1. NSF Research Awards Abstracts

This dataset comprises several paper abstracts, one per file, that were furnished by the NSF (National Science Foundation). A sample abstract is shown at the end.

Your task is developing an unsupervised model which classifies abstracts into a topic (discover them!). Indeed, your goal is to group abstracts based on their semantic similarity.

You can get a sample of abstracts here. Be creative and state clearly your approach. Although we don't expect accurate results but a good pipeline of your work.
It is affordable to create a notebook, like Jupyter (if you use python) or a Rmarkdown report (in case you use R) and make it available for us, i.e. github.

Hint to success in your quest: Develop and stay clear of the data science process you'll perform over the dataset and highlight important aspects you might consider affordable to discuss over.

Aside notes: All fields in every abstract file wouldn't be needed. Be keen.

Good luck and have fun.

Abstract sample:
==========
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.
==========

2. Purchases

Hi there,

In order to be assessed by the Globant's Data Science team, we encourage you to develop the following test in order to see your Data Scientist skills properly and how well you resolve a common data science task.

In this dataset you have a collection of purchase card transactions for the Birmingham City Council. This is a historical dataset, you're able to perform any of the following tasks:

1. **(Clustering) Discovering profiles (whether the case) or unusual transactions (anomalies detection) ...**
2. **(Forecasting) Try to guess future transactional behaviors. For instance, what would be the next purchase? Expenditures forecasting? ...**
3. **(Creativity) State a problem.**

It's up to you defining the time window in which your analysis will take place.

To do so, we suggest you create a notebook, like Jupyter (if you use python) or a Rmarkdown report (in case you use R) and make it available for us, i.e. github. Make sure your notebook can be properly loaded on a web browser or just send us a PDF.

Hint to success in your quest: Develop and state clear of the data science process you'll perform over the dataset and highlight important aspects you might consider affordable to discuss over. Use the mindset of a business and curious consultant.

You have up to a day before the technical interview to share your results of this test.

Good luck and have fun.

3. Diabetes

In this dataset you have 3 different outputs:

1. **No readmission;**
2. **A readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate);**
3. **A readmission in more than 30 days (this one is not so good as well the last one, however, the reason could be the state of the patient.**

Your task is either to classify a patient-hospital outcome or to cluster them aiming at finding patterns that give a distinct insight.

To do so, we suggest you create a notebook, like Jupyter (if you use python) or a Rmarkdown report (in case you use R) and make it available for us, i.e. github.

Hint to success in your quest: Develop and stay clear of the data science process you'll perform over the dataset and highlight important aspects you might consider affordable to discuss over.

You have up to a day before the technical interview to share your results of this test.

Good luck.