Using Regex Expressions to Analyze NSF Abstracts Data

Homework2

Natural Language Processing

Komal Gujarathi

211778351

Table of Contents

[1. Dataset 3](#_Toc509146082)

[2. Pre-processing (50%) 3](#_Toc509146083)

[3. Distribution of sentence lengths (50%) 5](#_Toc509146084)

[4. Appendix 8](#_Toc509146085)

# 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.

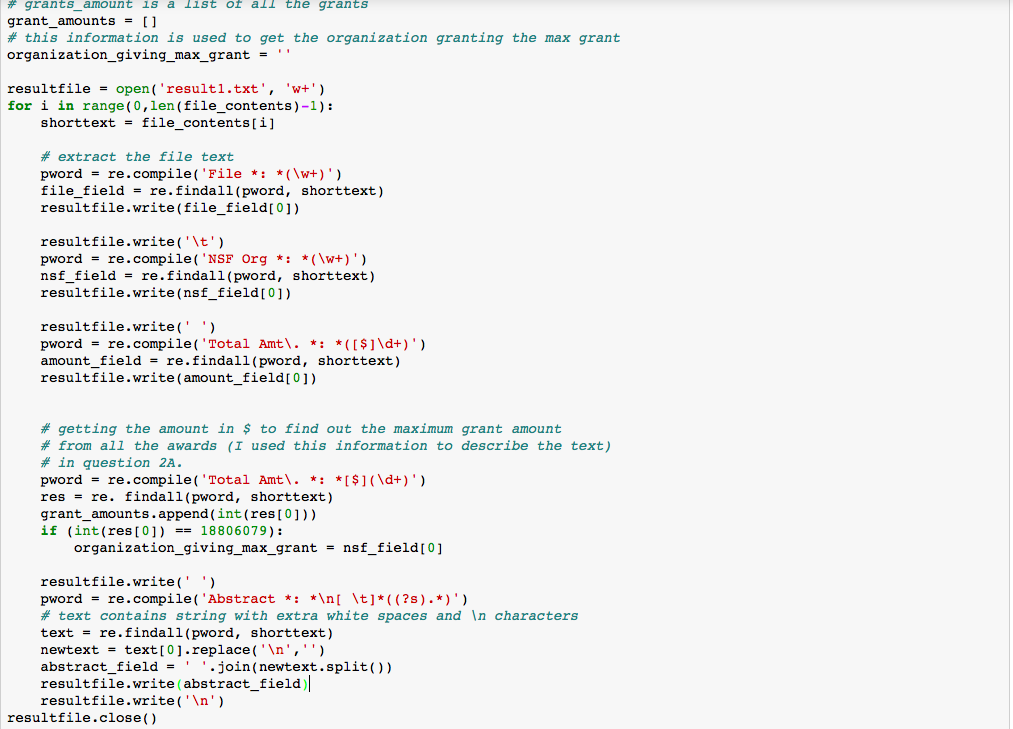
\*(all the figures are obtained by writing the python scripts)

**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.**

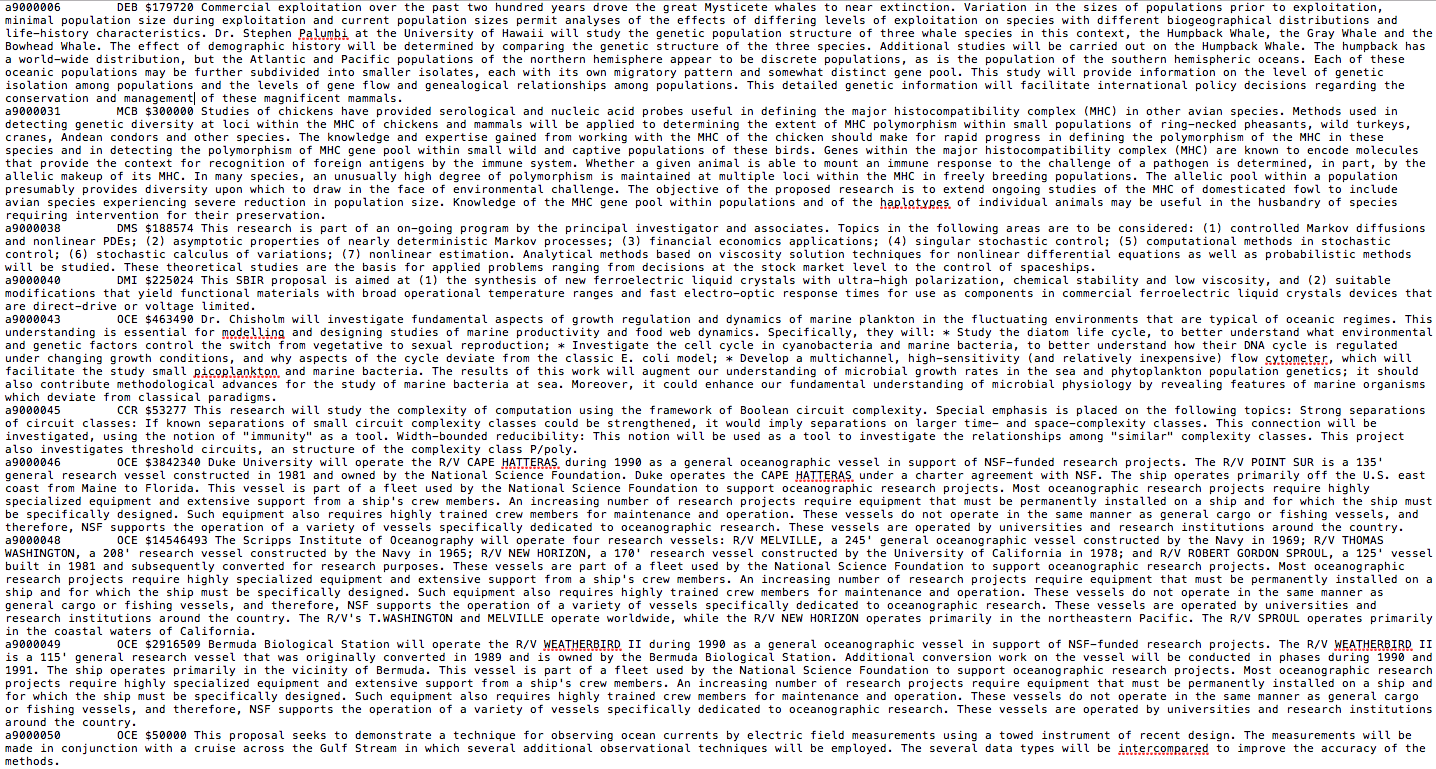
I have used regular expression to find out the required data from the input files. And then iterated over all the dataset to gather the data and dump that data in the ‘result1.txt’ file.

I have attached the screen shot of the python code and the sample output below.

Python Code



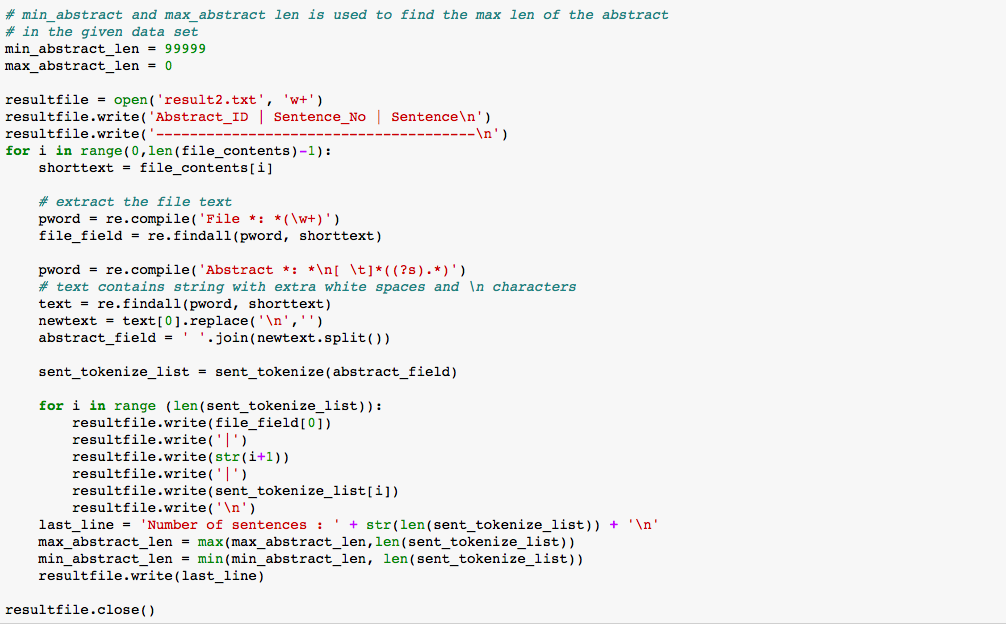
Output File Screen Shot (result1.txt)



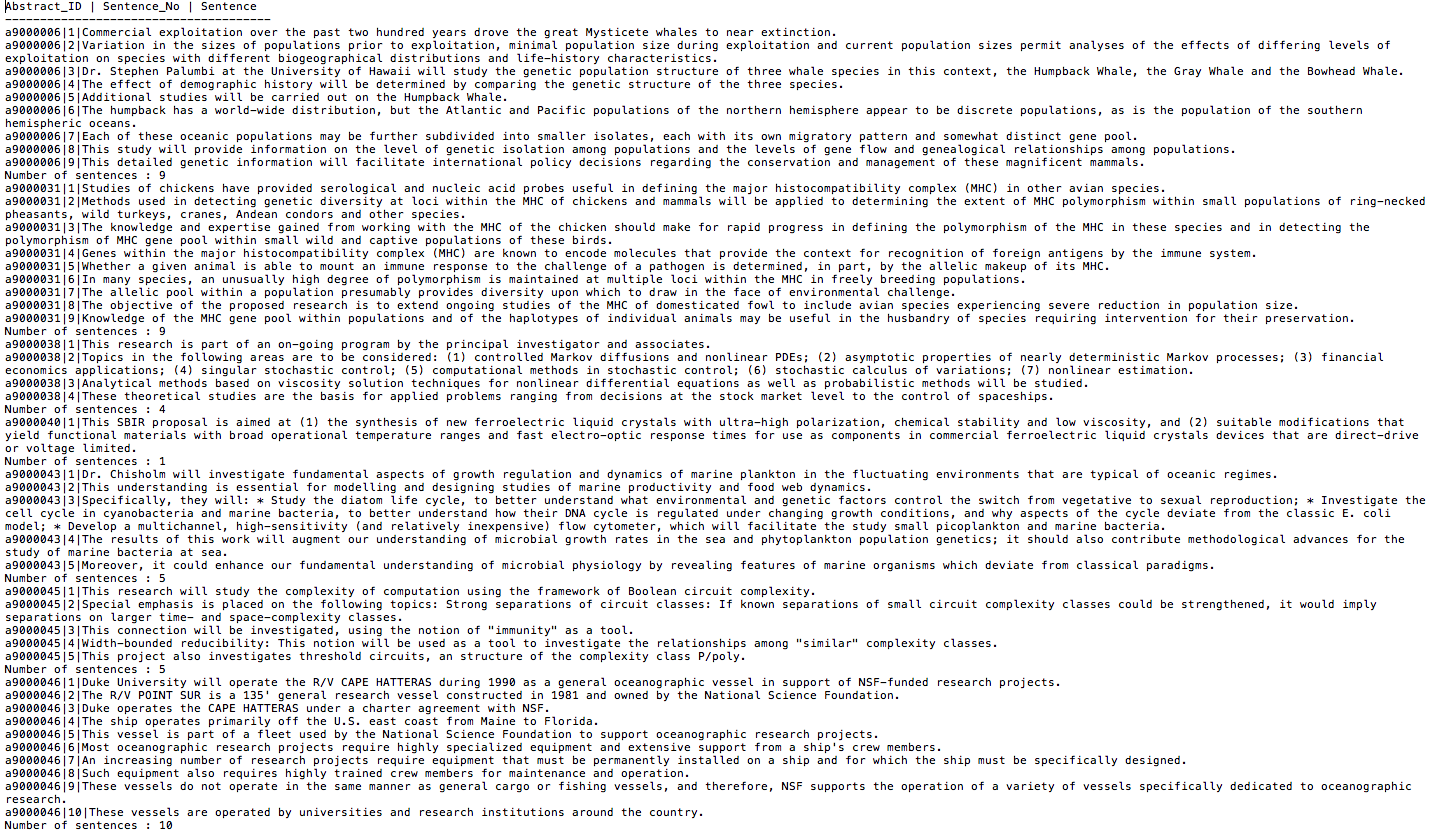
# 3. Distribution of sentence lengths (50%)

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 -



Output File Screen Shot (result2.txt) -



# 4. Appendix

Output and Python Processing

