NLP HOMEWORK-2

Using regular expressions to analyze NSF abstracts data

Submitted By

Sanjana Rajagopala

SUID - 607219462

INDEX

1. Description of NSF Abstract Dataset……..……..3

2. Pre-Processing – 2B……...4

a. Explanation……..4

b. Analysis…….5

3. Distribution of Sentence lengths……6

a. Explanation……...6

b. Analysis…….6

4. Appendix……..…7

DESCRIPTION OF NSF ABSTRACTS DATASET

BRIEF DESCRIPTION

The dataset mainly deals with the details of the NSF (National Science Foundation) research awards abstracts for the years 1990 to 2003. Each abstract contains the high-level information about the award. This includes the basic information such as Title, Type of Award, Award Number, Estimated Amount, Start and Expiry Date. Additionally, the specifics about research abstract, Sponsors and Field Applications are also stored. It mentions the respective Investigators, Sponsors and Program Managers involved in the research. The description under "Abstract" field gives a glimpse about the what and how aspects of the research work. All the above-mentioned attributes are distributed in a similar format across all the data files – the attributes aligned to the left and the corresponding details placed after specific number of tabs/spaces from the field. The 'NSF Org' attribute holds the abbreviation of the type of NSF organization or department to which the current research work belongs to. The expanded form of the same abbreviation is present as part of the 'Prgm Manager'.

Naming Convention

Each text data file is named based on the "Award" number prefixed by the letter 'a'. Some files are consecutively arranged. For example, a sample data file name is 'a9000006'. All the files are stored in the 'NSF_abstracts' folder.

Number of Documents

Although the complete dataset from the NSF organization consists of 134,161 abstracts, the current dataset that we are using for our analysis contains a total of 4015 documents. Each abstract is contained in one text data file.

PRE-PROCESSING - 2B

EXPLANATION

Before extraction of the required information from the abstract text file, we iterate through the 'NSF_abstracts' folder and read each line in the respective file. This line is passed against the following regular expressions to obtain the details required for the output.

STEP 1: The first pattern looks for the text following the "File" attribute in the file. Similarly, the second pattern looks for the name of the 'NSG Org' and the third attribute extracts the amount following the 'Total Amt.' attribute. The patterns for each of these regular expressions take into consideration the presence of spaces and characters like ":". Each of these matched values are stored in temporary variables.

STEP 2: For extracting the text following the 'Abstract' field, as an initial step the entire matching token is obtained. This is in turn passed into another regular expression pattern that can handle the three cases -

- Paragraph ending with the period '.' – Retains the entire paragraph
- Paragraphs ending with period '.' but followed by '//***' – Retain only the paragraph by stripping off the additional unnecessary characters.
- Paragraphs not ending with a period such as just 'Not Available' – Retains the paragraph as is.

STEP 3: The extracted paragraph contains additional unnecessary newline, tab and space characters. These are replaced a single space character and a well-formed paragraph text is created.

STEP 4: The above obtained File, NSG org, Amount and Abstract field values are joined with space as separator and written into the output file.

ANALYSIS FOR QUESTION 2B

The results of the above processing using Regular expressions presents a variety of insights. These were obtained using manual analysis and few extra code snippets.

Using Manual Analysis –

1. The estimated total amount of the research was diverse – ranging from a large value of 100 thousand dollars to the least value of 0 dollars.

Using Code Snippets –

1. Research works with Maximum and Minimum expected total amount – It was observed that the Research file 'a9000959' associated with the NSF Organization 'OCE' expected the largest amount of 18806079 dollars. Conversely, the research file 'a9000222' associated with INT organization expected the least amount of zero dollars
2. Number of NSF Organizations and research abstracts – It was found that we had a total of 4015 research files and grouped into a total of 41 NSF Organizations.
3. Analysis on NSF Organization -
    a. It was inferred that the 'OCE' organization held the maximum total amount of 124297900 dollars
    b. The 'DMS' organization had in its account the maximum number of research works – 552.
    c. The research files with expected total amount as zero dollars majorly belonged to the 'INT' organization when compared to the other organizations. This justifies the lower total amount under this organization.

DISTRIBUTION OF SENTENCE LENGTHS – 3

EXPLANATION

Most of the required processing is done in the previous step. In addition to it, we make use of the sent_tokenize of NLTK package in Python to tokenize the paragraph under 'Abstract field into individual sentences. This concatenated with the filename obtained previously and the number position of the sentence to form a single line separated by bar "|". The length of the sentence tokens list gives us the total number of sentences in the respective file.

ANALYSIS FOR QUESTION 3

1. By skimming through output file, it can be observed that a majority of the files contain research information in the "Abstract" while a couple of them lacked this. Such files included a "Not Available" text.
2. Around 72 research files contained "Not available" as the text in the Abstract.
3. The research file 'a9003074' with organization as 'SES' contained the maximum number of sentences – 26.
4. The organizations 'MCB', 'INT', 'PHY' consist of number of sentences as one when compared to other organizations in the dataset.
5. A significant difference between the maximum and minimum number of sentences could be observed. Although majority of the files contained number of sentences as single-digit, there were files with more number of sentences and particularly, longer sentences.
6. It was observed that the files with one as the number of sentences (excluding 'Not Available") included the sentence which is comparatively longer and more complex.

This exercise finds similarities to the process of text mining wherein the specific required parts (for example, the file name, abstract, number of sentences etc.) are extracted from the entire text data and used for further analysis.

APPENDIX

OUTPUT

NOTE: Considering the large output that are generated for Questions 2B and 3, only a glimpse of the results has been included here. The output files – **Output_2.txt** and **Output_3.txt** containing the entire set of results are attached with this report separately.

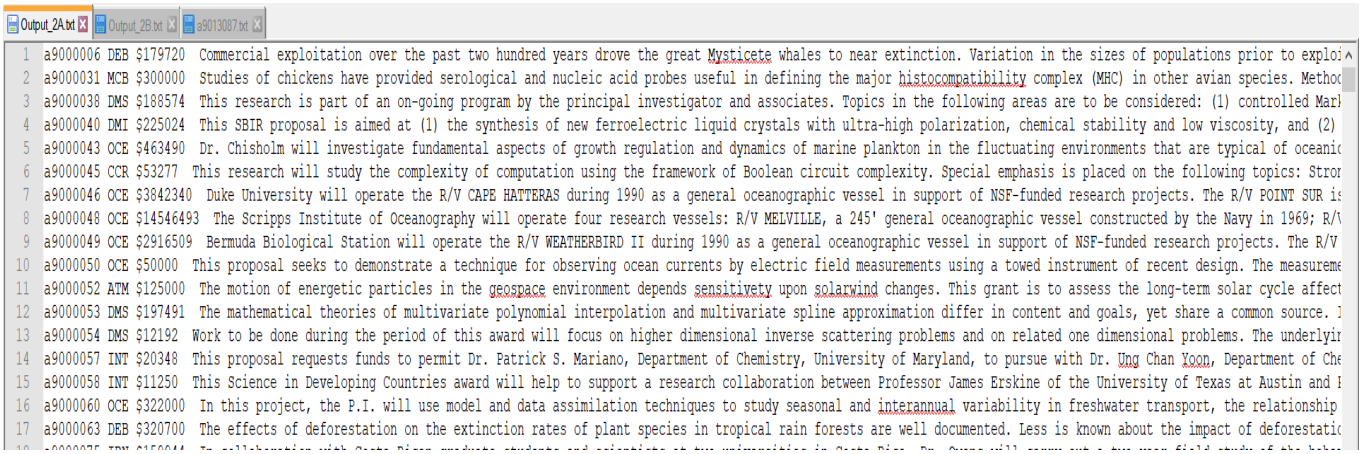
RESULTS FOR QUESTION 2B

a9000006 DEB $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

a9000031 MCB $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

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 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. Analytical methods based on viscosity solution techniques for nonlinear differential equations as well as probabilistic methods will be studied. These theoretical studies are the basis for applied problems ranging from decisions at the stock market level to the control of spaceships

From the text file,



RESULTS FOR QUESTION 3

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

---

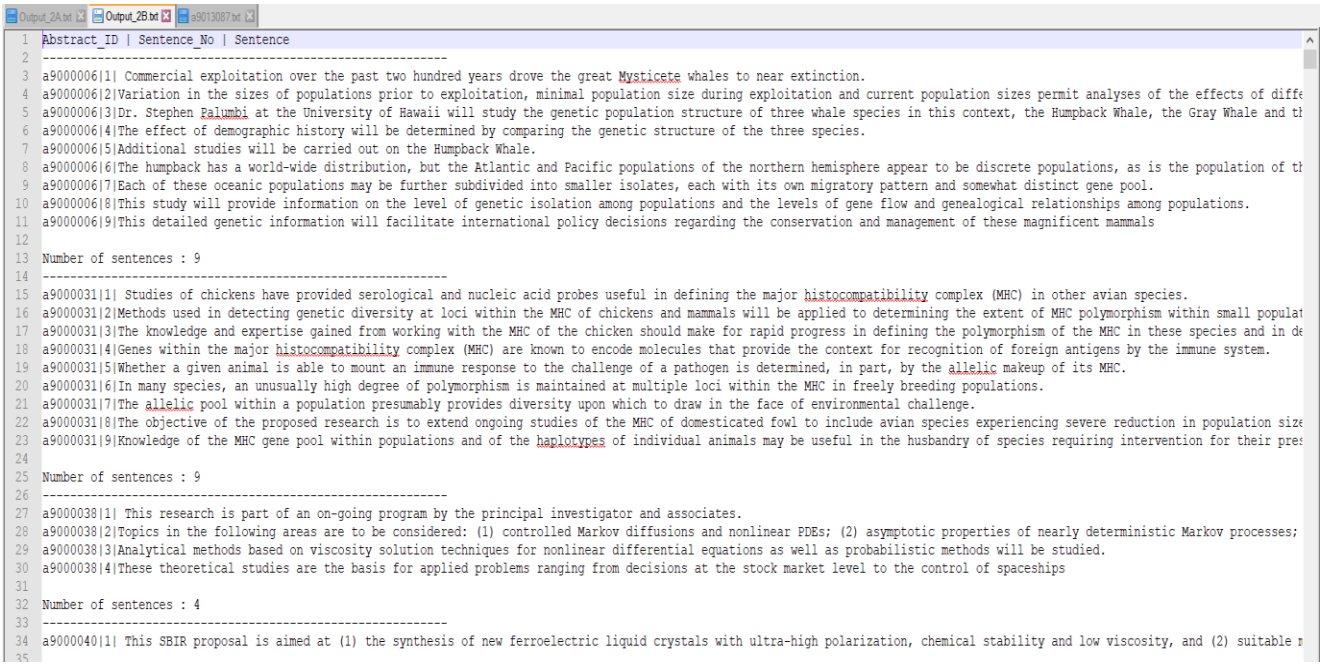
a9000031|1| Studies of chickens have provided serological and nucleic acid probes useful in defining the major histocompatibility complex (MHC) in other avian species.
a9000031|2|Methods used in detecting genetic diversity at loci within the MHC of chickens and mammals will be applied to determining 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
---
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

From the text file,



```
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 diffe
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 th
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 th
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
-----------------------------------------------------------------
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 populat
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 de
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 pres

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) asymptotic properties of nearly deterministic Markov processes;
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
-----------------------------------------------------------------
a9000040|1| This SBIR proposal is aimed at (1) the synthesis of new ferroelectric liquid crystals with ultra-high polarization, chemical stability and low viscosity, and (2) suitable m
```

PYTHON CODE

NOTE: Additionally, the python code file – **nlp_homework_2.py** is attached separately along with the report.

```
# -*- coding: utf-8 -*-
"""
Created on Sat Mar 17 13:22:06 2018

@author: sanja
NLP Homework-2 : Corpus Statistics and Python Programming

"""
#Import the required packages
import nltk
import re

#Define the patterns
#1. Extract the File name
#2. Extract the NSF Org
#3. Extract the Award Amount
#4. Extract the Abstract text

patterns = r''' (?x)
   (?:File\s+:\s*(\w+))
   | NSF\s+Org\s+:\s+(\w+)
   | Total\sAmt\.\s+:\s+(\$\d+)

     '''
#To obtain the entire Abstract text wiht the Abstract word
abs_pattern = "Abstract\s*:(.*)"

#To strip off the additional characters such as //*** after the end of
the paragraph in the Abstract text.
abs_pattern_2 = r''' (?x)
    Abstract\s*:(.*)[!\?\.]
    | Abstract\s*:([\w\s\.,;:]+)


    '''
#NOTE - Might require changing the path of the folder to ensure the
corrent folder is read or written

#Create the output file to write the first set of results
outputFile                                                          =
open("C://Users/sanja/Desktop/CoursesSpring18/IST664/Assignments/Assi
gnment_2/Output_2.txt","w")

#Create the output file to write the second set of results
outputFile_2                                                        =
open("C://Users/sanja/Desktop/CoursesSpring18/IST664/Assignments/Assi
gnment_2/Output_3.txt","w")
```

```
#Write the first initial heading sentence in the file

initial_heading = ["Abstract_ID","Sentence_No","Sentence"]
intitial_sent = " | ".join(initial_heading)
outputFile_2.write(intitial_sent)
outputFile_2.write("\n")
outputFile_2.write("-------------------------------------------------
----------")

#Import the required packages
from nltk.tokenize import sent_tokenize


#Add the abstract id \ text line in the first line
#Define the function that takes the entire file text and file name as
the inputs. It transforms the input into required ersult format
def identifySent(fileText, fileName):
    sentence_string = ""
    sent_count = 1
    sent_total = "Number of sentences : "
    #Use sentence tokenizer to obtain the sentences in the input file
text
    sentence_list = sent_tokenize(fileText)

    outputFile_2.write("\n")
    #Create the line with filename, sentence number and the sentence
    for sent in sentence_list:
        sentence_string += fileName
        sentence_string += "|" +(str(sent_count))
        sentence_string += "|" +sent

        outputFile_2.write(sentence_string)
        outputFile_2.write("\n")
        sent_count +=1
        sentence_string = ""


    outputFile_2.write("\n")
    outputFile_2.write(sent_total + str(len(sentence_list)))
    outputFile_2.write("\n")
    outputFile_2.write("-------------------------------------------
--------------")

    listOfSents.append(len(sentence_list))

#For Analysis of results
listOfFiles = []
listOfAmounts = []
listOfOrgs = []
listOfSents = []
```

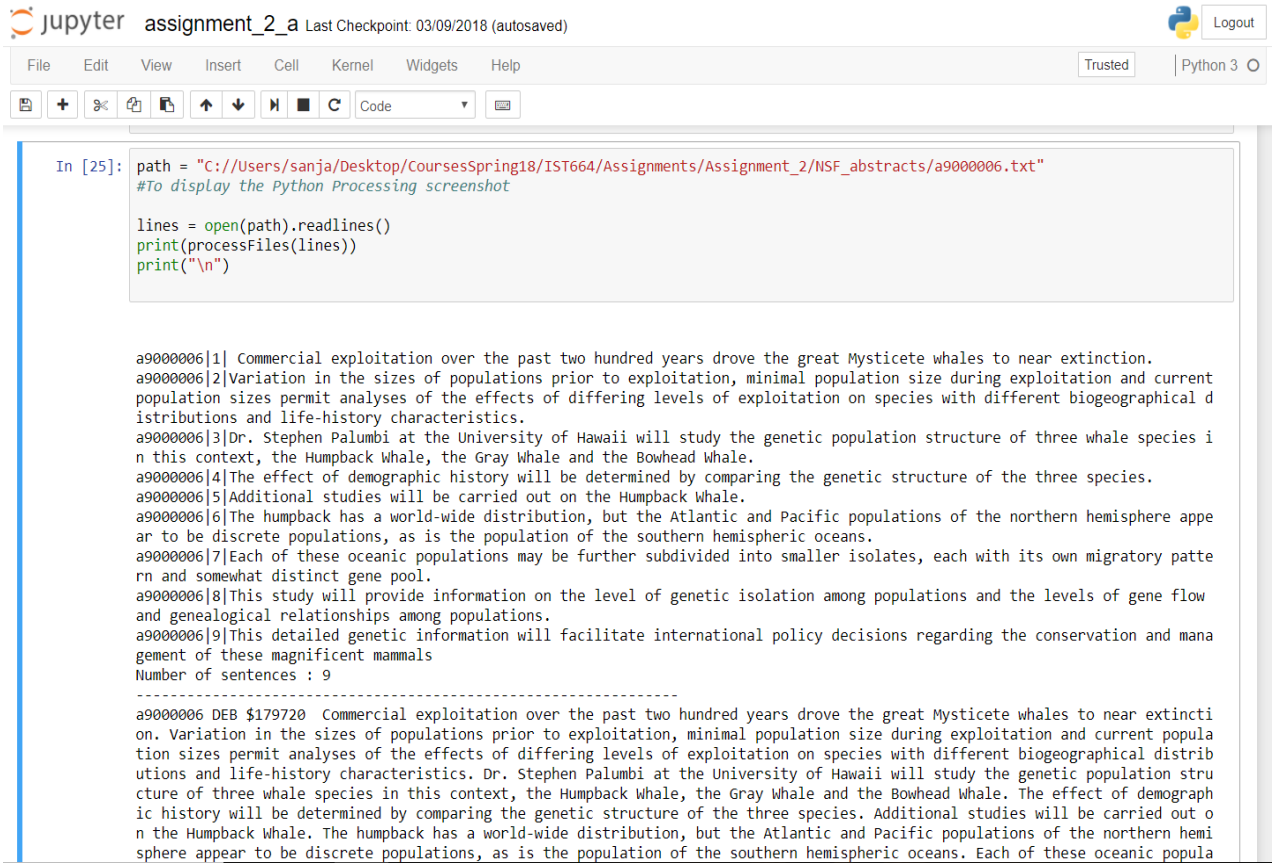
```
#Define the function that does the same processing on each of the files
def processFiles(lines):
    abstract_text = ""
    isAbs = False
    temp_file = ""
    temp_org = ""
    temp_amnt = ""

    #First, check for Abstract text, extract it and maintain in
separate string
    for line in lines:

        if(len(nltk.regexp_tokenize(line, abs_pattern))>0 or isAbs ==
True):
            isAbs = True
            abstract_text += line
    #Add the tokenized terms into the separate list
        else:
            token_list = nltk.regexp_tokenize(line, patterns)
            if(len(token_list)>0):
                if(token_list[0][0]!=''):
                    temp_file = token_list[0][0]
                if(token_list[0][1]!=''):
                    temp_org = token_list[0][1]
                if(token_list[0][2]!=''):
                    temp_amnt = token_list[0][2]


    #Return if the abstract text is empty
    if(abstract_text == ''):
        return ""



    #Check the matching tuple because some abstract text do not end
with . but some end with ***//.
    #To handle both the cases, this processing step is done
    #Replace the newline and extra space characters with single space
character.
    temp_abs = nltk.regexp_tokenize(abstract_text, abs_pattern_2)[0]
    if(temp_abs[0]!=''):
        final_abs_text = temp_abs[0]
    elif(temp_abs[1]!=''):
        final_abs_text = temp_abs[1]


    final_abs_text = final_abs_text.replace('\n',"")
    final_abs_text = re.sub("\s+",' ', final_abs_text)
```

```
    #Create the final string containing all the above terms with space
as delimiter
    final_output_line = [temp_file,temp_org,temp_amnt,final_abs_text]
    final_output_line = " ".join(final_output_line)


    #Call the function that writes the sentence details into the output
file 2B
    identifySent(final_abs_text, temp_file)

    listOfFiles.append(temp_file)
    listOfOrgs.append(temp_org)
    listOfAmounts.append(temp_amnt)

    #Return this modified string
    return final_output_line

import glob
#Read the text files from the NFS_abstracts folder
path                                                                    =
"C://Users/sanja/Desktop/CoursesSpring18/IST664/Assignments/Assignmen
t_2/NSF_abstracts/*.txt"


#For each file perform the same processing steps and write the returned
results into the file
for fl in glob.glob(path):
    lines = open(fl).readlines()
    outputFile.write(processFiles(lines))
    outputFile.write("\n")


#Close the output results file
outputFile.close()
outputFile_2.close()

##################################For           Analysis           of
Results#############################################################

#Getting teh max and min amounts abstracts
listOfAmounts = list(map(lambda x : x[1:], listOfAmounts))
listOfAmounts = list(map(int, listOfAmounts))
max_amnt = max(listOfAmounts)
min_amnt = min(listOfAmounts)

print(max_amnt)
print(min_amnt)

print(listOfAmounts.index(max_amnt))
print(listOfAmounts.index(min_amnt))
print(listOfFiles[364])
```

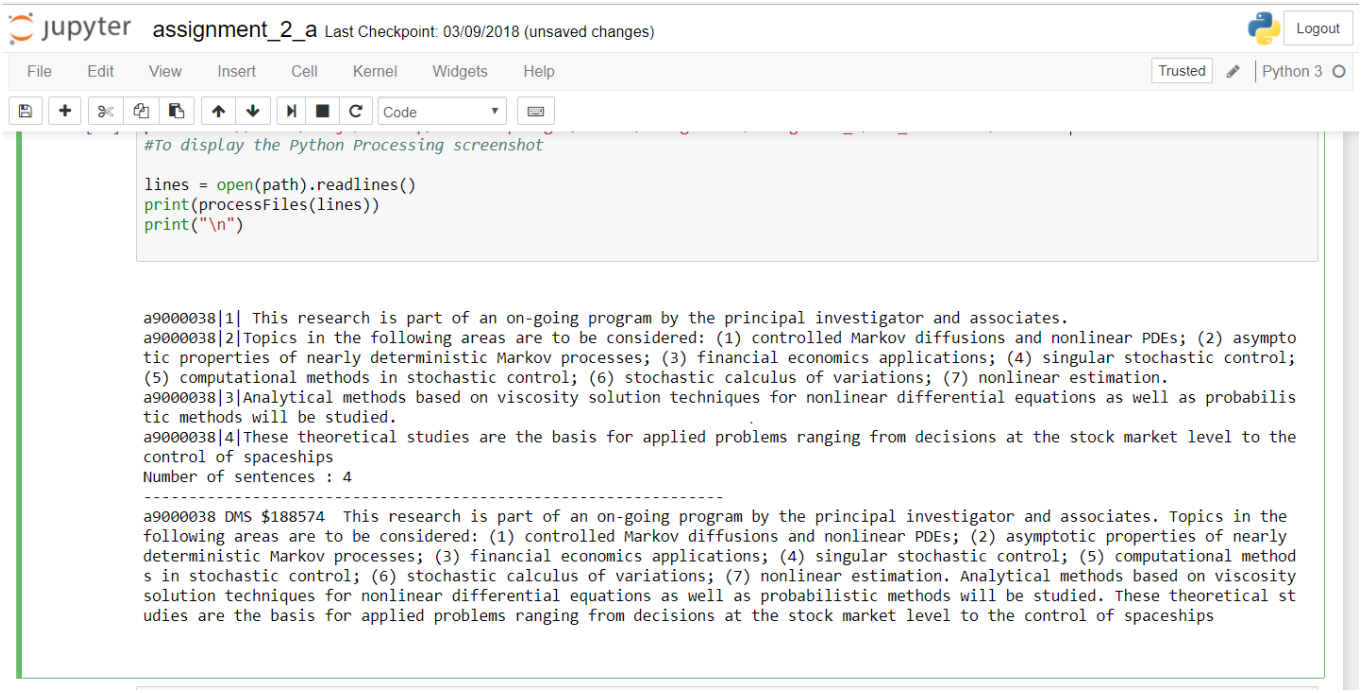
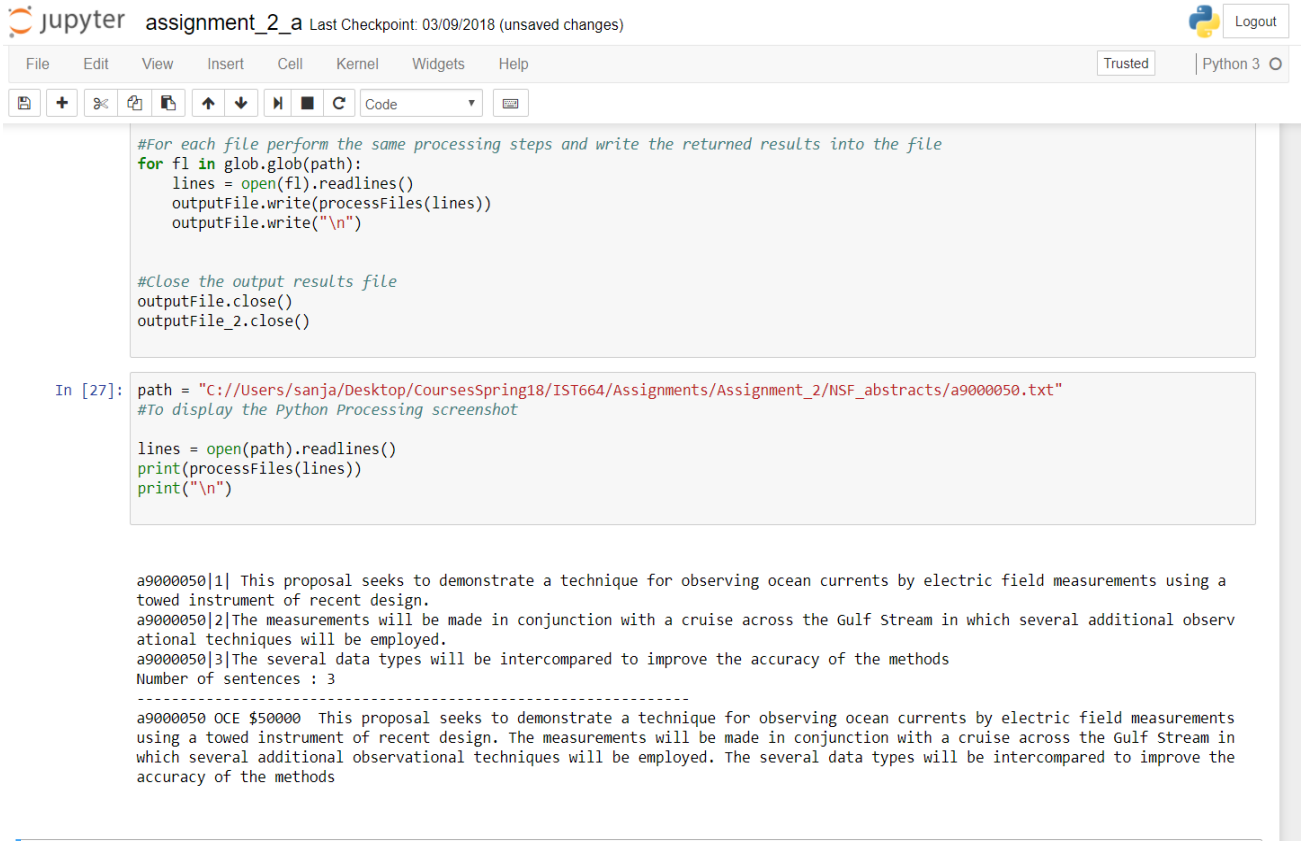
```
print(listOfFiles[68])
print(listOfOrgs[364])
print(listOfOrgs[68])

#Creating a data frame for the results
import pandas as pd

df = pd.DataFrame({'files':listOfFiles, 'orgs': listOfOrgs, 'amnts':
listOfAmounts, 'sents': listOfSents})
orgsGroup = df.groupby('orgs')
orgsGroup.apply(lambda x : max(x['amnts']))
sampl = orgsGroup.apply(lambda x : sum(x['amnts']))
print(sampl)
min(sampl)
max(sampl) #---> OCE

sampl_2 = orgsGroup.apply(lambda x : len(x['amnts']))
print(sampl_2)
max(sampl_2) #---->DMS

sampl_3 = (orgsGroup.apply(lambda x : max(x['sents'])))
print(sampl_3)
max(sampl_3) #---->SES

#Approximately 72 abstracts with Not Available abstract texts

for i in df.index:
    if(df.loc[i,'amnts']==0):
        print(df.loc[i,'orgs'])

#Majorly INT org with 0 amounts
```

PYTHON PROCESSING SCREENSHOTS

1. Output including the sentences for first text file in the NSF_abstracts folder



2. Output including the sentences for second text file in the NSF_abstracts folder



3. Output including the sentences for third text file in the NSF_abstracts folder