# Instructions

The idea of this challenge is to identify in a better way your capacities to translate data into assets, we expect a good pipeline and solution that you can understand and translate.

Select one of the below problems where it is affordable to create an asset using python (scripts files is preferd) where we can identify whole data process,  not only the modeling part, like  data pipeline/ flow, but you can develop a notebook file. Remember to Make it available over github

Hint:

- Highlight variables or patterns using EDA
- Be sure to clarify your hypothesis
- Validate your functions
- Be clear with the pipeline
- Readme and Requirements files are a plus

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

**Good luck and have fun.**

# 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 we want to identify your knowledge over traditional and newest method over NLP

**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.
==========

# Purchases

**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.**