Using Regex Expressions to Analyze NSF Abstracts Data

Homework2

Natural Language Processing

Komal Gujarathi
211778351

Table of Contents

1. Dataset

Analyze a subset of a publicly available collection of NSF (National Science Foundation) research awards abstracts spanning 1990-2003. The complete dataset consists of 134161 abstracts describing NSF Awards for basic research, bag of word data files extracted from the abstracts, list of words used for indexing the bag of word data. But for this assignment we will only use the part of the abstract data.

2. Pre-processing (50%)

2.1. Review the dataset and describe the characteristics of the corpus briefly such as naming conventions of its files, number of documents it contains, etc. (10%)

2.1.1. The dataset we will be analyzing has total 4016 documents. Each document is the abstract describing NSF Awards for the basic research.

2.1.2. Naming convention followed for these files is character 'a' followed by Award Number. e.g. if Award Number is 900006, then file name is 'a900006'.

2.1.3. In each abstract, there is abstract title, the name of NSF organization that gives this award, award number, sponsor, award amount, start date, end date and the abstract text.

2.1.4. Minimum and maximum number length of all these abstract texts is 1 and 26 sentences respectively.

2.1.5. The maximum grant is given by NSF Organization 'OCE' which amounts to $18806079, while minimum grant given is $0.

*(all the figures are obtained by writing the python scripts)

2.2. 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.

I have used regular expression to find out the required data from the input files. And then iterated over all the dataset to gather the data and dump that data in the 'result1.txt' file.
I have attached the screen shot of the python code and the sample output below.

Python Code

```
# grants_amount is a list of all the grants
grant_amounts = []
# this information is used to get the organization granting the max grant
organization_giving_max_grant = ''

resultfile = open('result1.txt', 'w+')
for i in range(0,len(file_contents)-1):
    shorttext = file_contents[i]

    # extract the file text
    pword = re.compile('File *: *(\w+)')
    file_field = re.findall(pword, shorttext)
    resultfile.write(file_field[0])

    resultfile.write('\t')
    pword = re.compile('NSF Org *: *(\w+)')
    nsf_field = re.findall(pword, shorttext)
    resultfile.write(nsf_field[0])

    resultfile.write(' ')
    pword = re.compile('Total Amt\. *: *([$]\d+)')
    amount_field = re.findall(pword, shorttext)
    resultfile.write(amount_field[0])


    # getting the amount in $ to find out the maximum grant amount
    # from all the awards (I used this information to describe the text)
    # in question 2A.
    pword = re.compile('Total Amt\. *: *[$](\d+)')
    res = re. findall(pword, shorttext)
    grant_amounts.append(int(res[0]))
    if (int(res[0]) == 18806079):
        organization_giving_max_grant = nsf_field[0]

    resultfile.write(' ')
    pword = re.compile('Abstract *: *\n[ \t]*((?s).*)')
    # text contains string with extra white spaces and \n characters
    text = re.findall(pword, shorttext)
    newtext = text[0].replace('\n','')
    abstract_field = ' '.join(newtext.split())
    resultfile.write(abstract_field)
    resultfile.write('\n')
resultfile.close()
```

Output File Screen Shot (result1.txt)

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.
a9000040 DMI $225024 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 modifications that yield functional materials with broad operational temperature ranges and fast electro-optic response times for use as components in commercial ferroelectric liquid crystals devices that are direct-drive or voltage limited.
a9000043 OCE $463490 Dr. Chisholm will investigate fundamental aspects of growth regulation and dynamics of marine plankton in the fluctuating environments that are typical of oceanic regimes. This understanding is essential for modelling and designing studies of marine productivity and food web dynamics. Specifically, they will: * Study the diatom life cycle, to better understand what environmental and genetic factors control the switch from vegetative to sexual reproduction; * Investigate the cell cycle in cyanobacteria and marine bacteria, to better understand how their DNA cycle is regulated under changing growth conditions, and why aspects of the cycle deviate from the classic E. coli model; * Develop a multichannel, high-sensitivity (and relatively inexpensive) flow cytometer, which will facilitate the study small picoplankton and marine bacteria. The results of this work will augment our understanding of microbial growth rates in the sea and phytoplankton population genetics; it should also contribute methodological advances for the study of marine bacteria at sea. Moreover, it could enhance our fundamental understanding of microbial physiology by revealing features of marine organisms which deviate from classical paradigms.
a9000045 CCR $53277 This research will study the complexity of computation using the framework of Boolean circuit complexity. Special emphasis is placed on the following topics: Strong separations of circuit classes: If known separations of small circuit complexity classes could be strengthened, it would imply separations on larger time- and space-complexity classes. This connection will be investigated, using the notion of "immunity" as a tool. Width-bounded reducibility: This notion will be used as a tool to investigate the relationships among "similar" complexity classes. This project also investigates threshold circuits, an structure of the complexity class P/poly.
a9000046 OCE $3842340 Duke University will operate the R/V CAPE HATTERAS during 1990 as a general oceanographic vessel in support of NSF-funded research projects. The R/V POINT SUR is a 135' general research vessel constructed in 1981 and owned by the National Science Foundation. Duke operates the CAPE HATTERAS under a charter agreement with NSF. The ship operates primarily off the U.S. east coast from Maine to Florida. This vessel is part of a fleet used by the National Science Foundation to support oceanographic research projects. Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members. An increasing number of research projects require equipment that must be permanently installed on a ship and for which the ship must be specifically designed. Such equipment also requires highly trained crew members for maintenance and operation. These vessels do not operate in the same manner as general cargo or fishing vessels, and therefore, NSF supports the operation of a variety of vessels specifically dedicated to oceanographic research. These vessels are operated by universities and research institutions around the country.
a9000048 OCE $14546493 The Scripps Institute of Oceanography will operate four research vessels: R/V MELVILLE, a 245' general oceanographic vessel constructed by the Navy in 1969; R/V THOMAS WASHINGTON, a 200' research vessel constructed by the Navy in 1965; R/V NEW HORIZON, a 170' research vessel constructed by the University of California in 1978; and R/V ROBERT GORDON SPROUL, a 125' vessel built in 1981 and subsequently converted for research purposes. These vessels are part of a fleet used by the National Science Foundation to support oceanographic research projects. Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members. An increasing number of research projects require equipment that must be permanently installed on a ship and for which the ship must be specifically designed. Such equipment also requires highly trained crew members for maintenance and operation. These vessels do not operate in the same manner as general cargo or fishing vessels, and therefore, NSF supports the operation of a variety of vessels specifically dedicated to oceanographic research. These vessels are operated by universities and research institutions around the country. The R/V's T.WASHINGTON and MELVILLE operate worldwide, while the R/V NEW HORIZON operates primarily in the northeastern Pacific. The R/V SPROUL operates primarily in the coastal waters of California.
a9000049 OCE $2916509 Bermuda Biological Station will operate the R/V WEATHERBIRD II during 1990 as a general oceanographic vessel in support of NSF-funded research projects. The R/V WEATHERBIRD II is a 115' general research vessel that was originally converted in 1989 and is owned by the Bermuda Biological Station. Additional conversion work on the vessel will be conducted in phases during 1990 and 1991. The ship operates primarily in the vicinity of Bermuda. This vessel is part of a fleet used by the National Science Foundation to support oceanographic research projects. Most oceanographic research projects require highly specialized equipment and extensive support from a ship's crew members. An increasing number of research projects require equipment that must be permanently installed on a ship and for which the ship must be specifically designed. Such equipment also requires highly trained crew members for maintenance and operation. These vessels do not operate in the same manner as general cargo or fishing vessels, and therefore, NSF supports the operation of a variety of vessels specifically dedicated to oceanographic research. These vessels are operated by universities and research institutions around the country.
a9000050 OCE $50000 This proposal seeks to demonstrate a technique for observing ocean currents by electric field measurements using a towed instrument of recent design. The measurements will be made in conjunction with a cruise across the Gulf Stream in which several additional observational techniques will be employed. The several data types will be intercompared to improve the accuracy of the methods.

3. Distribution of sentence lengths (50%)

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

Python Code -

```
# min_abstract and max_abstract len is used to find the max len of the abstract
# in the given data set
min_abstract_len = 99999
max_abstract_len = 0

resultfile = open('result2.txt', 'w+')
resultfile.write('Abstract_ID | Sentence_No | Sentence\n')
resultfile.write('-----------------------------------\n')
for i in range(0,len(file_contents)-1):
    shorttext = file_contents[i]

    # extract the file text
    pword = re.compile('File *: *(\w+)')
    file_field = re.findall(pword, shorttext)

    pword = re.compile('Abstract *: *\n[ \t]*((?s).*)')
    # text contains string with extra white spaces and \n characters
    text = re.findall(pword, shorttext)
    newtext = text[0].replace('\n','')
    abstract_field = ' '.join(newtext.split())

    sent_tokenize_list = sent_tokenize(abstract_field)

    for i in range (len(sent_tokenize_list)):
        resultfile.write(file_field[0])
        resultfile.write('|')
        resultfile.write(str(i+1))
        resultfile.write('|')
        resultfile.write(sent_tokenize_list[i])
        resultfile.write('\n')
    last_line = 'Number of sentences : ' + str(len(sent_tokenize_list)) + '\n'
    max_abstract_len = max(max_abstract_len,len(sent_tokenize_list))
    min_abstract_len = min(min_abstract_len, len(sent_tokenize_list))
    resultfile.write(last_line)

resultfile.close()
```

Output File Screen Shot (result2.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 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
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 modifications that yield functional materials with broad operational temperature ranges and fast electro-optic response times for use as components in commercial ferroelectric liquid crystals devices that are direct-drive or voltage limited.
Number of sentences : 1
a9000043|1|Dr. Chisholm will investigate fundamental aspects of growth regulation and dynamics of marine plankton in the fluctuating environments that are typical of oceanic regimes.
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 vegetative to sexual reproduction; * Investigate the cell cycle in cyanobacteria and marine bacteria, to better understand how their DNA cycle is regulated under changing growth conditions, and why aspects of the cycle deviate from the classic E. coli model; * Develop a multichannel, high-sensitivity (and relatively inexpensive) flow cytometer, which will facilitate the study small picoplankton and marine bacteria.
a9000043|4|The results of this work will augment our understanding of microbial growth rates in the sea and phytoplankton population genetics; it should also contribute methodological advances for the study of marine bacteria at sea.
a9000043|5|Moreover, it could enhance our fundamental understanding of microbial physiology by revealing features of marine organisms which deviate from classical paradigms.
Number of sentences : 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 classes could be strengthened, it would imply separations on larger time- and space-complexity classes.
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 sentences : 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 specifically designed.
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 of vessels specifically dedicated to oceanographic research.
a9000046|10|These vessels are operated by universities and research institutions around the country.
Number of sentences : 10
```

4. Appendix

Output and Python Processing

```
import nltk
import re
from nltk.tokenize import sent_tokenize


from nltk.corpus import PlaintextCorpusReader
mycorpus = PlaintextCorpusReader('.', '.*\.txt')
type(mycorpus.fileids())
```

```
num_docs = len(mycorpus.fileids())  # mycorpus.fileids() gives the list of
num_docs
```

```
4016
```

```
# extract the contents of all the files in the file_contents array
file_ids = mycorpus.fileids()
file_contents = []  # file_contents is the array of all the file contents
for i in range(len(file_ids)):
    file = file_ids[i]
    file_handle = open(file,'r',encoding = "ISO-8859-1")
    file_content = file_handle.read()
    # file content has the actual content of each of the file
    file_contents.append(file_content)
    file_handle.close()
```

```
# sample demo of pattern matching
import re
shorttext = file_contents[5]

pword = re.compile('File *: *(\w+)')
re.findall(pword, shorttext)
```

```
['a9000045']
```

```
pword = re.compile('NSF Org *: *(\w+)')
re.findall(pword, shorttext)
```

```
['CCR']
```

```
pword = re.compile('Total Amt\. *: *([$]\d+)')
re.findall(pword, shorttext)

pword = re.compile('Total Amt\. *: *[$](\d+)')
res = re.findall(pword, shorttext)
type(res[0])
int(res[0])
```

```
53277
```

```
# grants_amount is a list of all the grants
grant_amounts = []
# this information is used to get the organization granting the max grant
organization_giving_max_grant = ''

resultfile = open('result1.txt', 'w+')
for i in range(0,len(file_contents)-1):
    shorttext = file_contents[i]

    # extract the file text
    pword = re.compile('File *: *(\w+)')
    file_field = re.findall(pword, shorttext)
    resultfile.write(file_field[0])

    resultfile.write('\t')
    pword = re.compile('NSF Org *: *(\w+)')
    nsf_field = re.findall(pword, shorttext)
    resultfile.write(nsf_field[0])

    resultfile.write(' ')
    pword = re.compile('Total Amt\. *: *([$]\d+)')
    amount_field = re.findall(pword, shorttext)
    resultfile.write(amount_field[0])


    # getting the amount in $ to find out the maximum grant amount
    # from all the awards (I used this information to describe the text)
    # in question 2A.
    pword = re.compile('Total Amt\. *: *[$](\d+)')
    res = re. findall(pword, shorttext)
    grant_amounts.append(int(res[0]))
    if (int(res[0]) == 18806079):
        organization_giving_max_grant = nsf_field[0]

    resultfile.write(' ')
    pword = re.compile('Abstract *: *\n[ \t]*((?s).*)')
    # text contains string with extra white spaces and \n characters
    text = re.findall(pword, shorttext)
    newtext = text[0].replace('\n','')
    abstract_field = ' '.join(newtext.split())
    resultfile.write(abstract_field)
    resultfile.write('\n')
resultfile.close()
```

```
# min_abstract and max_abstract len is used to find the max len of the abstract
# in the given data set
min_abstract_len = 99999
max_abstract_len = 0

resultfile = open('result2.txt', 'w+')
resultfile.write('Abstract_ID | Sentence_No | Sentence\n')
resultfile.write('-------------------------------------\n')
for i in range(0,len(file_contents)-1):
    shorttext = file_contents[i]

    # extract the file text
    pword = re.compile('File *: *(\w+)')
    file_field = re.findall(pword, shorttext)

    pword = re.compile('Abstract *: *\n[ \t]*((?s).*)')
    # text contains string with extra white spaces and \n characters
    text = re.findall(pword, shorttext)
    newtext = text[0].replace('\n','')
    abstract_field = ' '.join(newtext.split())

    sent_tokenize_list = sent_tokenize(abstract_field)

    for i in range (len(sent_tokenize_list)):
        resultfile.write(file_field[0])
        resultfile.write('|')
        resultfile.write(str(i+1))
        resultfile.write('|')
        resultfile.write(sent_tokenize_list[i])
        resultfile.write('\n')
    last_line = 'Number of sentences : ' + str(len(sent_tokenize_list)) + '\n'
    max_abstract_len = max(max_abstract_len,len(sent_tokenize_list))
    min_abstract_len = min(min_abstract_len, len(sent_tokenize_list))
    resultfile.write(last_line)

resultfile.close()
```

```
print('max grant amount : '+ str(max(grant_amounts)))
print('min grant amount : '+ str(min(grant_amounts)))
print('organizarion giving max grant : ' + organization_giving_max_grant)
print('max abstract length : ' + str(max_abstract_len))
print('min abstract length : ' + str(min_abstract_len))
```

```
max grant amount : 18806079
min grant amount : 0
organizarion giving max grant : OCE
max abstract length : 26
min abstract length : 1
```