Assignment 2

Prof. Lu Xiao

Name: Raghav Raheja
Course: IST664
SUID: 840334338

Assignment 2

A. Review the dataset

First review the dataset and describe the characteristics of the corpus briefly such as naming convention of its files, number of documents it contains, etc.

The data we worked on this week has been taken from National Science Foundation (NSF). The data is in form of reports regarding the award won by the organisation in NSF. The fields in the report range from the title of research under the organisation, award amount, abstract, etc. The format of all the reports are more or less the same with specific entries. The dataset we have is a subpart of the huge data as mentioned in the question set. The dataset we have for this assignment date from 1990 to 1994. Each file has been named as 'a' followed by the award number which acts as unique identifier. The amount mentioned in the report is the amount in awards received till date. A normal report has almost 18 fields specifying the type of award, the organisation, the project amount received and so on. For our assignment we have 4016 award abstracts, out of which one file 'a9013087' is completely empty. There are also some awards that do not have any money associated with it. There are almost 80 files that have amount as either $0 or $1. Additionally, there are also some files where abstract is given as not available, to be exact there are 68 such files.

B. Please submit the output with a tabular format included

Next, you will write a Python code that reads in each abstract and extract the abstract identity ('File'), NSF organization ('NSF org'), the award amount, and abstract text. Please submit the output with a tabular format included. The output may look like as follows.

In this step I used the function *'listdir'* from library os to loop over all the files present under the folder for our assignment.

Assignment 2

Prof. Lu Xiao

Name: Raghav Raheja
Course: IST664
SUID: 840334338

path = "C://……./IST 664 Natural Language Processing/Assignment 2/Data for HW2/"
filelist = os.listdir(path)

The *'filelist'* returned the name of all the files (a9000006.txt, a90000031.txt, ..). In the next step I created an empty list in which I could append my result. I ran a loop over all the files to read each one of them and process it. I used *'with' statement* to run a pair of commands, first reading the file(award abstract) and the next applying the regular expression to get our results. After reading the file, I realized that there are too many irregular spaces in the file. I used regular expression to find more than 2 spaces and substituted it with one space to make all the files look in a common format.

```
for i in filelist:
   with open(path + i, 'r') as f:
      rawtext = f.read()
      a = re.compile('[\s]{2,}')
      rawtext = a.sub(" ",rawtext)
```

The next step I did was to apply the regular expression, which looked like:

```
re.compile('[F]ile\s:\s[a]\d+|[O]rg\s:\s[A-Z]{3,4}\s|\$+\d+|[A]bstract\s:\s[\(|\")]*[0-9]*\w+.*\w+')
```

I thought of multiple regular expression. I even thought of finding each parameter separately and then concatenating it but it seemed like extra steps. I found that by using the specific words like [A]bstract, [F]ile, [O]rg reduces the chances of error, I learned that by creating different regular expressions without them and finally settling with this. This Regular expression takes into consideration all the parameters asked by the question. I do realize that by using this regular expression, I am incurring extra text (Abstract : , Org : , File :), but I thought that these can be substituted using the *'replace'* function. I realize that these will add extra bit of steps, but I was more focused towards picking up the exact abstract identities (parameters) and this seemed like the perfect solution.

```
g = re.findall(pword,rawtext)
      g = [s.replace("Abstract : ","") for s in g]
      g = [s.replace("Org : ","") for s in g]
      g = [s.replace("File : ","") for s in g]
      x6.append(g)
```

The last step I did in the loop was to append the result to my empty list.

After going through all these steps, I got the output in the form of a huge list in *'x6'*, which looked something like appendix B1.

The next step was to make it look like the way it appears in the example provided in the question. I tried to use a couple of different methods like tabulate, concatenation. However, at this time, I was also thinking from the next question perspective, where I realized that our main focus is abstract, so I tried to look for an option that would help me locate each abstract individually. I thought of data frames, as I have

Assignment 2

Prof. Lu Xiao

Name: Raghav Raheja
Course: IST664
SUID: 840334338

worked with it in R and they are really helpful in arranging the data in tabular format. I used pandas and created a data frame from my list *'x6'*. There are a couple of reasons I used data frame, one of the biggest

was that It becomes a lot easier to look for a missing column values in a data frame rather that in a list. The other reason was that I wanted to create an excel sheet (csv file), so that I can filter the data and check if my regular expression works completely fine on the whole dataset. I also created a text file using *'iloc'* function. The next part of the assignment also becomes a lot easier as it just involves locating the data from the data frame.

```
for index,row in df0.iterrows():
    if df0.iloc[index,1] is not None:
        t = t+df0.iloc[index,1]+ '\t '+df0.iloc[index,0]+'  '+df0.iloc[index,2]+ '   '+df0.iloc[index,3][:990]
+'\n'
```

This gives me the data in a string format looking exactly like it looks in the question statement (B2). The final step in this part is writing the code in a text file.

```
f = open("part1.txt","w")
f.write(t)
f.close()
```

3. Identify sentences in the abstract

You may use sentence tokenizers in Python. Your code output should contain the abstract identity, the sentence number, and the sentence text delimited with a bar (|), and the total number of sentences per each file at the end.

In this step I started with creating an empty string to store the result. The next step, I started with a for loop that will run on all the rows in my data frame. Since my last column in the data frame has the abstract, I used *'sent_tokenize'* on the last column. I created another variable storing the length, to have the number of sentences. I started another loop that will go till the total number of sentences. I used an if statement to take in only the values where the value is not None (Although there is no empty entry, but just to capture the exceptions). In the next step, I used string concatenation to add my desired values in the empty string *'g'* and I also added the total number of sentences in the same string outside my second for loop. The code looked like this:

```
for index,row in df0.iterrows():
    j = nltk.sent_tokenize(str(df0.iloc[index,-1]))
```

Assignment 2

Prof. Lu Xiao

Name: Raghav Raheja
Course: IST664
SUID: 840334338

```
  k = len(j)
    for l in range(0, k):
       if df0.iloc[index,1] is not None:
          g = g + '\n' + df0.iloc[index,1] + '|' + str(l+1) + '|'+ j[l]

if df0.iloc[index,1] is not None:
       g = g + '\n' + "Number of sentences: " + str(l+1)
```

In the final step, I wrote *'g'* in a text file naming part2. The output generated by it looks exactly same the one given in example (3A)

Appendix

B1

```
[['DEB ',
  'a9000006',
  '$179720',
  '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 pattern and somewhat distinct gene pool. This study will
provide information on the level of genetic isolation among populations and the levels of gene flow and
genealogical relationships among populations. This detailed genetic information will facilitate
international policy decisions regarding the conservation and management of these magnificent mammals'],
 ['MCB ',
  'a9000031',
  '$300000',
  'Studies of chickens have provided serological and nucleic acid probes useful in defining the major
histocompatibility complex (MHC) in other avian species. Methods used in detecting genetic diversity at
loci within the MHC of chickens and mammals will be applied to determining the extent of MHC polymorphism
within small populations of ring-necked pheasants, wild turkeys, cranes, Andean condors and other species.
The knowledge and expertise gained from working with the MHC of the chicken should make for rapid progress
in defining the polymorphism of the MHC in these species and in detecting the polymorphism of MHC gene pool
within small wild and captive populations of these birds. Genes within the major histocompatibility complex
(MHC) are known to encode molecules that provide the context for recognition of foreign antigens by the
immune system. 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 its MHC. In many species, an unusually high degree of
polymorphism is maintained at multiple loci within the MHC in freely breeding populations. The allelic pool
within a population presumably provides diversity upon which to draw in the face of environmental
challenge. The objective of the proposed research is to extend ongoing studies of the MHC of domesticated
fowl to include avian species experiencing severe reduction in population size. Knowledge of the MHC gene
pool within populations and of the haplotypes of individual animals may be useful in the husbandry of
species requiring intervention for their preservation'],
 ['DMS ',
  'a9000038',
```

Figure 1: Result from x6

Assignment 2

Prof. Lu Xiao

Name: Raghav Raheja
Course: IST664
SUID: 840334338

B2

```
a9000006        DEB   $179720   Commercial exploitation over the past two hundred years drove the gre
a9000031        MCB   $300000   Studies of chickens have provided serological and nucleic acid probes
a9000038        DMS   $188574   This research is part of an on-going program by the principal investi
a9000040        DMI   $225024   This SBIR proposal is aimed at (1) the synthesis of new ferroelectric
a9000043        OCE   $463490   Dr. Chisholm will investigate fundamental aspects of growth regulatic
a9000045        CCR   $53277   This research will study the complexity of computation using the frame
a9000046        OCE   $3842340   Duke University will operate the R/V CAPE HATTERAS during 1990 as a
a9000048        OCE   $14546493   The Scripps Institute of Oceanography will operate four research ve
a9000049        OCE   $2916509   Bermuda Biological Station will operate the R/V WEATHERBIRD II durin
a9000050        OCE   $50000   This proposal seeks to demonstrate a technique for observing ocean cur
a9000052        ATM   $125000   The motion of energetic particles in the geospace environment depends
a9000053        DMS   $197491   The mathematical theories of multivariate polynomial interpolation an
a9000054        DMS   $12192   Work to be done during the period of this award will focus on higher d
a9000057        INT   $20348   This proposal requests funds to permit Dr. Patrick S. Mariano, Departm
a9000058        INT   $11250   This Science in Developing Countries award will help to support a rese
a9000060        OCE   $322000   In this project, the P.I. will use model and data assimilation techni
a9000063        DEB   $320700   The effects of deforestation on the extinction rates of plant species
a9000075        IBN   $159944   In collaboration with Costa Rican graduate students and scientists at
a9000089        DEB   $477000   Our ability to restore tropical ecosystems and to construct sustainab
a9000091        DEB   $169000   Optimizing the chances of survival of rare or endangered plants is a
```

Figure 2: Final result from part 2B

3A

```
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
a9000006|3|Dr. Stephen Palumbi at the University of Hawaii will study the genetic population structure of three whale s
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 hemisph
a9000006|7|Each of these oceanic populations may be further subdivided into smaller isolates, each with its own migratc
a9000006|8|This study will provide information on the level of genetic isolation among populations and the levels of ge
a9000006|9|This detailed genetic information will facilitate international policy decisions regarding the conservation
Number of sentences: 9
a9000031|1|Studies of chickens have provided serological and nucleic acid probes useful in defining the major histocomp
a9000031|2|Methods used in detecting genetic diversity at loci within the MHC of chickens and mammals will be applied t
a9000031|3|The knowledge and expertise gained from working with the MHC of the chicken should make for rapid progress i
a9000031|4|Genes within the major histocompatibility complex (MHC) are known to encode molecules that provide the conte
a9000031|5|Whether a given animal is able to mount an immune response to the challenge of a pathogen is determined, in
a9000031|6|In many species, an unusually high degree of polymorphism is maintained at multiple loci within the MHC in f
a9000031|7|The allelic pool within a population presumably provides diversity upon which to draw in the face of environ
a9000031|8|The objective of the proposed research is to extend ongoing studies of the MHC of domesticated fowl to inclu
a9000031|9|Knowledge of the MHC gene pool within populations and of the haplotypes of individual animals may be useful
Number of sentences: 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)
a9000038|3|Analytical methods based on viscosity solution techniques for nonlinear differential equations as well as pr
a9000038|4|These theoretical studies are the basis for applied problems ranging from decisions at the stock market leve
Number of sentences: 4
```

Figure 3: Result for Question 3