NLP Homework 2

Due Sunday, March 18, 11:59 pm.

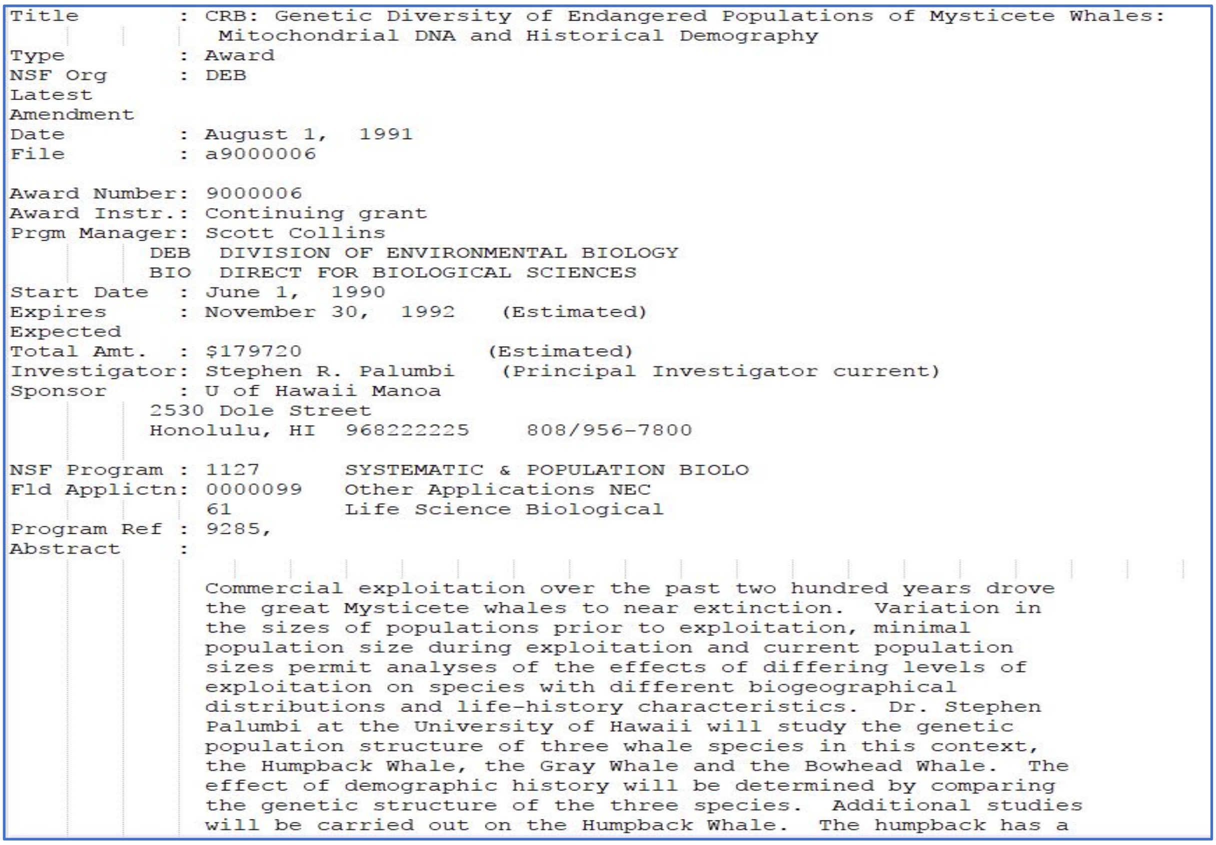
Using Regular Expressions to Analyze NSF abstracts Data

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

1) Dataset

In this assignment, you will analyze a subset of a publicly available collection of NSF (National Science Foundation) research awards abstracts spanning 1990 - 2003. The complete dataset consists of (1) 134,161 abstracts describing NSF awards for basic research, (b) bag-of-word data files extracted from the abstracts, (c) a list of words used for indexing the bag-of-word data. For this assignment, we will use only part of the abstracts data. For the complete details of this dataset, please refer to the following URL: https://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

Each abstract is contained in one txt file that is located in a sub-folder within a subset of the folders. For your convenience, the data set used for your tasks is already extracted and arranged in three folders. The dataset is available for download in the “Assignment” folder in the course web site at Blackboard. Here is a screenshot of an example input file ‘a9000006.txt’:

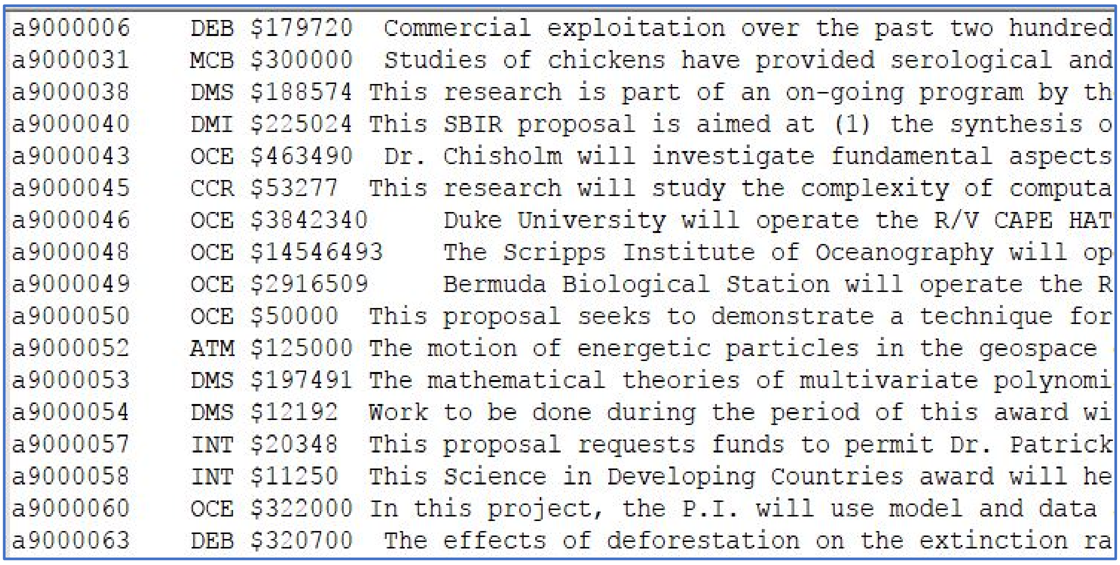


```
Title       : CRB: Genetic Diversity of Endangered Populations of Mysticete Whales:
               Mitochondrial DNA and Historical Demography
Type        : Award
NSF Org     : DEB
Latest
Amendment
Date        : August 1,  1991
File        : a9000006

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

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

              Commercial exploitation over the past two hundred years drove
              the great Mysticete whales to near extinction.  Variation in
              the sizes of populations prior to exploitation, minimal
              population size during exploitation and current population
              sizes permit analyses of the effects of differing levels of
              exploitation on species with different biogeographical
              distributions and life-history characteristics.  Dr. Stephen
              Palumbi at the University of Hawaii will study the genetic
              population structure of three whale species in this context,
              the Humpback Whale, the Gray Whale and the Bowhead Whale.  The
              effect of demographic history will be determined by comparing
              the genetic structure of the three species.  Additional studies
              will be carried out on the Humpback Whale.  The humpback has a
```

2) Pre-processing (50%)

A) 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. (10%, a short paragraph, no more than 300 words)

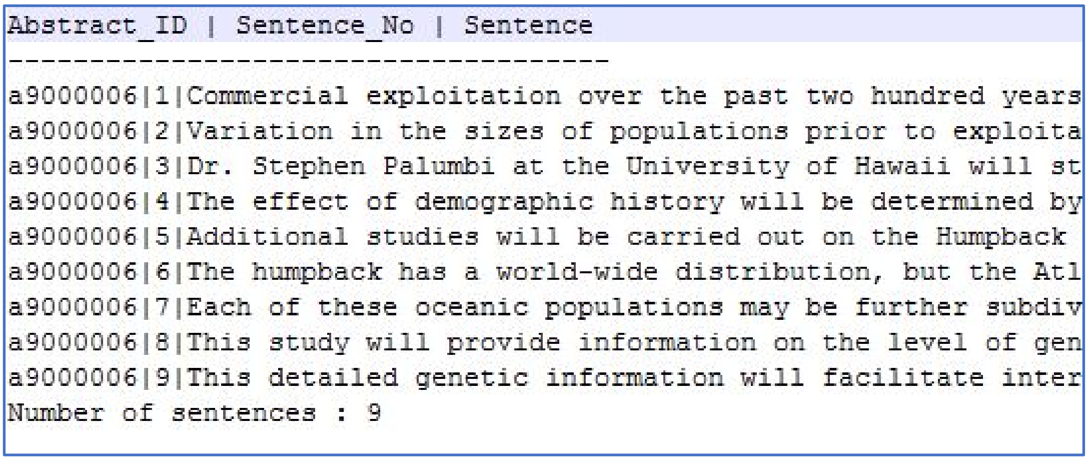
B) Next, you will write a Python code that reads in each abstract and extract the abstract identity (‘File’), NSF organization (‘NSF org’), the award amount, and abstract text. Please submit the output with a tabular format included. The output may look like as follows. (40%)



```
a9000006    DEB $179720  Commercial exploitation over the past two hundred
a9000031    MCB $300000  Studies of chickens have provided serological and
a9000038    DMS $188574 This research is part of an on-going program by th
a9000040    DMI $225024 This SBIR proposal is aimed at (1) the synthesis o
a9000043    OCE $463490  Dr. Chisholm will investigate fundamental aspects
a9000045    CCR $53277  This research will study the complexity of computa
a9000046    OCE $3842340    Duke University will operate the R/V CAPE HAT
a9000048    OCE $14546493   The Scripps Institute of Oceanography will op
a9000049    OCE $2916509    Bermuda Biological Station will operate the R
a9000050    OCE $50000  This proposal seeks to demonstrate a technique for
a9000052    ATM $125000 The motion of energetic particles in the geospace
a9000053    DMS $197491 The mathematical theories of multivariate polynomi
a9000054    DMS $12192  Work to be done during the period of this award wi
a9000057    INT $20348  This proposal requests funds to permit Dr. Patrick
a9000058    INT $11250  This Science in Developing Countries award will he
a9000060    OCE $322000 In this project, the P.I. will use model and data
a9000063    DEB $320700  The effects of deforestation on the extinction ra
```

3) Distribution of sentence lengths (50%)

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**. For example, the first file (a9000006.txt) has 9 sentences and the output would be as follows:



```
Abstract_ID | Sentence_No | Sentence
-------------------------------------
a9000006|1|Commercial exploitation over the past two hundred years
a9000006|2|Variation in the sizes of populations prior to exploita
a9000006|3|Dr. Stephen Palumbi at the University of Hawaii will st
a9000006|4|The effect of demographic history will be determined by
a9000006|5|Additional studies will be carried out on the Humpback
a9000006|6|The humpback has a world-wide distribution, but the Atl
a9000006|7|Each of these oceanic populations may be further subdiv
a9000006|8|This study will provide information on the level of gen
a9000006|9|This detailed genetic information will facilitate inter
Number of sentences : 9
```

How to Submit Homework:

Write a report to describe the output and the processing options you choose for the tasks. In the appendix, provide your Python code and Python output. Please submit your report in the PDF format. Go to the Blackboard system and the Assignment for Homework 2. Attach your report file and submit. Your report should include:

1) your output
2) the Python code
3) the Python processing screenshot (the sentences for three abstracts)