IS597MCL: Building Machine Learning Pipelines Assignment

Data Selection

This assignment will use data from NSF research awards abstracts between 1990 and 2003. The complete dataset contains (a) 129,000 abstracts describing NSF awards for basic research, (b) bag-of-word data files extracted from the abstracts, and (c) a list of words used for indexing the bag-of-word data. For this assignment, we will use only the first zipped file ("Abstracts_Part1.zip), which consists of 51,979 abstracts. You can find the data set and information about the data at http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html.

Instructions

1. Download a copy of the first dataset named "Abstracts_Part1.zip" from the archive.

2. Log in to AWS ALL (Academy Learning Lab) with your account and password. Create a AWS SageMaker Notebook Instance using the following information:

   Instance Name: IS597MLC-SP2024-Machine-Learning-Pipelines-Assignment
   Instance Type: ml.t3.xlarge

   Upload the zipped file with a Jupyter Notebook starter file provided by the instructor to the Notebook Instance you just created for your assignment. Or you may create a GitHub repository if you have an account and use a repo URL when you create an AWS Notebook Instance.

3. You might want to look into the first few files in your dataset to see what they look like. Also, detailed description of data can be found under 'Information files' section at the same site which you download your file from. For reference, here is a screenshot of a sample file.

```
Title       : CAREER: Markov Chain Monte Carlo Methods
Type: Award
NSF Org     : CCR
Latest
Amendment
Date        : May 5, 2003
File        : a0237834

Award Number: 0237834
Award Instr.: Continuing grant
Prgm Manager: Ding-Zhu Du
          CCR  DIV OF COMPUTER-COMMUNICATIONS RESEARCH
          CSE  DIRECT FOR COMPUTER & INFO SCIE & ENGINR
Start Date  : August 1,  2003
Expires     : May 31,  2008 (Estimated)
Expected
Total Amt.  : $400000             (Estimated)
Investigator: Eric Vigoda vigoda@cs.uchicago.edu  (Principal Investigator current)
Sponsor     : University of Chicago
          5801 South Ellis Avenue
          Chicago, IL  606371404    773/702-8602

NSF Program : 2860      THEORY OF COMPUTING
Fld Applictn:
Program Ref : 1045,1187,9216,HPCC,
Abstract    :

     Markov chain Monte Carlo (MCMC) methods are an important algorithmic
     device in a variety of fields.  This project studies techniques for rigorous
     analysis of the convergence properties of Markov chains.   The emphasis is on
     refining probabilistic, analytic and combinatorial tools (such as coupling,
     log-Sobolev, and canonical paths) to improve existing algorithms and develop
     efficient algorithms for important open problems.

     Problems arising in
     computer science, discrete mathematics, and physics are of particular interest,
     e.g., generating random colorings and independent sets of bounded-degree
     graphs, approximating the permanent, estimating the volume of a convex body,
     and sampling contingency tables.  The project also studies inherent connections
     between phase transitions in statistical physics models and convergence
     properties of associated Markov chains.

     The investigator is developing a
     new graduate course on MCMC methods.
```

4. Use the Jupyter Notebook starter file provided by the instructor to complete the exercises, following the instructions described in this notebook.

Submission

Submit your assignment to the UIUC Canvas Assignment section following the instructions in the notebook.