Assignment 1 – Manipulating Text
LIS590TX
Spring 2016

Motivation: Text mining applications identify patterns from vast reservoirs of existing information. The goal of this assignment is to give you practice with manipulating text in preparation for your final project. This assignment reflects the first three activities in the knowledge discovery framework presented in class – data selection, pre-processing, and transformation. The results of this assignment will feed into assignment 2, which will explore feature selection.

Instructions:

Part 1) Data selection

This assignment will use draw on NSF research awards abstracts between 1990 and 2003. The complete dataset comprises (a) 129,000 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. In this assignment we will use only the first zipped file (called part1.zip), which comprises 51979 abstracts. Your first task is to download a local copy of these abstracts. You can find the abstracts that we will use in moodle, but complete details of this dataset in moodle or at the following URL: http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

Example input file a9000006.txt, with sections required for this assignment highlighted.

```
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 ← Abstract Identity

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
              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.
```

Part 2) Preprocessing

Write a computer program that reads in each abstract and extracts the abstract identity the, NSF Organization, the award amount and abstract text. You may use any programming or scripting language you like. Your program should keep track of the number of awards in each NSF Organization and the total amount awarded to that organization.

Part 3) Transformation

Identify sentences in the abstract. You may use sentence tokenizers in python or java classes, or re-implement the sentence tokenizer algorithm described in the text book. Your program should output should contain the abstract identity, the sentence number, and the sentence text delimited with a bar (|) and the number of sentences in each file. For example, file a9000006.txt has 8 sentences and the first 5 sentences would be:

```
9000006|1|Commercial exploitation over the past two hundred years drove the great Mysticete
whales to near extinction.
9000006|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.
9000006|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.
9000006|4|The effect of demographic history will be determined by comparing the genetic structure
of the three species.
9000006|5|Additional studies will be carried out on the Humpback Whale.
```

What to submit

Submit your code, the number of awards in each NSF Organization and the total number of awards made to that NSF organization, the sentences for three abstracts, and complete the assignment 1 template in moodle with the distribution of sentence lengths that you found. The template looks like this:

```
  0 x
  1 y
  …
 10 z
  …
```

Where you will add x, which is the number of files without any abstracts (i.e. 0 sentences); y, the number of abstracts with 1 sentence; z, the number of abstracts with 10 sentences etc. Don’t forget to add the programming language that you used and your initials to the template so that I can pull these together in order for us to discuss further in class.

Note for students who have already worked with this data set in another course and have this framework in place. You should also process the NSF abstracts as described above and submit the same required materials as other students. In addition you should apply your existing program to one another collection and submit the same summary data (i.e. number of texts in each category and the distribution of sentence length). Ideally the second collection will be the data that you will use for your project, but use the Amazon beauty or health reviews if your project text doesn’t have sentences (e.g. twitter data).

Grading: This assignment is worth 10% of your grade and more importantly will give you the framework that you will extend in assignment 2 and use for your project.