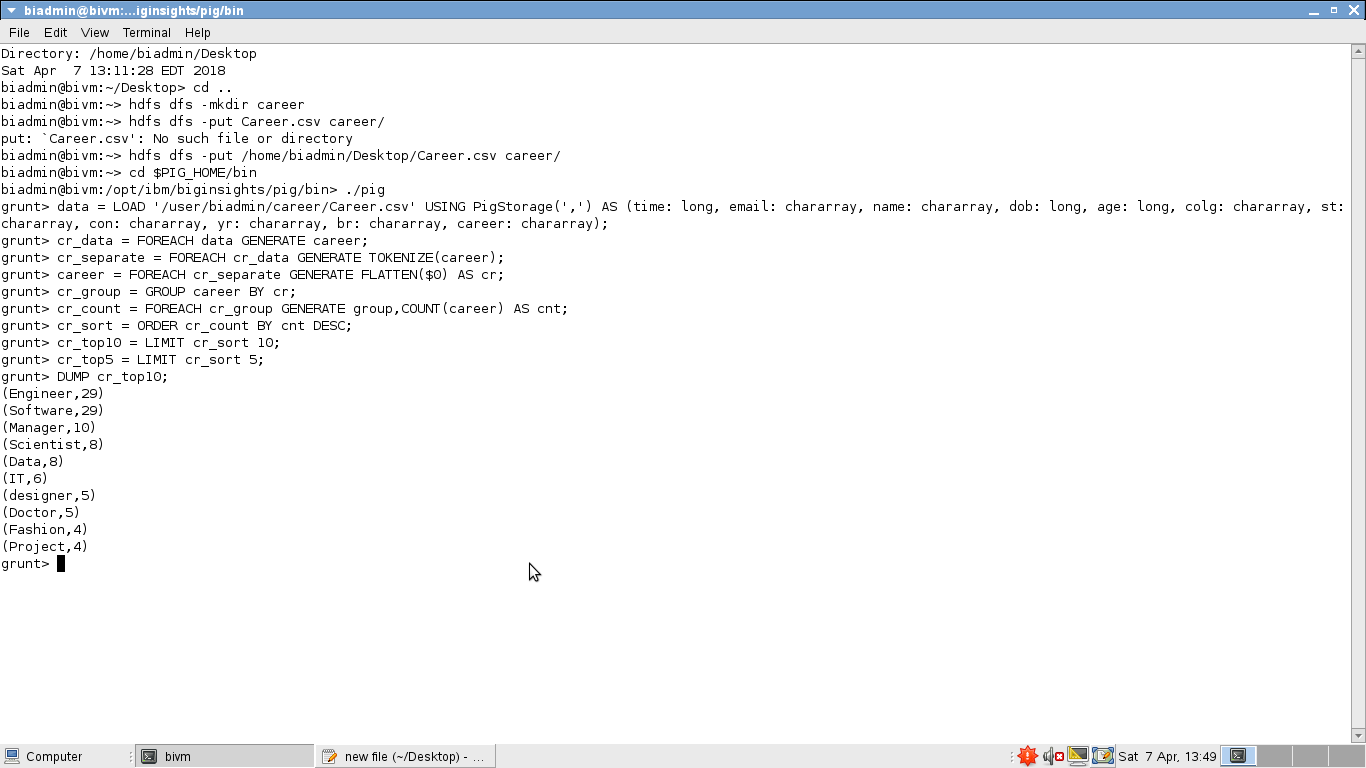
Compute the probability

**FIGURE 4**

**CHAPTER 9**

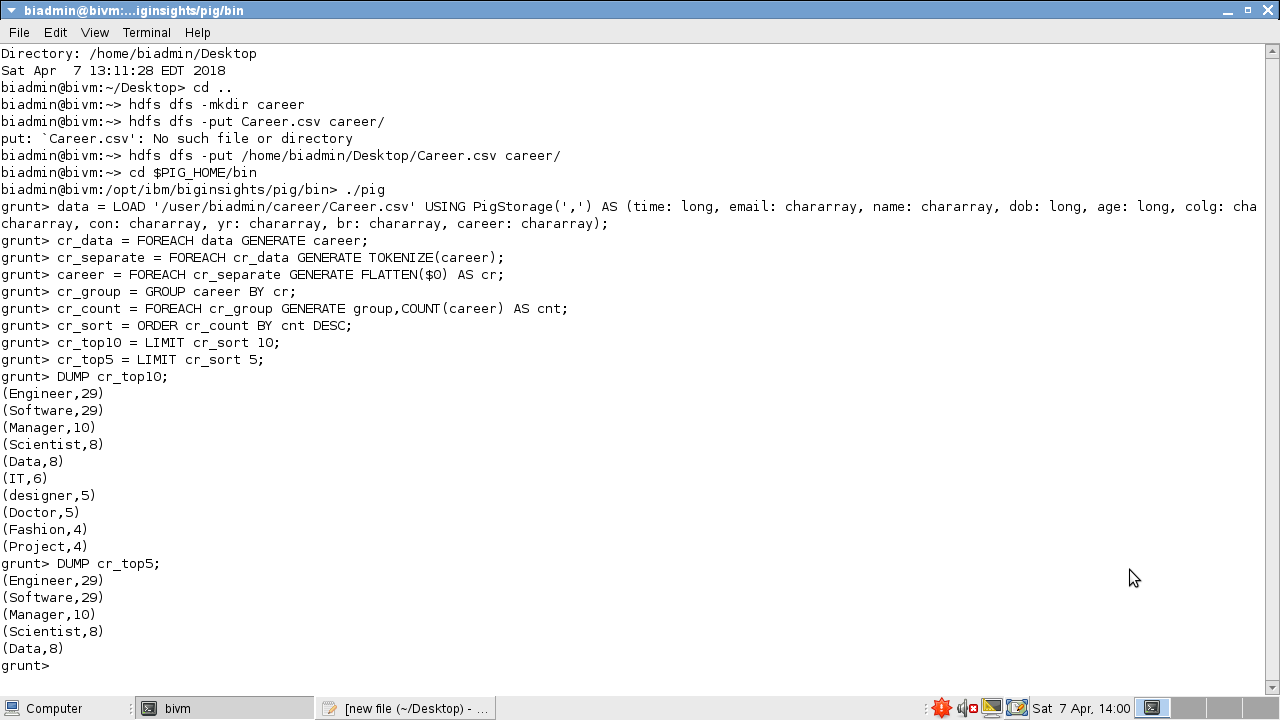
**OUTPUT**

**Question 1: Top 10 professions**



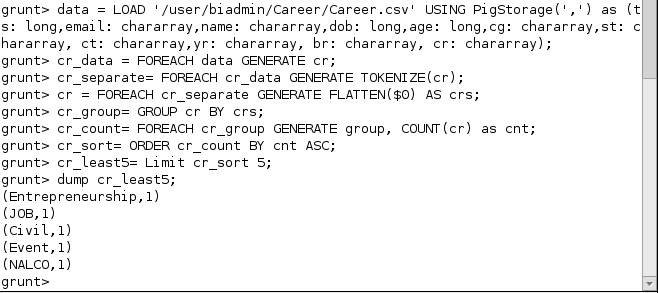
**FIGURE 5**

**Question 2:** Top 5 professions



**FIGURE 6**

**Question 3:** Least 5 professions

****

**Fig 7**

**References**

1. T. Wilson, J. Wiebe and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in Proceedings of HLT and EMNLP. ACL, (2005), pp. 347–354
2. A. Gates, “Programming Pig”, O'Reilley, (2011).
3. T. White, “Hadoop: The Definitive Guide”, Third Edition, O'Reilley, (2012).
4. S. Kumar, F. Morstatter and H. Liu, “Data Analytics”, Springer Science & Business Media, (2013).