NLP Homework 2

1. Dataset Characteristics

Analysis of a subset of a National Science Foundation dataset

The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..." NSF supports fundamental research and education in all the non-medical fields of science and engineering.

Information about research projects that NSF has funded since 1989 can be found by searching the Award Abstracts database. The information includes abstracts that describe the research, and names of principal investigators and their institutions. The database includes both completed and in-process research.

In this assignment, we will analyze a subset of a publicly available (Date Donated: November 18, 2003) collection of National Science Foundation research awards abstracts spanning 1990 - 2003. This data set consists of (a) 129,000 abstracts describing NSF awards for basic research, (b) bag-of-word data files extracted from the abstracts, (c) a list of words for indexing the bag-of-word data. Each abstract is contained in one txt file that is in a sub-folder within a subset of the folders. Datatype of abstracts is text and tabular.

- A sample abstract[Appendix 1] contains (a) Title of an abstract, (b) type of an award and NSF Org., (c) latest amendment date, (d) award number and instructions, (e) Program Manager, (f) start and expiry dates, (g) total expected amount (estimated), etc.…

Using Regular Expressions to Analyze NSF abstracts Data

This homework is mainly designed to help you exercise the power of regular expressions in information searching that we have learned in class.

Pre-processing:-

- The data is about National Science Foundation(NSF) abstract data where there is a list of files which includes the Abstract of the data, NSF organization who provided the award to its respective title, Award amount, Date, period of time of the research work, etc.
- Each file contain the name of Program Manager,Investigator,sponsor of each project.
- Each file has information of its NSF_Program.
- Some of them has Estimated Expire date of its research project, which included as Continuing grant from the organization.
- The NSF_abstract file contained around 4016 text files. One or two file are either empty or the format is different than other files.
- The NSF_abstract data is all about the research work done by the organization people and the details about the research as such when it has started and when its going to end then the thing where is about the grand that is provided in the files with respect to the reseach is continuing for further research or it's a fresh start.
- From this data we can get the information about all the organizations which were involved during 1989 to 2003.
- Most of the project were in future based as their end date was estimated in each text file. It also provided the information about sponsor and NSF Program as it related to which NSF department.
- There is a site on the Internet where you can search for the awards given by the organization. The site is given below. https://www.nsf.gov/awardsearch/showAward?AWD_ID=9702149&HistoricalAwards=false

Python Code Desription:- (Output_part1)

In the first Part Mainly contains 4 parts:-

- File name
- NSF_Org
- Total Amount
- Abstract Data

Here each file has formatted in specific sections. Like wise “Title : “, NSF_Org : “ , etc. So for extracting the data there were two ways.

- First I can use regular expression and serch in the file for the specific value.
- Second I can convert the file content in to dictionary and and from the keys I can get the values.

In this Homework we strictly had to use regular expression for fetch the data so I used the regular expression to get the content I want to fetch.

Here I used to find the **File name** of each file.

```
data=f.read()

File=re.findall('File\s+: a[0-9]*', data)
   if len(File) > 0:
      File_name = re.findall('a[0-9]+', File[0])
      File_name=str(File_name[0])
```

I know the naming convection of each file format in the folder so I accordingly given the regular expression.

NSF_Org:-

```
NSF_Org=re.findall('NSF Org\s+: [A-Z]{3}',data)
   if len(NSF_Org) > 0:
      NSF_name = re.findall(' [A-Z]{3}', NSF_Org[0])
      NSF_name=str(NSF_name[0])
```

The regular expression to fetch the NSF_org name contains just 3 letters and in capital. That was the requirement to fetch for each file for out first file output.

Award Amount:-

```
total_amt_line=re.findall('Total Amt\.\s+: \$[0-9]*', data)
   if len(total_amt_line) > 0:
      total_amt = re.findall("\$[0-9]+", total_amt_line[0])
```

```
    total_amt_int=re.findall("[0-9]+",total_amt_line[0])
    Amount=str(total_amt[0])
```

The regular expression to fetch the data for Award amount It includes the $ symbol and the digits.

Abstract data:-

```
pat_abstract=re.compile('Abstract.*',re.M|re.DOTALL)
  abstract=pat_abstract.findall(data)
  abstract=''.join(abstract)
  abstract=" ".join(abstract.split())
  abstract_content=abstract[11:]
```

This will find the Abstract key word in the text file and fetch the whole data after that.
After getting all the data, I am writing it in to a file to display in the format which was asked in the assignment definition.

Here I am using string slicing to get rid of “Abstact” word from the whole content.

Distribution of sentence lengths :- (Output_part2)

The second part is about the abstract data. In that we have to calculate how many lines are there in abstract data for each file. According to lines, writing the filename and number of lines in an abstract data and the line with its number.

Here I used tokenizer from nltk to fetch the data line by line and according to that I can print in the file.

```
sent_text = nltk.sent_tokenize(abstract_content)

 cnt=len(sent_text)
```

I am talking the length of the toekniser to get write the number of lines at the end of each document.

Here in the output_part 2 file the file name and abstract data line are there with respect to their line numbers and at the end of each text file the number of lines in the abstract is there in text file.

Analysis Part:-

Here we mainly have 3 categories of each file.

- NSF_ Org
- Award Amount
- Abstract data

From the Abstract data we already did the lines by line separation of the whole paragraph.

So we mainly focused on NSF_Org and Award Amount in the text files. There are several senarios to use those two key data from the files. Some of them are listed below:-

- Get the list of awards_amount given by organization.
- Get the list of number of awards given by each organization.
- The maximum award of amount given by each organization.

These are the main three scenarios which we can get from these 2 keyword.

❖ Get the list of awards_amount given by organization.

For this I made a dictionary of NSF_Org and the award_amount.

NSF_Org as Key and Award_amount as Value.

The code to add in to the dictionary :-

```
if dic1.get(NSF_name):
    dic1[NSF_name].append(Amount_int)
else:
    dic1.setdefault(NSF_name, []).append(Amount_int)
```

Here if the NSF_name is already existed in the dictionary then it will just add the value in the respective of its key index.

And if the key is not in the dictionary then it will create the new index for the key and add the value.

The output of this is added in the Appendix.

❖ Get the list of number of awards given by each organization.

I added all the NSF_name in a list and then use the counter from the collections package.

From the result of this list we can get the information who got the maximum awards and who got the lowest.

I used the frequency distribution concept to retrieve this data.

From this we got the information which organization got the maximum awards as per that whose research work was impressive according to NSF.

❖ The maximum award of amount given by each organization.

The dictionary I made to get the NSF_name and the award_amount I used the same for this operation.

I used the **max** keyword to get the maximum value from the list of awards_amount.

The code is given below:-

```
for k , v in dic1.items():

    print (k,max(v))
```

Output of this code is given below:-

```
DEB 4373619
 MCB 567000
 DMS 2027422
 DMI 324395
 OCE 18806079
 CCR 667000
 ATM 1203339
 INT 930233
 IBN 361000
 BCS 410000
 CTS 555000
 CHE 693000
 DMR 2634600
 EAR 400000
 SES 491856
 OPP 3550650
 ANI 338509
 BES 616720
 PHY 6432676
 DBI 1247182
 SRS 1097959
 CMS 924737
 EID 100783
 DAS 2899686
 IIS 688500
 HRM 1069986
 ECS 499230
 AST 485000
 EIA 3670965
 EEC 740000
```

ACI 810000
DOB 649402
NON 1
ENG 151129
LPA 132077
HRD 683416
MIP 188510
BIO 109250
IRM 4000
REC 854642
GEO 50000

From this we got the information as OCE got the maximum award of 18806079.

❖ Extra analysis:-

I did the best I can to get the information from the NSF_name and awards_amount. Yet I felt there should be something more from the files I can get. So I thought one more scenario from the data as delivering the the NSF_Org name with respect to the time period it delivered.

So what I did is get the date when the text document is created and from that I fetch the year and make the dictionary of year and NSF_Org.

The code to get the date from the text files:-

```
pat_year=re.compile('Date.*File',re.M|re.DOTALL)

   Date_term=pat_year.findall(data)

### Converting list to string

Date_term=''.join(Date_term)

 ### Finding the start year. The result of the findall is a list

year=re.findall('[1-2][0-9][0-9][0-9]',Date_term)

if(len(year)>0):

    year_int=int(year[0])

else:

    year_int=123
```

I wrote the year=123 if it couldn't find the date in the text file. Because as I mentioned earlier there are couple of files, which are empty or incomplete, compare to others.

So there are some data which are not useful but we can remove that by filtering data.

```
dic2[year_int].append(NSF_name)
```

```
for k , v in dic2.items():

    print (k,Counter(v))
```

Output of this code will be:-

```
1991 Counter({' DMS': 135, ' OCE': 65, ' EAR': 45, ' SES': 42, ' IIS': 34, ' CCR': 33, ' BCS': 31, ' CMS': 29, ' DEB': 27, ' CHE': 25, ' INT': 23, ' ATM': 20, ' PHY': 19, ' BES': 17, ' IBN': 16, ' DBI': 14, ' CTS': 14, ' DMI': 12, ' ECS': 9, ' ANI': 8, ' MCB': 8, ' HRD': 7, ' EID': 6, ' EIA': 5, ' AST': 5, ' DMR': 4, ' OPP': 3, ' ACI': 2, ' SRS': 2, ' EEC': 1, ' HRM': 1})
1990 Counter({' DMS': 257, ' EAR': 161, ' INT': 151, ' CMS': 130, ' DEB': 119, ' SES': 104, ' DBI': 100, ' IBN': 84, ' CCR': 82, ' BCS': 81, ' CTS': 79, ' MCB': 74, ' ECS': 57, ' OCE': 55, ' CHE': 53, ' HRD': 46, ' PHY': 38, ' DMR': 34, ' BES': 33, ' DMI': 27, ' EID': 26, ' IIS': 22, ' ATM': 20, ' OPP': 16, ' EIA': 15, ' ANI': 14, ' AST': 12, ' ACI': 10, ' ENG': 2, ' IRM': 2, ' MIP': 2, ' SRS': 2, ' NON': 1, ' BIO': 1, ' GEO': 1})
1992 Counter({' DMS': 132, ' CHE': 97, ' MCB': 73, ' IBN': 63, ' OCE': 62, ' DMR': 59, ' DEB': 39, ' PHY': 37, ' CCR': 36, ' ATM': 33, ' CMS': 31, ' CTS': 24, ' EAR': 22, ' IIS': 20, ' BCS': 18, ' SES': 17, ' DBI': 16, ' BES': 12, ' ECS': 11, ' INT': 10, ' DMI': 9, ' AST': 7, ' OPP': 5, ' ANI': 5, ' EIA': 5, ' HRD': 2, ' LPA': 1, ' SRS': 1, ' ACI': 1, ' MIP': 1, ' BIO': 1, ' REC': 1, ' DAS': 1})
1995 Counter({' DMI': 5, ' CHE': 5, ' ATM': 5, ' OCE': 5, ' DEB': 4, ' INT': 4, ' MCB': 3, ' IBN': 3, ' PHY': 3, ' BCS': 1, ' SRS': 1, ' DAS': 1, ' HRM': 1, ' EEC': 1, ' ECS': 1, ' CCR': 1, ' SES': 1, ' EAR': 1, ' HRD': 1, ' DMS': 1, ' DBI': 1})
1993 Counter({' MCB': 33, ' IBN': 29, ' DEB': 28, ' INT': 24, ' CMS': 21, ' DMR': 21, ' OCE': 17, ' BCS': 17, ' CCR': 15, ' SES': 14, ' DBI': 13, ' ATM': 13, ' CHE': 12, ' EAR': 12, ' DMS': 11, ' ECS': 11, ' PHY': 10, ' IIS': 8, ' CTS': 8, ' DMI': 7, ' BES': 5, ' AST': 4, ' HRD': 3, ' DOB': 1, ' ANI': 1, ' ACI': 1, ' EIA': 1, ' SRS': 1})
1989 Counter({' CMS': 8, ' INT': 3, ' DEB': 2, ' BCS': 2, ' EAR': 2, ' OCE': 1, ' SRS': 1, ' DMI': 1, ' PHY': 1, ' CTS': 1, ' BES': 1, ' MCB': 1, ' CCR': 1})
1994 Counter({' CHE': 18, ' DMS': 16, ' OCE': 13, ' IBN': 13, ' BCS': 12, ' MCB': 11, ' DEB': 10, ' ATM': 9, ' INT': 8, ' DMR': 7, ' BES': 4, ' CTS': 4, ' AST': 4, ' CMS': 4, ' EAR': 3, ' SES': 3, ' DMI': 2, ' PHY': 2, ' IIS': 1, ' EEC': 1, ' ECS': 1, ' OPP': 1, ' HRD': 1, ' DBI': 1, ' CCR': 1, ' EIA': 1})
1996 Counter({' DEB': 6, ' DMI': 4, ' OCE': 4, ' ATM': 3, ' DBI': 1, ' EEC': 1, ' INT': 1, ' CMS': 1})
1997 Counter({' EEC': 1, ' DEB': 1})
1998 Counter({' OPP': 1})
```

So from this analysis we can say that the DMS Organization got the most number of awards in 1990 year.

Appendix:- (Screenshots)

1):-

```
Title       : CRB: Genetic Diversity of Endangered Populations of Mysticete Whales:
               Mitochondrial DNA and Historical Demography
Type        : Award
NSF Org     : DEB
Latest
Amendment
Date        : August 1,  1991
File        : a9000006

Award Number: 9000006
Award Instr.: Continuing grant
Prgm Manager: Scott Collins
              DEB  DIVISION OF ENVIRONMENTAL BIOLOGY
              BIO  DIRECT FOR BIOLOGICAL SCIENCES
Start Date  : June 1,  1990
Expires     : November 30,  1992   (Estimated)
Expected
Total Amt.  : $179720             (Estimated)
Investigator: Stephen R. Palumbi   (Principal Investigator current)
Sponsor     : U of Hawaii Manoa
              2530 Dole Street
              Honolulu, HI  968222225    808/956-7800

NSF Program : 1127      SYSTEMATIC & POPULATION BIOLO
Fld Applictn: 0000099   Other Applications NEC
              61        Life Science Biological
Program Ref : 9285,
Abstract    :

              Commercial exploitation over the past two hundred years drove
              the great Mysticete whales to near extinction.  Variation in
              the sizes of populations prior to exploitation, minimal
              population size during exploitation and current population
              sizes permit analyses of the effects of differing levels of
              exploitation on species with different biogeographical
              distributions and life-history characteristics.  Dr. Stephen
              Palumbi at the University of Hawaii will study the genetic
              population structure of three whale species in this context,
              the Humpback Whale, the Gray Whale and the Bowhead Whale.  The
              effect of demographic history will be determined by comparing
              the genetic structure of the three species.  Additional studies
              will be carried out on the Humpback Whale.  The humpback has a
              world-wide distribution, but the Atlantic and Pacific
              populations of the northern hemisphere appear to be discrete
              populations, as is the population of the southern hemispheric
              oceans.  Each of these oceanic populations may be further
              subdivided into smaller isolates, each with its own migratory
```

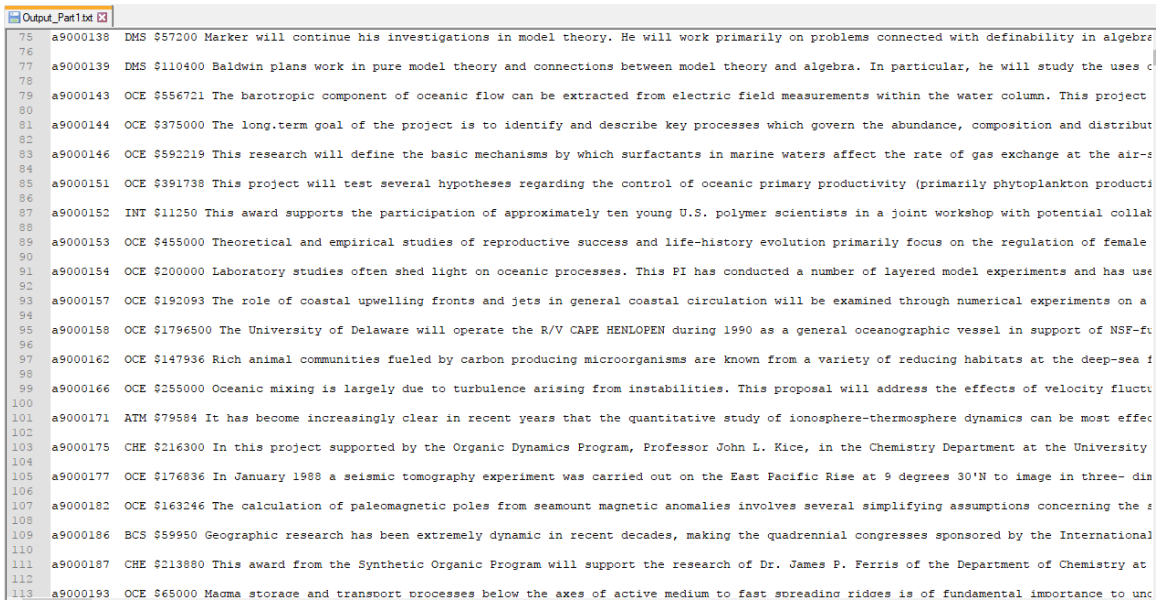
Output_part1:-

Output_Part1.txt

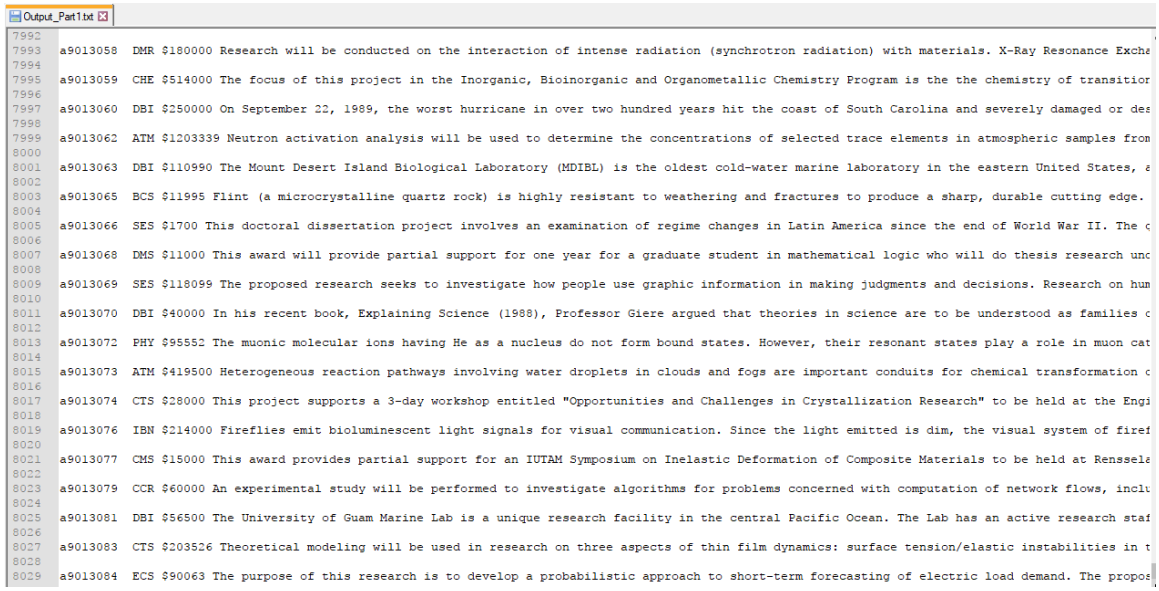
```
a9000006  DEB $179720 Commercial exploitation over the past two hundred years drove the great Mysticete whales to near extinction. Variation in the siz

a9000031  MCB $300000 Studies of chickens have provided serological and nucleic acid probes useful in defining the major histocompatibility complex (MH

a9000038  DMS $188574 This research is part of an on-going program by the principal investigator and associates. Topics in the following areas are to b

a9000040  DMI $225024 This SBIR proposal is aimed at (1) the synthesis of new ferroelectric liquid crystals with ultra-high polarization, chemical stab

a9000043  OCE $463490 Dr. Chisholm will investigate fundamental aspects of growth regulation and dynamics of marine plankton in the fluctuating environ

a9000045  CCR $53277 This research will study the complexity of computation using the framework of Boolean circuit complexity. Special emphasis is plac

a9000046  OCE $3842340 Duke University will operate the R/V CAPE HATTERAS during 1990 as a general oceanographic vessel in support of NSF-funded resear

a9000048  OCE $14546493 The Scripps Institute of Oceanography will operate four research vessels: R/V MELVILLE, a 245' general oceanographic vessel con

a9000049  OCE $2916509 Bermuda Biological Station will operate the R/V WEATHERBIRD II during 1990 as a general oceanographic vessel in support of NSF-f

a9000050  OCE $50000 This proposal seeks to demonstrate a technique for observing ocean currents by electric field measurements using a towed instrumen

a9000052  ATM $125000 The motion of energetic particles in the geospace environment depends sensitivety upon solarwind changes. This grant is to assess

a9000053  DMS $197491 The mathematical theories of multivariate polynomial interpolation and multivariate spline approximation differ in content and go

a9000054  DMS $12192 Work to be done during the period of this award will focus on higher dimensional inverse scattering problems and on related one di

a9000057  INT $20348 This proposal requests funds to permit Dr. Patrick S. Mariano, Department of Chemistry, University of Maryland, to pursue with Dr.

a9000058  INT $11250 This Science in Developing Countries award will help to support a research collaboration between Professor James Erskine of the Un

a9000060  OCE $322000 In this project, the P.I. will use model and data assimilation techniques to study seasonal and interannual variability in freshw

a9000063  DEB $320700 The effects of deforestation on the extinction rates of plant species in tropical rain forests are well documented. Less is knowr

a9000075  IBN $159944 In collaboration with Costa Rican graduate students and scientists at two universities in Costa Rica, Dr. Owens will carry out a

a9000089  DEB $477000 Our ability to restore tropical ecosystems and to construct sustainable, useful analogs of tropical forests depends on our abilit

a9000091  DEB $169000 Optimizing the chances of survival of rare or endangered plants is a fundamental concern of plant conservation biologists. Part c
```

Output_Part1.txt

```
a9000091  DEB $169000 Optimizing the chances of survival of rare or endangered plants is a fundamental concern of plant conservation biologists. Part c

a9000094  IBN $53563 The continued destruction of the coastal and tropical forest in South and Central America threatens the survival of natural animal

a9000099  BCS $199979 9000099 Blumenschine With National science Foundation support, Dr. Robert Blumenschine and his colleagues will conduct two seasor

a9000100  IBN $49219 All 37 species of monitor lizards are considered as threatened or endangered. Habitat destruction and human intervention have caus

a9000102  DEB $140070 Traditional forestry practices in the Northeast have led to the dispersion of forest openings and scattered clear cuts throughout

a9000110  DMS $71000 Work to be done on this project continues mathematical research on nonlinear elliptic problems arising in perfect-fluid hydrodynam

a9000111  ATM $40000 In order to fully understand how energetic particles that precipitate into the earth's atmosphere loose energy requires a combinat

a9000112  OCE $228603 The Woods Hole Oceanographic Institution will continue an oceanographic instrumentation development project to develop, construct

a9000117  CTS $252238 Nonsteady state performance of chemical reactors are difficult to predict and analyze. Many industrial reactors exhibit such beha

a9000121  DMI $250039 Multilayer coatings can vastly improve the performance of X-ray optical elements. They have a variety of current applications in

a9000127  OCE $480000 In this project, the P.I.'s will investigate numerical methods leading to development of improved global ocean models suitable fo

a9000129  DMS $35883 This project is concerned with the relationship between the torsion product and various classes of abelian p-groups. Let C be a cl

a9000130  OCE $465457 The University of Michigan will operate the R/V LAURENTIAN during 1990 as a general oceanographic vessel in support of NSF-funded

a9000132  DMS $29565 The principal investigator will continue his research in probability in infinite dimensional spaces, particularly on convergence o

a9000133  DMS $43490 The general objective of this project is to understand and to predict the surface structures and geometries of crystals which have

a9000134  OCE $100917 Previous work by the Dr. Wells has demonstrated that organically bound aggregates of mineral grains are relatively common to estu

a9000135  OCE $135000 Since 1982, the California coast has seen an extraordinary number of very large scale disturbances in kelp beds. These resulted f

a9000137  DMS $43500 The principal investigator will continue her studies of hydrodynamic and hydromagnetic waves of the type that occur in the earth's

a9000138  DMS $57200 Marker will continue his investigations in model theory. He will work primarily on problems connected with definability in algebra
```


Output_Part1.txt

```
a9000138  DMS $57200 Marker will continue his investigations in model theory. He will work primarily on problems connected with definability in algebra

a9000139  DMS $110400 Baldwin plans work in pure model theory and connections between model theory and algebra. In particular, he will study the uses o

a9000143  OCE $556721 The barotropic component of oceanic flow can be extracted from electric field measurements within the water column. This project

a9000144  OCE $375000 The long.term goal of the project is to identify and describe key processes which govern the abundance, composition and distribut

a9000146  OCE $592219 This research will define the basic mechanisms by which surfactants in marine waters affect the rate of gas exchange at the air-s

a9000151  OCE $391738 This project will test several hypotheses regarding the control of oceanic primary productivity (primarily phytoplankton producti

a9000152  INT $11250 This award supports the participation of approximately ten young U.S. polymer scientists in a joint workshop with potential collab

a9000153  OCE $455000 Theoretical and empirical studies of reproductive success and life-history evolution primarily focus on the regulation of female

a9000154  OCE $200000 Laboratory studies often shed light on oceanic processes. This PI has conducted a number of layered model experiments and has use

a9000157  OCE $192093 The role of coastal upwelling fronts and jets in general coastal circulation will be examined through numerical experiments on a

a9000158  OCE $1796500 The University of Delaware will operate the R/V CAPE HENLOPEN during 1990 as a general oceanographic vessel in support of NSF-fu

a9000162  OCE $147936 Rich animal communities fueled by carbon producing microorganisms are known from a variety of reducing habitats at the deep-sea f

a9000166  OCE $255000 Oceanic mixing is largely due to turbulence arising from instabilities. This proposal will address the effects of velocity fluctu

a9000171  ATM $79584 It has become increasingly clear in recent years that the quantitative study of ionosphere-thermosphere dynamics can be most effec

a9000175  CHE $216300 In this project supported by the Organic Dynamics Program, Professor John L. Kice, in the Chemistry Department at the University

a9000177  OCE $176836 In January 1988 a seismic tomography experiment was carried out on the East Pacific Rise at 9 degrees 30'N to image in three- dim

a9000182  OCE $163246 The calculation of paleomagnetic poles from seamount magnetic anomalies involves several simplifying assumptions concerning the s

a9000186  BCS $59950 Geographic research has been extremely dynamic in recent decades, making the quadrennial congresses sponsored by the International

a9000187  CHE $213880 This award from the Synthetic Organic Program will support the research of Dr. James P. Ferris of the Department of Chemistry at

a9000193  OCE $65000 Magma storage and transport processes below the axes of active medium to fast spreading ridges is of fundamental importance to unc
```


Output_Part1.txt

```
a9013058  DMR $180000 Research will be conducted on the interaction of intense radiation (synchrotron radiation) with materials. X-Ray Resonance Excha

a9013059  CHE $514000 The focus of this project in the Inorganic, Bioinorganic and Organometallic Chemistry Program is the the chemistry of transitior

a9013060  DBI $250000 On September 22, 1989, the worst hurricane in over two hundred years hit the coast of South Carolina and severely damaged or des

a9013062  ATM $1203339 Neutron activation analysis will be used to determine the concentrations of selected trace elements in atmospheric samples from

a9013063  DBI $110990 The Mount Desert Island Biological Laboratory (MDIBL) is the oldest cold-water marine laboratory in the eastern United States, a

a9013065  BCS $11995 Flint (a microcrystalline quartz rock) is highly resistant to weathering and fractures to produce a sharp, durable cutting edge.

a9013066  SES $1700 This doctoral dissertation project involves an examination of regime changes in Latin America since the end of World War II. The c

a9013068  DMS $11000 This award will provide partial support for one year for a graduate student in mathematical logic who will do thesis research unc

a9013069  SES $118099 The proposed research seeks to investigate how people use graphic information in making judgments and decisions. Research on hun

a9013070  DBI $40000 In his recent book, Explaining Science (1988), Professor Giere argued that theories in science are to be understood as families o

a9013072  PHY $95552 The muonic molecular ions having He as a nucleus do not form bound states. However, their resonant states play a role in muon cat

a9013073  ATM $419500 Heterogeneous reaction pathways involving water droplets in clouds and fogs are important conduits for chemical transformation o

a9013074  CTS $28000 This project supports a 3-day workshop entitled "Opportunities and Challenges in Crystallization Research" to be held at the Engi

a9013076  IBN $214000 Fireflies emit bioluminescent light signals for visual communication. Since the light emitted is dim, the visual system of firef

a9013077  CMS $15000 This award provides partial support for an IUTAM Symposium on Inelastic Deformation of Composite Materials to be held at Rensselа

a9013079  CCR $60000 An experimental study will be performed to investigate algorithms for problems concerned with computation of network flows, inclu

a9013081  DBI $56500 The University of Guam Marine Lab is a unique research facility in the central Pacific Ocean. The Lab has an active research staf

a9013083  CTS $203526 Theoretical modeling will be used in research on three aspects of thin film dynamics: surface tension/elastic instabilities in t

a9013084  ECS $90063 The purpose of this research is to develop a probabilistic approach to short-term forecasting of electric load demand. The propos
```

Output_part2:-

Output_Part2.txt

```
Abstract_ID|Sentence_No|Sentence

a9000006|1|Commercial exploitation over the past two hundred years drove the great Mysticete whales to near extinction.
a9000006|2|Variation in the sizes of populations prior to exploitation, minimal population size during exploitation and current population sizes permit
a9000006|3|Dr. Stephen Palumbi at the University of Hawaii will study the genetic population structure of three whale species in this context, the Hump
a9000006|4|The effect of demographic history will be determined by comparing the genetic structure of the three species.
a9000006|5|Additional studies will be carried out on the Humpback Whale.
a9000006|6|The humpback has a world-wide distribution, but the Atlantic and Pacific populations of the northern hemisphere appear to be discrete popula
a9000006|7|Each of these oceanic populations may be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct
a9000006|8|This study will provide information on the level of genetic isolation among populations and the levels of gene flow and genealogical relatic
a9000006|9|This detailed genetic information will facilitate international policy decisions regarding the conservation and management of these magnific
Number of Lines in file a9000006    is :- 9
-----------------------------------------------------------
a9000031|1|Studies of chickens have provided serological and nucleic acid probes useful in defining the major histocompatibility complex (MHC) in other
a9000031|2|Methods used in detecting genetic diversity at loci within the MHC of chickens and mammals will be applied to determining the extent of MHC
a9000031|3|The knowledge and expertise gained from working with the MHC of the chicken should make for rapid progress in defining the polymorphism of t
a9000031|4|Genes within the major histocompatibility complex (MHC) are known to encode molecules that provide the context for recognition of foreign ar
a9000031|5|Whether a given animal is able to mount an immune response to the challenge of a pathogen is determined, in part, by the allelic makeup of i
a9000031|6|In many species, an unusually high degree of polymorphism is maintained at multiple loci within the MHC in freely breeding populations.
a9000031|7|The allelic pool within a population presumably provides diversity upon which to draw in the face of environmental challenge.
a9000031|8|The objective of the proposed research is to extend ongoing studies of the MHC of domesticated fowl to include avian species experiencing se
a9000031|9|Knowledge of the MHC gene pool within populations and of the haplotypes of individual animals may be useful in the husbandry of species requ
Number of Lines in file a9000031    is :- 9
-----------------------------------------------------------
a9000038|1|This research is part of an on-going program by the principal investigator and associates.
a9000038|2|Topics in the following areas are to be considered: (1) controlled Markov diffusions and nonlinear PDEs; (2) asymptotic properties of nearly
a9000038|3|Analytical methods based on viscosity solution techniques for nonlinear differential equations as well as probabilistic methods will be stud
a9000038|4|These theoretical studies are the basis for applied problems ranging from decisions at the stock market level to the control of spaceships.
Number of Lines in file a9000038    is :- 4
-----------------------------------------------------------
a9000040|1|This SBIR proposal is aimed at (1) the synthesis of new ferroelectric liquid crystals with ultra-high polarization, chemical stability and l
Number of Lines in file a9000040    is :- 1
-----------------------------------------------------------
a9000043|1|Dr. Chisholm will investigate fundamental aspects of growth regulation and dynamics of marine plankton in the fluctuating environments that
a9000043|2|This understanding is essential for modelling and designing studies of marine productivity and food web dynamics.
a9000043|3|Specifically, they will: * Study the diatom life cycle, to better understand what environmental and genetic factors control the switch from
a9000043|4|The results of this work will augment our understanding of microbial growth rates in the sea and phytoplankton population genetics; it shoul
a9000043|5|Moreover, it could enhance our fundamental understanding of microbial physiology by revealing features of marine organisms which deviate fro
Number of Lines in file a9000043    is :- 5
```

Output_Part2.txt

```
a9000043|5|Moreover, it could enhance our fundamental understanding of microbial physiology by revealing features of marine organisms which deviate fro
Number of Lines in file a9000043    is :- 5
-----------------------------------------------------------
a9000045|1|This research will study the complexity of computation using the framework of Boolean circuit complexity.
a9000045|2|Special emphasis is placed on the following topics: Strong separations of circuit classes: If known separations of small circuit complexity
a9000045|3|This connection will be investigated, using the notion of "immunity" as a tool.
a9000045|4|Width-bounded reducibility: This notion will be used as a tool to investigate the relationships among "similar" complexity classes.
a9000045|5|This project also investigates threshold circuits, an structure of the complexity class P/poly.
Number of Lines in file a9000045    is :- 5
-----------------------------------------------------------
a9000046|1|Duke University will operate the R/V CAPE HATTERAS during 1990 as a general oceanographic vessel in support of NSF-funded research projects.
a9000046|2|The R/V POINT SUR is a 135' general research vessel constructed in 1981 and owned by the National Science Foundation.
a9000046|3|Duke operates the CAPE HATTERAS under a charter agreement with NSF.
a9000046|4|The ship operates primarily off the U.S. east coast from Maine to Florida.
a9000046|5|This vessel is part of a fleet used by the National Science Foundation to support oceanographic research projects.
a9000046|6|Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members.
a9000046|7|An increasing number of research projects require equipment that must be permanently installed on a ship and for which the ship must be spec
a9000046|8|Such equipment also requires highly trained crew members for maintenance and operation.
a9000046|9|These vessels do not operate in the same manner as general cargo or fishing vessels, and therefore, NSF supports the operation of a variety
a9000046|10|These vessels are operated by universities and research institutions around the country.
Number of Lines in file a9000046    is :- 10
-----------------------------------------------------------
a9000048|1|The Scripps Institute of Oceanography will operate four research vessels: R/V MELVILLE, a 245' general oceanographic vessel constructed by t
a9000048|2|These vessels are part of a fleet used by the National Science Foundation to support oceanographic research projects.
a9000048|3|Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members.
a9000048|4|An increasing number of research projects require equipment that must be permanently installed on a ship and for which the ship must be spec
a9000048|5|Such equipment also requires highly trained crew members for maintenance and operation.
a9000048|6|These vessels do not operate in the same manner as general cargo or fishing vessels, and therefore, NSF supports the operation of a variety
a9000048|7|These vessels are operated by universities and research institutions around the country.
a9000048|8|The R/V's T.WASHINGTON and MELVILLE operate worldwide, while the R/V NEW HORIZON operates primarily in the northeastern Pacific.
a9000048|9|The R/V SPROUL operates primarily in the coastal waters of California.
Number of Lines in file a9000048    is :- 9
-----------------------------------------------------------
a9000049|1|Bermuda Biological Station will operate the R/V WEATHERBIRD II during 1990 as a general oceanographic vessel in support of NSF-funded resear
a9000049|2|The R/V WEATHERBIRD II is a 115' general research vessel that was originally converted in 1989 and is owned by the Bermuda Biological Static
a9000049|3|Additional conversion work on the vessel will be conducted in phases during 1990 and 1991.
a9000049|4|The ship operates primarily in the vicinity of Bermuda.
a9000049|5|This vessel is part of a fleet used by the National Science Foundation to support oceanographic research projects.
a9000049|6|Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members.
```

Output_Part2.txt

```
a9000049|5|This vessel is part of a fleet used by the National Science Foundation to support oceanographic research projects.
a9000049|6|Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members.
a9000049|7|An increasing number of research projects require equipment that must be permanently installed on a ship and for which the ship must be spec
a9000049|8|Such equipment also requires highly trained crew members for maintenance and operation.
a9000049|9|These vessels do not operate in the same manner as general cargo or fishing vessels, and therefore, NSF supports the operation of a variety
a9000049|10|These vessels are operated by universities and research institutions around the country.
Number of Lines in file a9000049    is :- 10
-----------------------------------------------------------
a9000050|1|This proposal seeks to demonstrate a technique for observing ocean currents by electric field measurements using a towed instrument of recer
a9000050|2|The measurements will be made in conjunction with a cruise across the Gulf Stream in which several additional observational techniques will
a9000050|3|The several data types will be intercompared to improve the accuracy of the methods.
Number of Lines in file a9000050    is :- 3
-----------------------------------------------------------
a9000052|1|The motion of energetic particles in the geospace environment depends sensitivety upon solarwind changes.
a9000052|2|This grant is to assess the long-term solar cycle affects on the characteristics of energetic particles in the near-earth space and the cont
a9000052|3|It is also planned to investigate energetic particle processes during substorms through participation of CDAW-9 and by comparison between su
a9000052|4|An additional project will be to examine the stability of the neutral sheet in the so-called current disruption region in the magnetotail.
a9000052|5|*** //
Number of Lines in file a9000052    is :- 5
-----------------------------------------------------------
a9000053|1|The mathematical theories of multivariate polynomial interpolation and multivariate spline approximation differ in content and goals, yet sh
a9000053|2|In addition, many of the mathematical tools used to analyze basic questions are similar.
a9000053|3|Underlying much of this work has been the problem of developing a strategy for developing a theory of splines in several dimensions which is
a9000053|4|Work on multivariate polynomial interpolation derives from problems in what became known as box spline theory.
a9000053|5|A surprisingly simple and general method for choosing, for any given finite set of points in a space of several variables, a good polynomial
a9000053|6|Work will now be done exploring the theoretical and practical ramifications of this discovery.
a9000053|7|A long-term objective is to construct a coherent theory of interpolation, one which will play a more important role in multivariate numerica
a9000053|8|There are many approaches to spline approximation in higher dimensions currently in use.
a9000053|9|Each is associated with a certain type of mesh along which the elements are joined.
a9000053|10|The focus of this work will be that of approximation order.
a9000053|11|A better understanding of what makes for a good approximation order is expected to lead to the construction of better approximation methods
a9000053|12|Ultimately, one would like to develop a unification of the various theories and techniques now extant.
a9000053|13|A particularly important area of application of this work is in providing mathematical models of surfaces, often from a given set of boundi
a9000053|14|The mainstay of industrial work at this time is a method which only works for relatively flat surfaces.
a9000053|15|One immediate goal is to obtain a better understanding of how one can tell whether a given surface can be well represented by a small numbe
Number of Lines in file a9000053    is :- 15
-----------------------------------------------------------
a9000054|1|Work to be done during the period of this award will focus on higher dimensional inverse scattering problems and on related one dimensional
a9000054|2|The underlying idea for work of this nature is that of constructing obstacles from data measured as particles pass the obstacle.
```

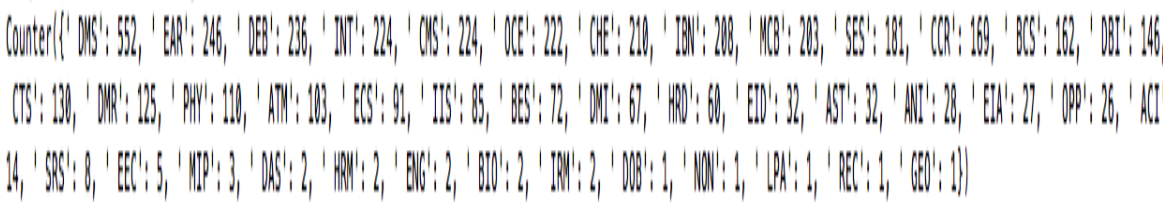
Output_Part2.txt

```
35581  -----------------------------------------------------------
35582  a9013079|1|An experimental study will be performed to investigate algorithms for problems concerned with computation of network flows, including the
35583  a9013079|2|The experimental study will proceed as a cooperative competition among members of the research community.
35584  a9013079|3|Participants will submit code for evaluation or will perform specific experiments at their home sites.
35585  a9013079|4|The competition will culminate in a workshop where participants will describe their results; the best implementations for several input cl
35586  a9013079|5|Competitors will generate a large amount of data about the performance of these algorithms.
35587  a9013079|6|Parametric and nonparametric methods of data analysis will be applied to obtain quantitative descriptions of performance as a function of
35588  a9013079|7|Some particular results to be obtained include: analysis of the relationship between running time and standard combinatorial measures of p
35589  a9013079|8|Besides providing substantial new knowledge about the performance of these algorithms, the project will also give insight into experimenta
35590  Number of Lines in file a9013079    is :- 8
35591  -----------------------------------------------------------
35592  a9013081|1|The University of Guam Marine Lab is a unique research facility in the central Pacific Ocean.
35593  a9013081|2|The Lab has an active research staff and hosts a steady stream of international investigators working in diverse disciplines.
35594  a9013081|3|The Lab also has an excellent record of graduate and undergraduate training.
35595  a9013081|4|Resident and visiting scientists have access to an array of shared equipment and instrumentation, and Dr. Ernest Matson proposes installat
35596  a9013081|5|The proposed multi-user equipment acquisitions will have a significant impact on the quantity and quality of research conducted at the Lab
35597  a9013081|6|The generator will provide backup electricity during nighttime and weekend shutdowns, and the intermittent power failures caused by tropic
35598  a9013081|7|The spectrophotometer will be an important addition to a diverse array of projects in marine ecology, systematic biology, biochemistry and
35599  Number of Lines in file a9013081    is :- 7
35600  -----------------------------------------------------------
35601  a9013083|1|Theoretical modeling will be used in research on three aspects of thin film dynamics: surface tension/elastic instabilities in thin flexib
35602  a9013083|2|The research is motivated by biomechanical issues related to the lung, namely the stability of small airways and the delivery of medicatic
35603  a9013083|3|The research is also applicable to surfactant effects in industrial thin films and the micropores of porous media, particularly poroelasti
35604  Number of Lines in file a9013083    is :- 3
35605  -----------------------------------------------------------
35606  a9013084|1|The purpose of this research is to develop a probabilistic approach to short-term forecasting of electric load demand.
35607  a9013084|2|The proposed approach accounts explicitly for weather and lifestyle influences.
35608  a9013084|3|The single customer load demand is decomposed into two components: stochastic and periodic.
35609  a9013084|4|The stochastic components (heating/cooling and water heater) are modeled via stochastic differential equations.
35610  a9013084|5|Semi-Markov theory is used to determine the load of these components.
35611  a9013084|6|The periodic part (the remaining part of the load, excluding the stochastic components) is determined using time series methods.
35612  a9013084|7|The second phase of this research is to identify the parameters of the models of the different components.
35613  a9013084|8|A validation of the proposed approach will also be performed.
35614  a9013084|9|The second phase will use the data gathered during the Department of Energy experimental project on distribution automation and control at
35615  a9013084|10|Since the basic building block of the proposed approach is the forecast at the single customer level, the accuracy of the forecast of the
35616  a9013084|11|The results of these investigations will provide the utilities with a more accurate technique for short-term load forecast at any desirec
35617  a9013084|12|The availability of an accurate forecast will enable the utilities to make more confident decisions of operation and planning.
35618  Number of Lines in file a9013084    is :- 12
35619  -----------------------------------------------------------
```

Output of Scenario 1:-

```
Rishis-MacBook-Pro:Assignment_2 rishi$ python3 Modified.py
Dictionary of Organization and its awards
defaultdict(<class 'list'>, {' DEB': [179720, 320700, 477000, 169000, 140070, 14400, 240375, 45422, 9845, 8000, 8589, 6817, 8248, 8500, 5400, 8500, 10000, 4421, 9274, 9000, 8000,
10000, 10080, 7000, 8487, 7520, 8000, 3080, 7500, 5061, 9499, 8375, 6054, 10946, 77577, 7000, 58800, 58800, 58800, 7000, 49996, 58800, 58800, 58800, 58800, 58800, 58800, 58800, 58
800, 58800, 117126, 324292, 274603, 56721, 34796, 10000, 117170, 25000, 40381, 59475, 42638, 64463, 38600, 73587, 229999, 7398, 10372, 305950, 160000, 160000, 80000, 5800, 255000,
 52624, 15577, 65005, 76502, 46528, 30000, 25000, 123542, 5353, 105000, 45000, 271000, 47106, 55000, 120026, 190016, 74985, 99180, 185000, 39760, 16109, 95000, 37208, 145000, 1464
31, 63221, 364188, 15000, 184440, 65000, 75000, 180000, 160757, 99084, 225000, 94987, 15000, 156237, 143182, 64839, 37269, 103000, 107130, 13450, 33172, 363756, 90980, 81474, 3050
00, 15000, 93009, 184996, 24999, 73276, 100000, 14605, 118030, 184973, 139564, 117388, 82163, 20000, 40000, 110000, 30000, 72845, 123000, 300010, 215738, 262995, 239774, 60000, 59
912, 200000, 46483, 61519, 288000, 310000, 157000, 234670, 95613, 85011, 212000, 184900, 185000, 199994, 350963, 180000, 1902197, 204894, 147200, 400000, 100000, 330000, 300000, 1
89295, 274550, 634998, 12513, 90441, 15000, 110000, 149828, 106362, 159834, 149982, 25000, 74135, 260100, 200185, 99000, 174718, 163800, 65743, 263000, 15000, 3504166, 163800, 100
00, 300000, 300000, 73830, 17908, 406500, 190000, 41979, 227328, 87379, 99999, 269905, 12000, 18000, 316009, 146397, 32165, 300000, 253387, 22000, 16000, 316600, 40000, 18000, 560
92, 11771, 50000, 31353, 42327, 141457, 971370, 3483858, 4359827, 4373619, 3517313, 3491499, 1264362, 121560, 49910, 4000, 352846, 24000, 1380009, 50000, 24520], ' MCB': [300000,
5000, 5000, 470431, 5000, 15000, 5000, 87500, 285000, 3000, 10000, 15000, 335000, 521890, 219000, 48964, 402944, 49640, 240050, 291000, 120000, 5000, 225000, 40000, 35000, 280000,
 222000, 342000, 49779, 413102, 252000, 100000, 259000, 270000, 195000, 6000, 270000, 285000, 268000, 355625, 196142, 138000, 246000, 270000, 273100, 219237, 260000, 140000, 25500
0, 200000, 4000, 274521, 411616, 243000, 272000, 409000, 294800, 276000, 337040, 321067, 225000, 170000, 216000, 95000, 427905, 218900, 288784, 567000, 120000, 100000, 286000, 175
000, 239872, 262000, 50000, 178300, 95000, 255000, 307000, 40000, 333338, 99800, 290000, 235000, 284922, 54830, 285000, 304000, 321000, 180000, 194642, 309000, 219601, 90000, 1899
97, 265000, 135000, 240000, 278616, 300000, 285000, 234000, 113000, 306121, 263000, 259045, 403000, 315000, 316000, 279879, 272555, 278000, 286000, 98000, 281250, 85000, 267000, 7
200, 170000, 249000, 290652, 29781, 270000, 252500, 125000, 300000, 75879, 231000, 265000, 340463, 105288, 105075, 105000, 105000, 105000, 105000, 105000, 219765, 345000, 300000,
229000, 170000, 150000, 263000, 85000, 321725, 290000, 4000, 300000, 296000, 15714, 264000, 18000, 50000, 6000, 10000, 289000, 395000, 12000, 195000, 261000, 254000, 97469, 97548,
 97200, 97406, 98600, 97272, 97200, 97200, 64800, 97299, 97632, 97870, 97200, 16416, 256763, 14355, 18000, 59572, 18000, 61234, 26841, 49768, 50000, 18000, 59800, 17424, 17975, 18
000, 18000, 18000, 13769, 270000, 60000, 72000, 5000, 159679, 280000, 255000, 6000, 49808, 10000], ' DMS': [188574, 197491, 12192, 71000, 35883, 29565, 43490, 43500, 57200, 110400
, 130750, 32400, 20000, 33270, 171300, 59500, 35837, 11000, 32497, 52600, 8000, 38015, 21700, 36226, 49200, 106250, 37680, 56100, 171150, 180728, 46823, 18500, 16325, 165000, 2400
0, 58850, 19910, 32000, 72000, 64000, 24000, 85500, 117400, 24000, 126158, 163250, 32352, 38097, 23000, 60000, 36425, 18390, 67554, 87000, 0, 128948, 38900, 0, 40180, 467000, 4132
5, 5000, 60784, 69000, 42800, 59600, 33214, 190600, 61800, 151300, 82590, 82600, 86700, 37700, 72000, 38290, 141120, 45990, 42600, 83210, 67700, 212700, 142950, 38280, 49423, 1155
00, 147072, 46550, 38000, 73500, 47745, 213100, 102377, 20000, 111600, 66319, 47764, 52550, 51000, 147880, 59500, 26543, 50300, 88000, 20000, 29800, 123150, 61850, 94700, 65000, 2
05450, 576415, 25000, 75879, 60200, 110000, 46100, 42100, 26374, 63350, 92404, 30000, 144000, 96350, 92284, 68577, 44680, 75975, 124470, 37398, 42000, 180000, 51292, 68701, 51164,
 47650, 86000, 9200, 104300, 46100, 127501, 25950, 57250, 102250, 0, 10000, 107000, 466038, 92420, 41089, 227800, 176769, 40418, 51250, 49550, 42100, 150100, 19650, 59600, 952451,
 68400, 37253, 42927, 31400, 45900, 72750, 198000, 79500, 43099, 148780, 115000, 125057, 29995, 36550, 48800, 52250, 87600, 75876, 51000, 53305, 102042, 15000, 114000, 31600, 1000
0, 12852, 58710, 51650, 147450, 32203, 65600, 34000, 46650, 8000, 286293, 340000, 180900, 45180, 42000, 50300, 168450, 24000, 61000, 23815, 126451, 74454, 56500, 67200, 339813, 76
160, 59241, 110855, 47300, 64000, 53900, 174381, 54400, 19890, 39800, 9670, 39000, 55600, 174450, 82250, 43440, 80330, 214900, 61524, 219700, 180000, 43500, 166200, 189900, 15600,
 110759, 14582, 33500, 85434, 15585, 83131, 83909, 26179, 48024, 87354, 45300, 31350, 43700, 56275, 219434, 40286, 43690, 33010, 217500, 15800, 178700, 142700, 164806, 29000, 5850
0, 43950, 236019, 46451, 49750, 94500, 35000, 98016, 51840, 47252, 16000, 38654, 81300, 68600, 80350, 42300, 27750, 45000, 40890, 491950, 170520, 52250, 37288, 136786, 164000, 190
40, 44885, 33500, 47291, 52000, 64000, 174100, 58646, 37874, 53750, 47255, 40529, 43900, 55515, 195530, 103047, 180000, 40620, 39000, 155600, 77480, 49400, 46080, 57795, 20000, 40
000, 40000, 260261, 28013, 30088, 0, 220100, 44300, 38207, 159942, 20000, 13792, 39260, 29700, 124500, 48050, 20000, 74503, 49847, 129950, 20000, 44910, 12687, 210000, 259900, 351
00, 42800, 80000, 328350, 53100, 36522, 66850, 314133, 44000, 34000, 16200, 15466, 222067, 23800, 48810, 30000, 64150, 158235, 20000, 451000, 43150, 48000, 35000, 10500, 17261, 42
686, 100778, 10250, 36585, 50000, 177019, 27000, 27000, 57200, 4000, 116600, 70100, 35000, 20000, 22000, 133339, 37500, 35000, 49316, 27500, 26000, 17000, 1382850, 20000, 287109,
66900, 242000, 18000, 50000, 38050, 48000, 21000, 50000, 42000, 100000, 17000, 35400, 36250, 30668, 30044, 45407, 18000, 46328, 42592, 291250, 71200, 61508, 185900, 11500, 56630,
193500, 59883, 3500, 11025, 34600, 53640, 110150, 16710, 156000, 6000, 33825, 62200, 68200, 35000, 43702, 160000, 17651, 180000, 23400, 218400, 151001, 93260, 97100, 15121, 420000
, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75000, 75
000, 189500, 91135, 266205, 41300, 70575, 20600, 368800, 190490, 41000, 68450, 56870, 27924, 95000, 37400, 44147, 75000, 256050, 134815, 15000, 35000, 29880, 43600, 38900, 12000,
35000, 46000, 8500, 30774, 38398, 72650, 30000, 11005, 15000, 95277, 48350, 231734, 39800, 82500, 195000, 0, 15000, 29320, 18000, 130550, 22700, 15000, 42000, 5210, 10800, 74000,
80580, 50000, 7700, 15000, 15000, 14000, 17000, 15000, 195400, 17000, 15500, 16259, 15000, 366900, 10000, 65040, 10000, 34052, 89255, 48500, 332889, 62615, 414412, 101790, 33731,
33700, 46244, 63733, 81561, 55819, 16900, 2027422, 11811, 83100, 136200, 30000, 65930, 1599768, 43537, 37700, 29830, 50000, 11000], ' DMI': [225024, 250039, 240397, 19458, 19822,
190336, 30000, 234250, 250000, 134721, 120000, 270431, 148461, 260162, 232640, 249930, 249668, 225000, 232084, 250000, 224505, 12000, 35000, 225000, 324395, 50081, 147350, 225000,
 4962, 100000, 100000, 110000, 250994, 100000, 100005, 99985, 249588, 250000, 130000, 225000, 99958, 12000, 12294, 205435, 230206, 88000, 70000, 60000, 69985, 74950, 50000,
 68269, 50000, 64397, 88615, 50000, 50000, 155253, 59526, 70000, 190888, 61500, 130000, 173303, 236574, 120001, ' OCE': [463490, 3842340, 14546493, 2916589, 50000, 322000, 228603
, 480000, 465457, 100917, 135000, 556721, 375000, 592219, 391738, 455000, 200000, 192093, 1796500, 147936, 255000, 176836, 163246, 65000, 250000, 159250, 622048, 4278539, 122940,
317970, 147000, 31250, 87784, 35000, 226344, 2980558, 188297, 110869, 117454, 37522, 1153600, 40000, 6503356, 126800, 33230, 69970, 173865, 344412, 52147, 159710, 56451, 385094, 7
5021, 65391, 709853, 129585, 10715, 200000, 152000, 82202, 216050, 139557, 57884, 44759, 86700, 86550, 189621, 72000, 140000, 78616, 141667, 107629, 12975, 18806079, 63401, 90392,
 113442, 4996327, 10177800, 352558, 23847, 37021, 14600, 160000, 195679, 60000, 141390, 172700, 47000, 165500, 323000, 550000, 195000, 1288406, 280000, 189017, 565000, 1158000, 69
878, 44849, 425000, 2550000, 935225, 680000, 31538, 1713507, 150000, 1210141, 49598, 207527, 263000, 374300, 581228, 50000, 143788, 90000, 39800, 158000, 48624, 250000, 97999, 120
00, 699306, 184298, 179000, 64990, 182576, 89490, 241000, 284671, 151530, 29985, 55271, 313068, 430302, 95276, 315000, 100000, 196072, 175264, 172793, 189432, 255000, 28000, 39400
0, 332848, 531654, 585000, 141164, 278138, 120486, 345000, 243860, 239004, 94539, 102998, 92900, 74031, 160000, 287733, 50052, 64947, 77531, 152000, 165000, 123982, 77749, 259502,
 579039, 244798, 250402, 150000, 327627, 110843, 162179, 107015, 770550, 922244, 778279, 142418, 137049, 53000, 121090, 225000, 264791, 45699, 133528, 37794, 43201, 120063, 137980
, 100000, 90026, 492200, 87000, 330000, 99000, 204451, 209934, 41278, 337229, 120612, 304536, 385000, 466634, 66031, 1072683, 844844, 213574, 91761, 1033712, 86494, 270865, 177770
, 248800, 251099, 86158, 62751, 150000, 125874, 27030, 59029], ' CCR': [53277, 130935, 54975, 193498, 80752, 99990, 104157, 185929, 4500, 96255, 154531, 35000, 6500, 55915, 100000
```

Output of Scenario 2:-



Output of Scenario 3:-

```
Max value from each key
 DEB 4373619
 MCB 567000
 DMS 2027422
 DMI 324395
 OCE 18806079
 CCR 667000
 ATM 1203339
 INT 930233
 IBN 361000
 BCS 410000
 CTS 555000
 CHE 693000
 DMR 2634600
 EAR 400000
 SES 491856
 OPP 3550650
 ANI 338509
 BES 616720
 PHY 6432676
 DBI 1247182
 SRS 1097959
 CMS 924737
 EID 100783
 DAS 2899686
 IIS 688500
 HRM 1069986
 ECS 499230
 AST 485000
 EIA 3670965
 EEC 740000
 ACI 810000
 DOB 649402
 NON 1
 ENG 151129
 LPA 132077
 HRD 683416
 MIP 188510
 BIO 109250
 IRM 4000
 REC 854642
 GEO 50000
```

Output of Scenario 4:- (Extra analysis)

```
Year:- NSF_Org
1991 Counter({' DMS': 135, ' OCE': 65, ' EAR': 45, ' SES': 42, ' IIS': 34, ' CCR': 33, ' BCS': 31, ' CMS': 29, ' DEB': 27, ' CHE': 25, ' INT': 23, ' ATM': 20, ' PHY': 19, ' BES':
17, ' IBN': 16, ' DBI': 14, ' CTS': 14, ' DMI': 12, ' ECS': 9, ' ANI': 8, ' MCB': 8, ' HRD': 7, ' EID': 6, ' EIA': 5, ' AST': 5, ' DMR': 4, ' OPP': 3, ' ACI': 2, ' SRS': 2, ' EEC'
: 1, ' HRM': 1})
1990 Counter({' DMS': 257, ' EAR': 161, ' INT': 151, ' CMS': 130, ' DEB': 119, ' SES': 104, ' DBI': 100, ' IBN': 84, ' CCR': 82, ' BCS': 81, ' CTS': 79, ' MCB': 74, ' ECS': 57, '
OCE': 55, ' CHE': 53, ' HRD': 46, ' PHY': 38, ' DMR': 34, ' BES': 33, ' DMI': 27, ' EID': 26, ' IIS': 22, ' ATM': 20, ' OPP': 16, ' EIA': 15, ' ANI': 14, ' AST': 12, ' ACI': 10, '
 ENG': 2, ' IRM': 2, ' MIP': 2, ' SRS': 2, ' NON': 1, ' BIO': 1, ' GEO': 1})
1992 Counter({' DMS': 132, ' CHE': 97, ' MCB': 73, ' IBN': 63, ' OCE': 62, ' DMR': 59, ' DEB': 39, ' PHY': 37, ' CCR': 36, ' ATM': 33, ' CMS': 31, ' CTS': 24, ' EAR': 22, ' IIS':
20, ' BCS': 18, ' SES': 17, ' DBI': 16, ' BES': 12, ' ECS': 11, ' INT': 10, ' DMI': 9, ' AST': 7, ' OPP': 5, ' ANI': 5, ' EIA': 5, ' HRD': 2, ' LPA': 1, ' SRS': 1, ' ACI': 1, ' MI
P': 1, ' BIO': 1, ' REC': 1, ' DAS': 1})
1995 Counter({' DMI': 5, ' CHE': 5, ' ATM': 5, ' OCE': 5, ' DEB': 4, ' INT': 4, ' MCB': 3, ' IBN': 3, ' PHY': 3, ' BCS': 1, ' SRS': 1, ' DAS': 1, ' HRM': 1, ' EEC': 1, ' ECS': 1,
' CCR': 1, ' SES': 1, ' EAR': 1, ' HRD': 1, ' DMS': 1, ' DBI': 1})
1993 Counter({' MCB': 33, ' IBN': 29, ' DEB': 28, ' INT': 24, ' CMS': 21, ' DMR': 21, ' OCE': 17, ' BCS': 17, ' CCR': 15, ' SES': 14, ' DBI': 13, ' ATM': 13, ' CHE': 12, ' EAR': 1
2, ' DMS': 11, ' ECS': 11, ' PHY': 10, ' IIS': 8, ' CTS': 8, ' DMI': 7, ' BES': 5, ' AST': 4, ' HRD': 3, ' DOB': 1, ' ANI': 1, ' ACI': 1, ' EIA': 1, ' SRS': 1})
1989 Counter({' CMS': 8, ' INT': 3, ' DEB': 2, ' BCS': 2, ' EAR': 2, ' OCE': 1, ' SRS': 1, ' DMI': 1, ' PHY': 1, ' CTS': 1, ' BES': 1, ' MCB': 1, ' CCR': 1})
1994 Counter({' CHE': 18, ' DMS': 16, ' OCE': 13, ' IBN': 13, ' BCS': 12, ' MCB': 11, ' DEB': 10, ' ATM': 9, ' INT': 8, ' DMR': 7, ' BES': 4, ' CTS': 4, ' AST': 4, ' CMS': 4, ' EA
R': 3, ' SES': 3, ' DMI': 2, ' PHY': 2, ' IIS': 1, ' EEC': 1, ' ECS': 1, ' OPP': 1, ' HRD': 1, ' DBI': 1, ' CCR': 1, ' EIA': 1})
1996 Counter({' DEB': 6, ' DMI': 4, ' OCE': 4, ' ATM': 3, ' DBI': 1, ' EEC': 1, ' INT': 1, ' CMS': 1})
1997 Counter({' EEC': 1, ' DEB': 1})
1998 Counter({' OPP': 1})
```

Note:-Here, Now I can see the output screenshots are not quite clear but I tried my best to do it. Still to re-check I request to you to run the code and see the output on the terminal. I have written down the code to print the data whatever its there in the screenshot.

Conclusion:-

Here, I analyzed the NSF_abstract data which has more then 4000 files in it. Fetch the require data and write it in to the file. Fetched the data using regular expression which gives us the most efficient data according to the requirements. After getting the data I make the most possible scenarios to get meaningful information from the given data. There were some issues with the data which was retrieved from the files but we can filter that later according to the requirement. I retrieved some extra information about the date factor and according to that I analyzed the given data with that and get some more information about the NSF Organization with respect to the year.

References:-

- https://www.nsf.gov/awards/about.jsp
- https://stackoverflow.com