



IST 664
Natural Language Processing

Assignment:
Homework 2

Rashmitha Varma Pandati
SUID: 622666081

Table of Contents

1. Introduction.....	1
2. Preprocessing	
A) Corpus Characteristics	
i. History.....	1
ii. Data Description.....	1
iii. Number of Documents.....	1
iv. Naming Convention.....	1
B) Reading and Extraction.....	2
3. Distribution of Sentence Lengths.....	2
4. Output Observations	
A) Tabular Output.....	2
B) Sentence Output.....	3
5. Appendix One: Outputs	
A) Tabular Output.....	3
B) Sentence Output.....	3
6. Appendix Two: Python Code.....	4
7. Appendix Three: IDE Screenshots	
A) Tabular Output.....	5
B) Sentence Output.....	6
8. References.....	6

REGULAR EXPRESSIONS TO ANALYZE NATIONAL SCIENCE FOUNDATION ABSTRACTS DATA

1. Introduction

The National Science Foundation Research Awards Abstracts corpus makes an audit of collection of different aspects of research and grants awarded for research being carried out across all domains and fields of study.

2. Pre-processing

A. Corpus Characteristics

The publicly available open source corpus collection of NSF research awards abstracts spanning 1990 – 2003 is analyzed as follows:

❖ History

The type of data consisting in this corpus is of text and tabular type. This corpus is the work from UCI [University of California, Irvine] and is originally owned by Michael J. Pazzani, who provided the abstracts, and Amnon Meyers, who provided the bag of words. These were donated on November 18, 2003.

❖ Data Description

The NSF Research Awards Abstracts have three main documents and they are:

➤ **134,161 abstracts describing NSF awards for basic research**

All the abstracts, one per file, were furnished by the NSF (National Science Foundation).

➤ **Bag-of-word data files extracted from the abstracts**

The bag-of-word data was produced by automatically processing the abstracts with a text analyzer called NSFAbst, built using VisualText. While most fields of the output are very accurate, the authors were not extracted from the Investigator field with 100% accuracy, due to wide variability.

➤ **List of words used for indexing the bag-of-word data**

The word list came from a separate process and may not include all the words of interest in the abstracts.

❖ Number of Documents

There are total of 4016 awards abstracts in the NSF Awards Abstracts dataset and its folder structure is as follows:

```
| RootDirectory
    | \ NSF_abstracts
        | \ \ Text File(.txt)
```

❖ Naming Convention

The files have been named in the format ‘*.txt’ where * represents the unique ID generated as the file was processed. The files consist of information like Title, File, Award Grant, Award department, Award start and end date, Sponsor etc. as fields followed by its values. For example:

- “Title” : “Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demography”
- “File” : “a9000006”
- “Award Number”: “9000006”
- “Sponsor” : “U of Hawaii Manoa”.

B. Reading and Extraction of Abstracts

- ❖ To overcome the opened file descriptor for reading and then extracting necessary items from large entries of directories, we copy the contents of each abstract to a list. We use list comprehension with Pathlib library to process the directory entries fast and efficient in object -oriented manner.
- ❖ Since, the fields are identical across all abstracts, we use regular expression to extract from them.
 - File/AbstractID : “*File.*:(.*)*” - To extract File ID. The regular expression extracts any character followed by “File :” field.
 - NSF Organization: “*NSF Org.*:(.*)*” - To extract NSF Org type. The regular expression extracts any character followed by “NSF Org :” field.
 - Award Amount : “*Total Amt.*(\\$[0-9]*)*” - To extract award grant. The regular expression extracts any digits followed by “Total Amt. :” field.
 - Abstract : “*\bAbstract\b.*([s\S]+)*” - To extract the abstract field. The regular expression matches any character (digits, alphanumeric, special, newline, tab, return, multiline) followed by “Abstract :” field.
- ❖ Once we extract the data using regular expression , we print them in tabular format and write to the text file – “TabularOutput.txt”.

3. Distribution of Sentence Lengths

- ❖ Once we have extracted the fields and values from each abstract, we perform tokenization using nltk tokenizer.
- ❖ We choose *Punkt* library for sentence tokenization. *Punkt.PunktTokenizer* provides a way of handling special characters like period within respective words (such as prefix) and actual end of sentence.
 - For eg: “Dr. Smith” is not end of sentence but a prefix.
- ❖ Each token identifies a sentence among each abstract. We print the details of sentence like sentence number, belonging file, corresponding sentence. Each of these fields are separated by “|” symbol.
- ❖ We also print number of sentences for first 3 abstracts in tabular format on console and each abstracts are written to file “SentenceOutput.txt”.

4. Output Observations

A) TabularOutput.txt

The required details namely, File Identity, NSF Org type, Total Amount and Abstract Content, of all the 4016 documents, are displayed in tabular format and saved in the file.

B) SentenceOutput.txt

The effect of sentence tokenization is shown by the display of each sentence individually along with the total number of sentences in that abstract.

5. Appendix One: Outputs

A) Tablular Output

```
TabularOutput - Notepad
File Edit Format View Help
a9000006 DEB $179720 Commercial exploitation over the past two hundred years drove the great Mysticete
be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct
a9000031 MCB $300000 Studies of chickens have provided serological and nucleic acid probes useful in de
C. In many species, an unusually high degree of polymorphism is maintained at multiple loci within the
a9000038 DMS $188574 This research is part of an on-going program by the principal investigator and ass
a9000040 DMI $225024 This SBIR proposal is aimed at (1) the synthesis of new ferroelectric liquid cryst
a9000043 OCE $463490 Dr. Chisholm will investigate fundamental aspects of growth regulation and dynamic
methodological advances for the study of marine bacteria at sea. Moreover, it could enhance our fundame
a9000045 CCR $53277 This research will study the complexity of computation using the framework of Boole
a9000046 OCE $3842340 Duke University will operate the R/V CAPE HATTERAS during 1990 as a general ocean
he operation of a variety of vessels specifically dedicated to oceanographic research. These vessels ar
a9000048 OCE $14546493 The Scripps Institute of Oceanography will operate four research vessels: R/V ME
vessels, and therefore, NSF supports the operation of a variety of vessels specifically dedicated to oc
a9000049 OCE $2916509 Bermuda Biological Station will operate the R/V WEATHERBIRD II during 1990 as a g
ng vessels, and therefore, NSF supports the operation of a variety of vessels specifically dedicated to
a9000050 OCE $50000 This proposal seeks to demonstrate a technique for observing ocean currents by elec
a9000052 ATM $125000 The motion of energetic particles in the geospace environment depends sensitively
a9000053 DMS $197491 The mathematical theories of multivariate polynomial interpolation and multivariat
ore important role in multivariate numerical analysis. There are many approaches to spline approximatio
a9000054 DMS $12192 Work to be done during the period of this award will focus on higher dimensional in
which are not necessarily bounded in time. This method has been used to achieve the complete solution (
a9000057 INT $20348 This proposal requests funds to permit Dr. Patrick S. Mariano, Department of Chemis
d in several joint publications in refereed journals. This successful collaboration has generated ideas
a9000058 INT $11250 This Science in Developing Countries award will help to support a research collabor
```

B) Sentence Output

```
SentenceOutput - Notepad
File Edit Format View Help

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 analyses of the
a9000006|3|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 G
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 populations, as is the
a9000006|7|Each of these oceanic populations may be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct gene pool.
a9000006|8|This study will provide information on the level of genetic isolation among populations and the levels of gene flow and genealogical relationships among popu
a9000006|9|This detailed genetic information will facilitate international policy decisions regarding the conservation and management of these magnificent mammals.
Number of sentences: 9

Abstract_ID | Sentence_No | Sentence
.....
a9000031|1|Studies of chickens have provided serological and nucleic acid probes useful in defining the major histocompatibility complex (MHC) in other avian species.
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 polymorphism with
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 the MHC in these s
a9000031|4|Genes within the major histocompatibility complex (MHC) are known to encode molecules that provide the context for recognition of foreign antigens by the imm
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 its MHC.
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 severe reduction in
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 requiring interventio
Number of sentences: 9

Abstract_ID | Sentence_No | Sentence
.....
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 deterministic Ma
a9000038|3|Analytical methods based on viscosity solution techniques for nonlinear differential equations as well as probabilistic methods will be studied.
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 sentences: 4

Abstract_ID | Sentence_No | Sentence
.....
```

6. Appendix Two: Python Code

```
1. #importing packages
2. import nltk
3. import sys
4. import os
5. import zipfile
6. from zipfile import ZipFile
7. import glob
8. import nltk
9. import re
10. import pathlib
11.
12.
13. #Initializing the folder structure
14. path = pathlib.Path("C:/Users/girl1/OneDrive/Documents/664/Homework 2/NSF_abstracts").r
    esolve().parent
15. NSFPath = path/"NSF_abstracts"
16.
17.
18. #Reading the contents of the text files present in the folder
19. AbstractRead = [file.read_text(encoding="ISO-8859-
    1") for file in NSFPath.rglob('*.txt')]
20.
21.
22. #Declaring an Empty List
23. AbstractStore = []
24. #Looping the extraction functions together for all the abstract files
25. for text in AbstractRead:
26.     if text:
27.         #Extracting the Filename from the abstract
28.         Abstract_Identity = re.findall(r"File.*:(.*)", text)[0].strip()
29.         #Extracting the Organization Type from the abstract
30.         NSFOrgType = re.findall(r"NSF Org.*:(.*)", text)[0].strip()
31.         #Extracting the Amount from the abstract
32.         Amount=re.findall(r"Total Amt.*(\$[0-9]*)", text)[0].strip()
33.         #Extracting the content from the abstract
34.         AbstractContent = " ".join(re.findall(r"\bAbstract\b.*([\s\S]+)", text)[0].spli
    t()))
35.         #Appending all the obtained outputs into the empty list
36.         AbstractStore.append({"File":Abstract_Identity,"OrgType":NSFOrgType,"Amount":Am
    ount,"Abstract":AbstractContent})
37.
38.
39. #Printing the new list
40. print(AbstractStore)
41.
42.
43. #Printing the output into a text file
44. with open('TabularOutput.txt','w+', encoding = 'UTF-8') as file:
45.     for item in AbstractStore:
46.         Data = item["File"]+" "+item["OrgType"]+" "+item["Amount"]+" "+item["Ab
    stract"]
47.         file.write(Data+"\n")
48.         print(Data)
49.
50.
51. #Declaring the punkt sentence tokenizer
52. SentenceDetector = nltk.data.load('tokenizers/punkt/english.pickle')
53. #Initializing count
```

```

54. count = 0
55. print("\n Sentence Tokenizer\n")
56. #Printing the output into a text file
57. with open("SentenceOutput.txt","w+", encoding = 'UTF-8') as file:
58.     for item in AbstractStore:
59.         #Distribution of Sentence Lengths for the first 3 abstracts
60.         if count < 3:
61.             print(" Abstract_ID | Sentence_No | Sentence ")
62.             print(".....")
63.
64.             file.write("\n Abstract_ID | Sentence_No | Sentence \n.....\n")
65.             SentenceList = SentenceDetector.tokenize(item["Abstract"].strip())
66.             for idx,sentence in enumerate(SentenceList):
67.                 print(item["File"]+"|"+str(idx+1)+"|"+sentence)
68.                 file.write(item["File"]+"|"+str(idx+1)+"|"+sentence+"\n")
69.             print("Number of sentences: ",len(SentenceList))
70.             file.write("Number of sentences: "+str(len(SentenceList))+"\n")
71.             print()
72.             count = count + 1
73.         #Distribution of Sentence Lengths for remaining abstracts
74.         else:
75.             file.write("\n Abstract_ID | Sentence_No | Sentence \n.....\n")
76.             SentenceList = SentenceDetector.tokenize(item["Abstract"].strip())
77.             for idx, sentence in enumerate(SentenceList):
78.                 file.write(item["File"] + "|" + str(idx + 1) + "|" + sentence + "\n")
79.             file.write("Number of sentences: " + str(len(SentenceList)) + "\n")
80.

```

7. Appendix Three: IDE Screenshots

A) Tabular Output

```

In [13]: #Printing the output into a text file
with open('TabularOutput.txt','w+', encoding = 'UTF-8') as file:
    for item in AbstractStore:
        Data = item["File"]+"|"+item["OrgType"]+"|"+item["Amount"]+"|"+item["Abstract"]
        file.write(Data+"\n")
        print(Data)

```

Abstract_ID	Sentence_No	Sentence
a9000006 DEB \$179720	1	Commercial exploitation over the past two hundred years drove the great Mysticete whales to near extinct ion. Variation in the sizes of populations prior to exploitation, minimal population size during exploitation and current pop
a9000006 DEB \$179720	2	ulation sizes permit analyses of the effects of differing levels of exploitation on species with different biogeographical di
a9000006 DEB \$179720	3	stributions and life-history characteristics. Dr. Stephen Palumbi at the University of Hawaii will study the genetic populati
a9000006 DEB \$179720	1	on structure of three whale species in this context, the Humpback Whale, the Gray Whale and the Bowhead Whale. The effect of
a9000006 DEB \$179720	2	demographic history will be determined by comparing the genetic structure of the three species. Additional studies will be ca
a9000006 DEB \$179720	3	ried out on the Humpback Whale. The humpback has a world-wide distribution, but the Atlantic and Pacific populations of the
a9000006 DEB \$179720	1	northern hemisphere appear to be discrete populations, as is the population of the southern hemispheric oceans. Each of these
a9000006 DEB \$179720	2	oceanic populations may be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct
a9000006 DEB \$179720	3	t gene pool. This study will provide information on the level of genetic isolation among populations and the levels of gene f
a9000006 DEB \$179720	1	low and genealogical relationships among populations. This detailed genetic information will facilitate international policy
a9000006 DEB \$179720	2	decisions regarding the conservation and management of these magnificent mammals.
a90000031 MCB \$300000	1	Studies of chickens have provided serological and nucleic acid probes useful in defining the major histo
a90000031 MCB \$300000	2	compatibility complex (MHC) in other avian species. Methods used in detecting genetic diversity at loci within the MHC of chi
a90000031 MCB \$300000	3	ckens and mammals will be applied to determining the extent of MHC polymorphism within small populations of ring-necked pheas
a90000031 MCB \$300000	1	ants, wild turkeys, cranes, Andean condors and other species. The knowledge and expertise gained from working with the MHC of
a90000031 MCB \$300000	2	the chicken should make for rapid progress in defining the polymorphism of the MHC in these species and in detecting the poly
a90000031 MCB \$300000	3	morphism of MHC gene pool within small wild and captive populations of these birds. Genes within the major histocompatibility
a90000031 MCB \$300000	1	complex (MHC) are known to encode molecules that provide the context for recognition of foreign antigens by the immune syste
a90000031 MCB \$300000	2	m. Whether a given animal is able to mount an immune response to the challenge of a pathogen is determined, in part, by the a
a90000031 MCB \$300000	3	

B) Sentence Output [For 3 abstracts]

```
In [15]: #Declaring the punkt sentence tokenizer
SentenceDetector = nltk.data.load('tokenizers/punkt/english.pickle')
#Initializing count
count = 0
print("\n Sentence Tokenizer\n")
#Printing the output into a text file
with open("SentenceOutput.txt", "w+", encoding = 'UTF-8') as file:
    for item in AbstractStore:
        #Distribution of Sentence Lengths for the first 3 abstracts
        if count < 3:
            print(" Abstract_ID | Sentence_No | Sentence ")
            print(".....\n")

            file.write("\n Abstract_ID | Sentence_No | Sentence \n.....\n")
            SentenceList = SentenceDetector.tokenize(item["Abstract"].strip())
            for idx, sentence in enumerate(SentenceList):
                print(item["File"] + "|" + str(idx+1) + "|" + sentence)
                file.write(item["File"] + "|" + str(idx+1) + "|" + sentence + "\n")
            print("Number of sentences: ", len(SentenceList))
            file.write("Number of sentences: " + str(len(SentenceList)) + "\n")
            print()
            count = count + 1
        #Distribution of Sentence Lengths for remaining abstracts
        else:
            file.write("\n Abstract_ID | Sentence_No | Sentence \n.....\n")
            SentenceList = SentenceDetector.tokenize(item["Abstract"].strip())
            for idx, sentence in enumerate(SentenceList):
                file.write(item["File"] + "|" + str(idx + 1) + "|" + sentence + "\n")
            file.write("Number of sentences: " + str(len(SentenceList)) + "\n")
```

Sentence Tokenizer

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 analyses of the effects of differing levels of exploitation on species with different biogeographical distributions and life-history characteristics.
a9000006|3|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.
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 populations, as is the population of the southern hemispheric oceans.
a9000006|7|Each of these oceanic populations may be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct gene pool.
a9000006|8|This study will provide information on the level of genetic isolation among populations and the levels of gene flow and genealogical relationships among populations.
a9000006|9|This detailed genetic information will facilitate international policy decisions regarding the conservation and management of these magnificent mammals.
Number of sentences: 9

Abstract_ID | Sentence_No | Sentence

.....
a9000031|1|Studies of chickens have provided serological and nucleic acid probes useful in defining the major histocompatibility complex (MHC) in other avian species.
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 polymorphism within small populations of ring-necked pheasants, wild turkeys, cranes, Andean condors and other species.
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 the MHC in these species and in detecting the polymorphism of MHC gene pool within small wild and captive populations of these birds.

a9000031|4|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.
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 its MHC.
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 severe reduction in population size.
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 requiring intervention for their preservation.
Number of sentences: 9

Abstract_ID | Sentence_No | Sentence

.....
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 deterministic Markov processes; (3) financial economics applications; (4) singular stochastic control; (5) computational methods in stochastic control; (6) stochastic calculus of variations; (7) nonlinear estimation.
a9000038|3|Analytical methods based on viscosity solution techniques for nonlinear differential equations as well as probabilistic methods will be studied.
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 sentences: 4

8. References:

- ❖ <https://realpython.com/working-with-files-in-python/>
- ❖ <https://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html>
- ❖ Course content