PaperScraper

MSCI-446 Term Project.

Andrew Andrade,
Andy Toulis, Baha Nurlybayev

Research Question

How has Artificial Intelligence and Predictive Analytics been used in the Petroleum Industry?

Problem

Data mining approach to aid multi-document classification for literature review.

Treating Uncertainties in Reservoir Performance Prediction with Neural Networks

☆☆☆☆☆ Average from 0 ratings

Authors	Johann Peter Lechner (OMV A.G.) | Georg Zangl (Schlumberger)

DOI	http://dx.doi.org/10.2118/94357-MS
Document ID	SPE-94357-MS
Publisher	Society of Petroleum Engineers
Source	SPE Europec/EAGE Annual Conference, 13-16 June, Madrid, Spain
Publication Date	2005
Document Type	Conference Paper
Language	English
ISBN	978-1-55563-943-3

Disciplines	6.7.4 Probabilistic Methods, 6.5.5 Evaluation of Uncertainties

Show less detail

SPE Member Price: USD 8.50

SPE Non-Member Price: USD 25.00

Add to cart

Abstract

In development projects reservoir parameters are only known within certain ranges - a fact that allows various realisations of the subsurface. Because of the computational time involved, not all of the possible parameter combinations can be covered by simulation models to obtain a probability distribution of possible outcomes. Creating a response surface that is based on a reduced number of simulation runs becomes necessary. Such a response surface can be utilized to approximate results for numerous different variations of input parameters. In contrast to the widely used methodology of fitting a polynomial model to the results of a limited number of simulation runs, an approach, where reservoir response is captured by an Artificial Neural Network (ANN) has been investigated.

Data

Features (Paper metadata, in text form):

- Abstracts
- Year
- Keywords
- Disciplines
- other meta data

Classes:

- Disciplines

Data

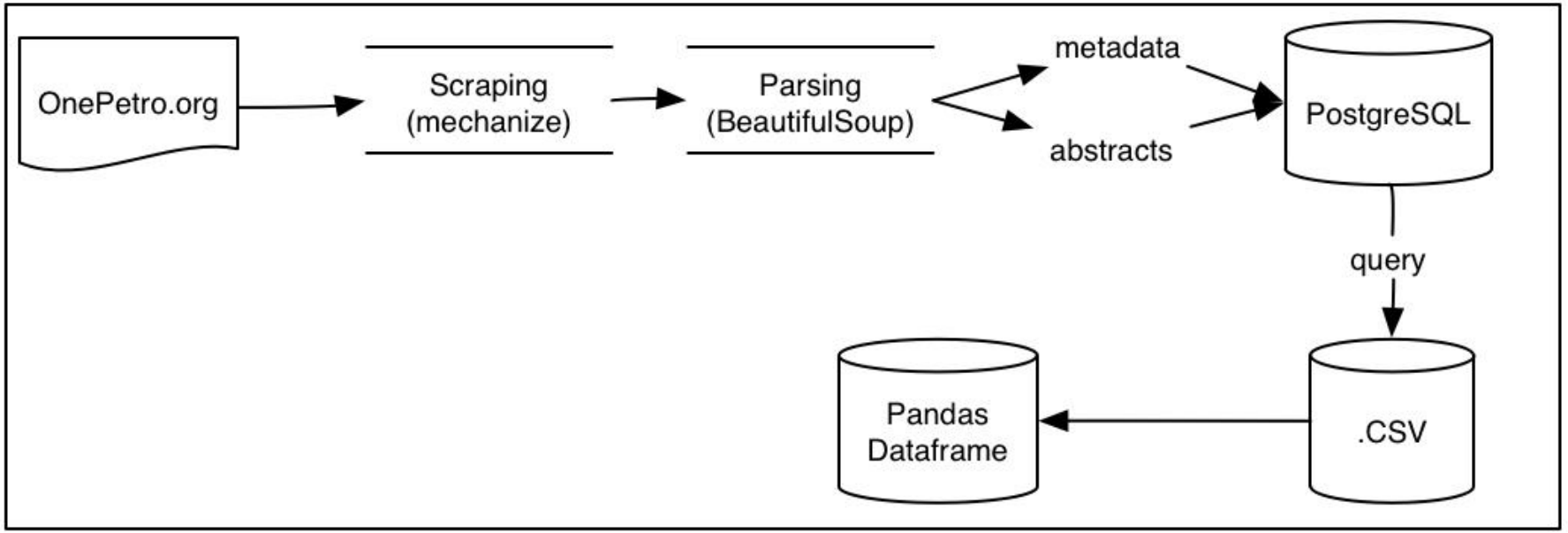
- 180 000+ papers on OnePetro
- 117 000 scraped, parsed*
- 22 250 labelled data
- Real world data is dirty (nightmarish)

*Scraper was originally written by Jonathon “Jay” Estrella as an intern at PetroPredict, later modified to include hidden metadata

Dirty Data

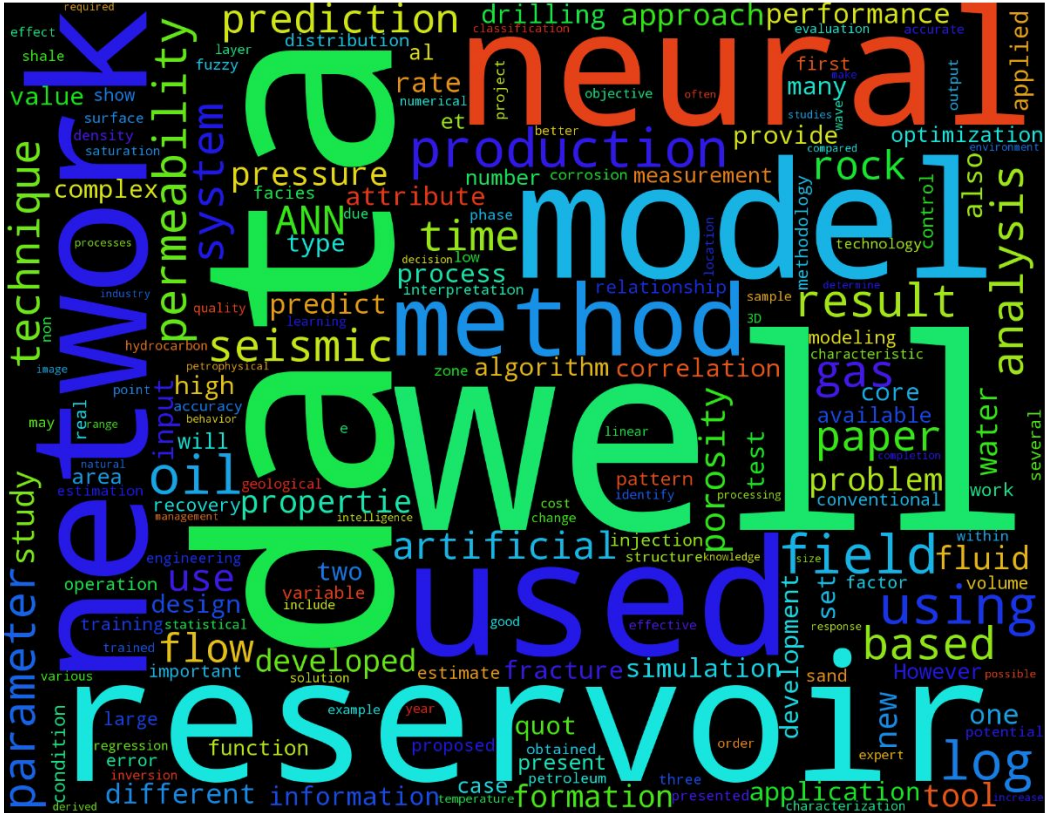
- unicode (special characters)
- Missing values
- non-sense
- mislabels
- inconsistencies

Data Collection



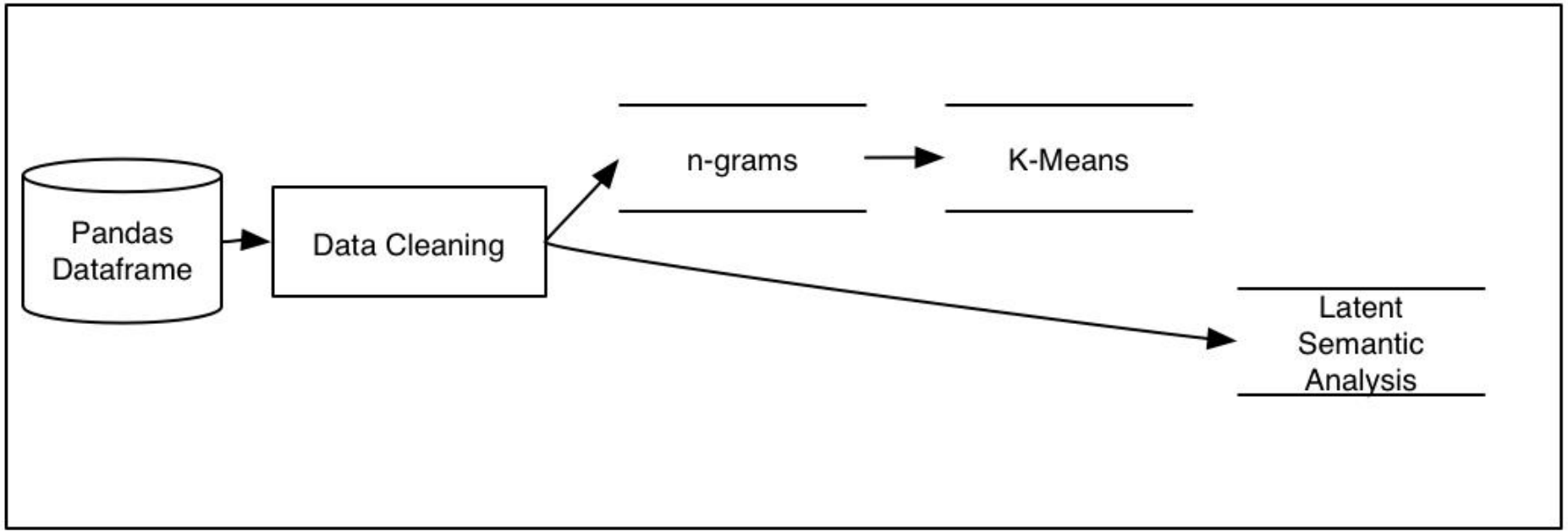
Approach

1. Data Collection and Storage
    a. Scraping
    b. Pandas (because we can)
2. **EDA & Clustering**



Full results: http://mrandrewandrade.com/blog/2015/11/13/paper-scraper-update-1s.html

EDA


Pandas Dataframe
Data Cleaning
n-grams
K-Means
Latent Semantic Analysis

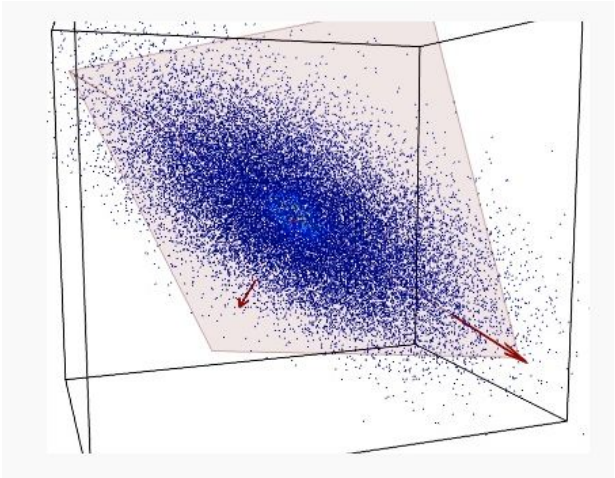
Latent Semantic Analysis

Feature: Abstracts (TF-IDF vector)

Dimensionality Reduction: Principal Component Analysis

Metric: Cosine Similarity

2 to 12 Clusters



http://earlywarn.blogspot.ca/2010/10/extracting-signal-from-drought-noise.html

Example: 12 Clusters

Cluster 1: steam crude polymer viscosity flooding produced sand surfactant

Cluster 2: strength waves structures results ice numerical soil tests structure failure study test loading

Cluster 3: safety spe management industry health risk companies business conference prepared training environmental

Cluster 4: drilling mud bit hole drill drilled wellbore casing rig fluids cuttings

Cluster 5: subsea offshore systems equipment project platform completion cement pipeline installation tubing

Cluster 6 : porosity models simulation properties log core saturation pore

Cluster 7: seismic inversion velocity migration imaging noise image processing acquisition survey source

Cluster 8: natural hydrate liquid condensate energy reserves methane coal

Cluster 9: preview atce commercially otc technologies readily limited engineers houston currently discussion

12 Clusters Cont.

Fracking:

Cluster 10: fracture fracturing fractures hydraulic proppant fractured stimulation treatment treatments

French!**:**

Cluster 11: eacute des les der egrave die agrave und dans pour une par von est ouml que sur sont qui mit

Cluster 12: corrosion steel alloy alloys resistance steels stainless metal coating materials hydrogen cracking

Approach

1. Data Collection and Storage
    a. Scraping
    b. SQL database
    c. Pandas (because we can)
2. Exploratory Data Analysis
    a. Data Visualization
    b. Determine papers using AI techniques
    c. K-means clustering Topic modeling
3. **Document Classification**

MultiLabel Classification (6 Disciplines of SPE)

Class 0: Drilling and Completions

Class 1: Health, Safety, Security, Environment and Social Responsibility

Class 2: Management and Information

Class 3: Project Facilities and Construction

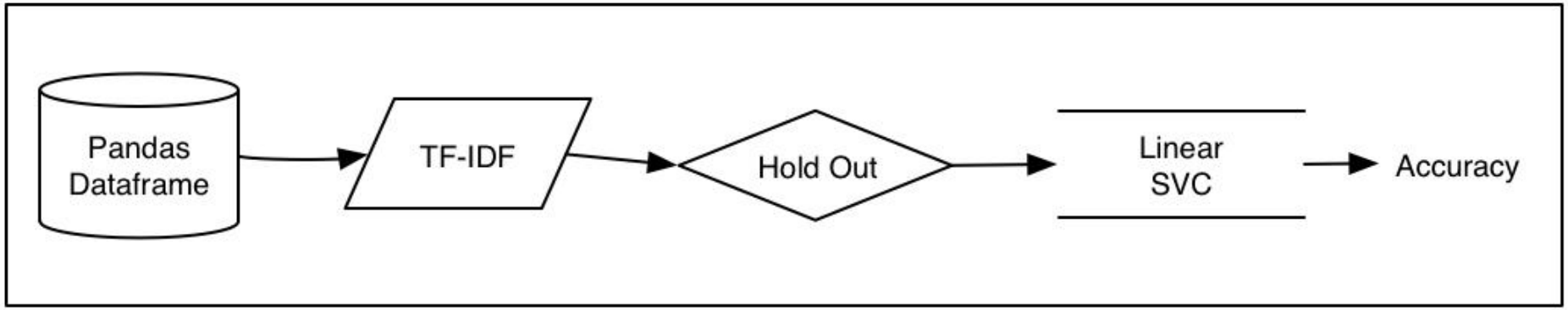
Class 4: Production and Operations

Class 5: Reservoir Description and Dynamics

Approach

1. Data Collection and Storage
    a. Scraping
    b. SQL database
    c. Pandas (because we can)
2. Exploratory Data Analysis
    a. Data Visualization
    b. Determine papers using AI techniques
    c. K-means clustering
3. Multi Label Document Classification
    a. Linear SVM

Data Classification 1



Very High Accuracy!

71% - 86%

```
Class0accuracy:
0.813251201923
Class1accuracy:
0.842247596154
Class2accuracy:
0.861628605769
Class3accuracy:
0.815504807692
Class4accuracy:
0.714242788462
Class5accuracy:
0.792518028846
```

Management and Information

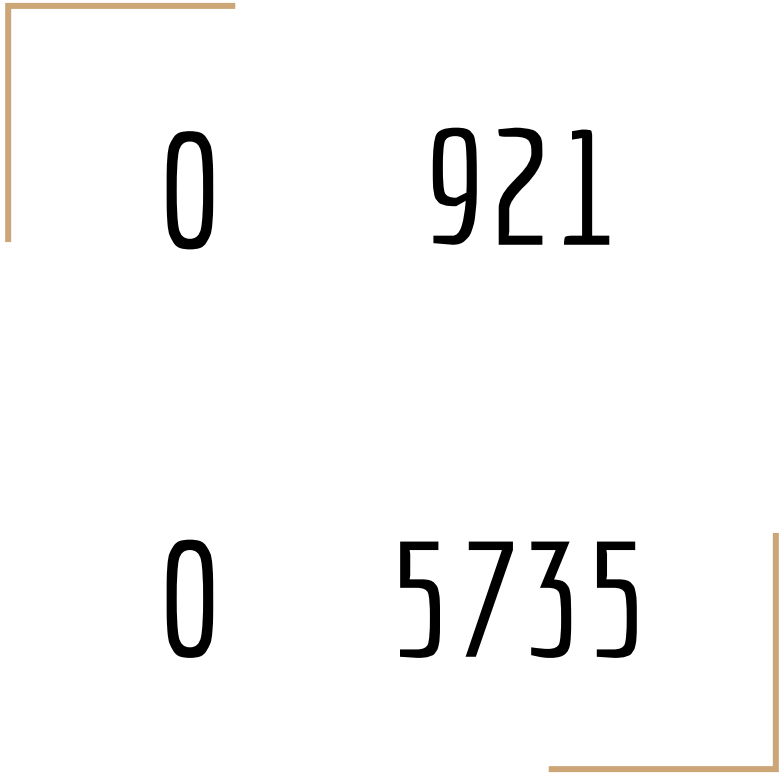
CM

$$\begin{bmatrix} 0 & 921 \\ 0 & 5735 \end{bmatrix}$$

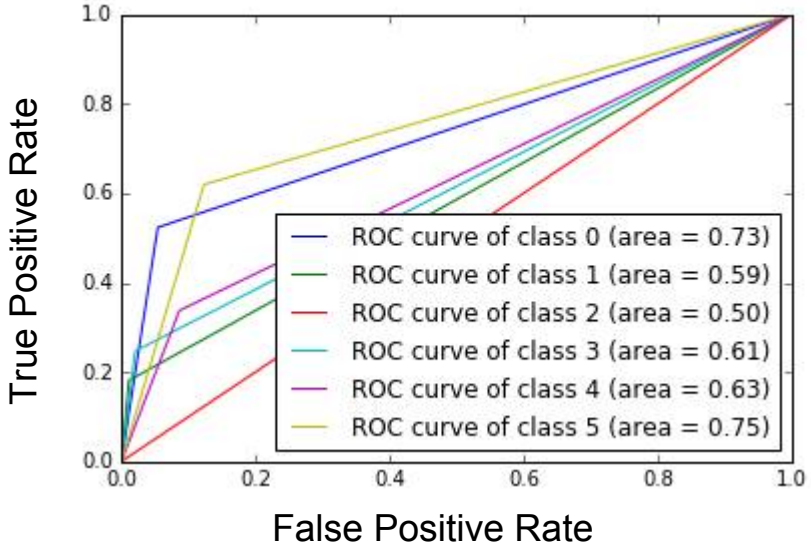
86% Accuracy!

Only 14% are about Management

Labeling everything as not Management!



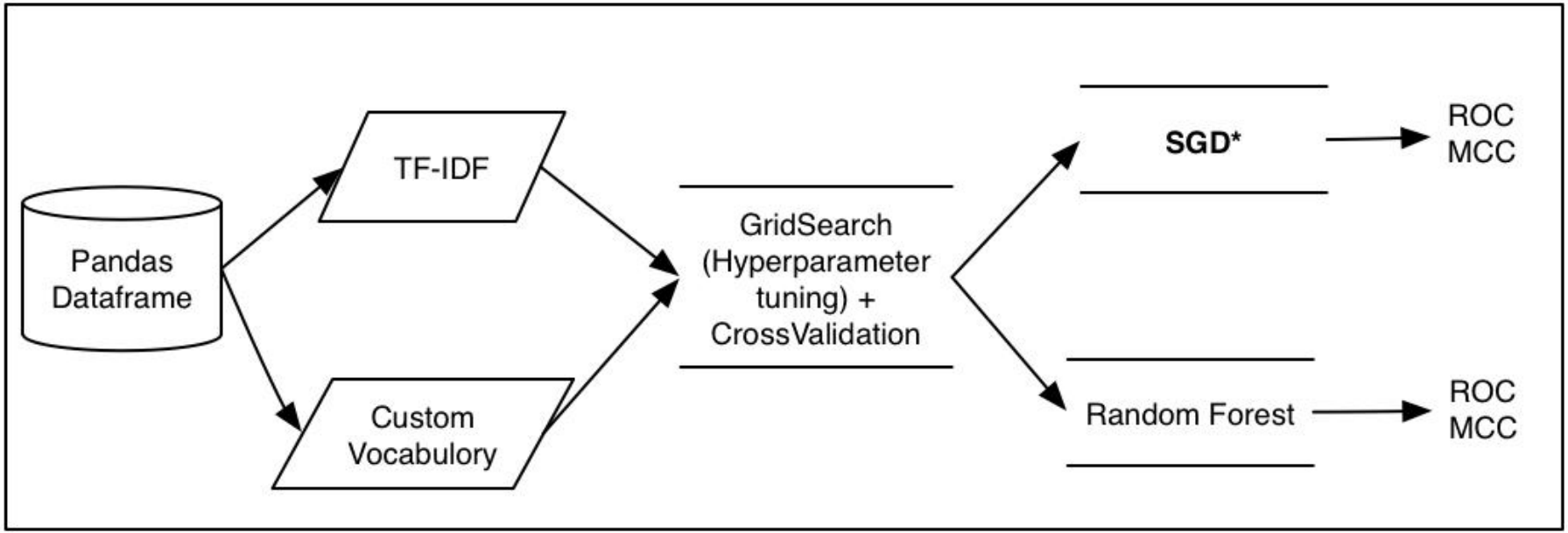
Receiver Operating Characteristic and AUC



Approach

1. Data Collection and Storage
    a. Scraping
    b. SQL database
    c. Pandas (because we can)
2. Exploratory Data Analysis
    a. Data Visualization
    b. Determine papers using AI techniques
    c. K-means clustering
3. Multi Label Document Classification
    a. Linear Support Vector Machine
    b. Grid search + SGD

Data Classification 2

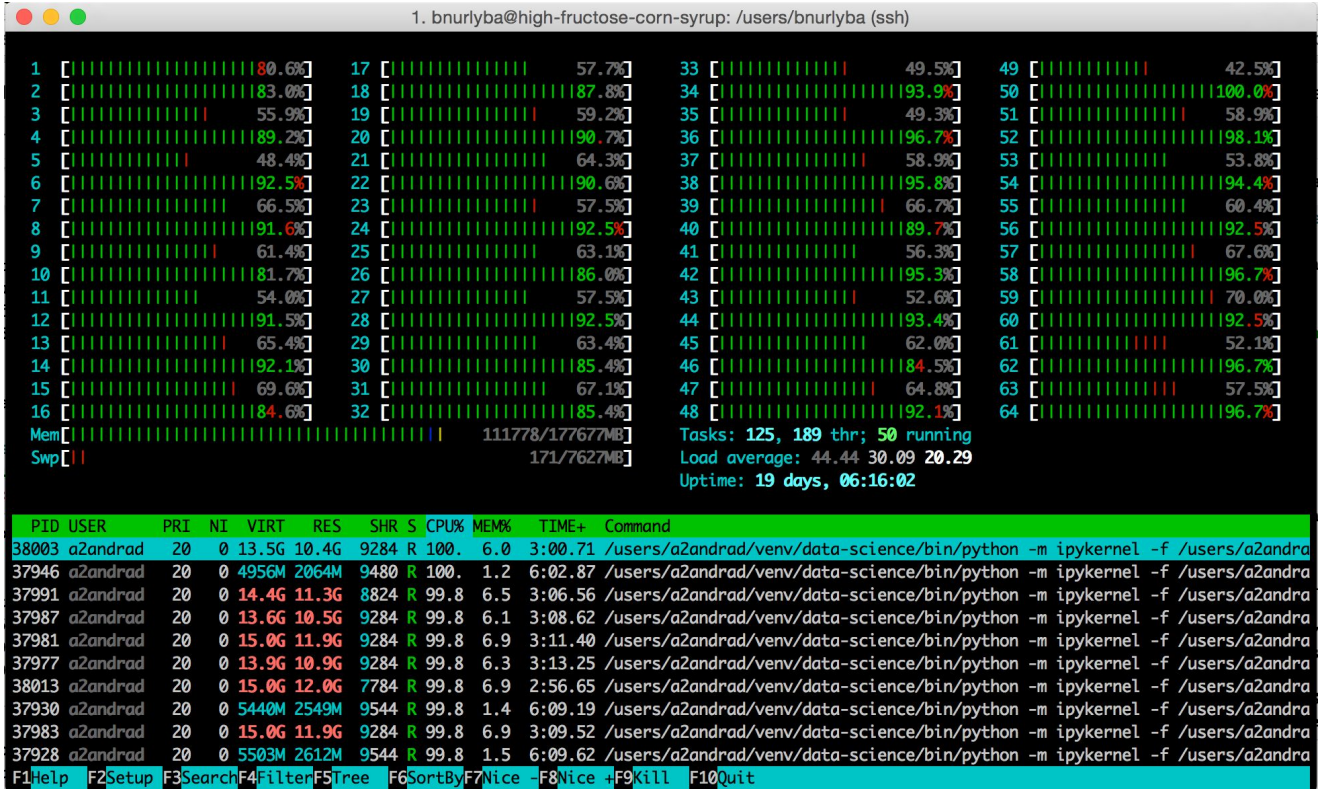


Pipelines and Hyperparameters

```
pipeline = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf', SGDClassifier(loss='log', n_iter=10, alpha=0.0001)),
])

parameters = {
    'vect__min_df': (0.1, 0.2),
    'vect__max_df': (0.8, 0.9),
    #'vect__max_features': (None, 5000, 10000, 50000),
    'vect__ngram_range': ((1, 1), (1, 2), (1,3)),
    #'tfidf__use_idf': (True, False),
    #'tfidf__norm': ('l1', 'l2'),
    'clf__penalty': ('l2', 'elasticnet')
}
```

GridSearch


1. bnurlyba@high-fructose-corn-syrup: /users/bnurlyba (ssh)
Mem[111778/177677MB]
Swp[171/7627MB]
Tasks: 125, 189 thr; 50 running
Load average: 44.44 30.09 20.29
Uptime: 19 days, 06:16:02
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
38003 a2andrad 20 0 13.5G 10.4G 9284 R 100. 6.0 3:00.71 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37946 a2andrad 20 0 4956M 2064M 9480 R 100. 1.2 6:02.87 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37991 a2andrad 20 0 14.4G 11.3G 8824 R 99.8 6.5 3:06.56 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37987 a2andrad 20 0 13.6G 10.5G 9284 R 99.8 6.1 3:08.62 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37981 a2andrad 20 0 15.0G 11.9G 9284 R 99.8 6.9 3:11.40 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37977 a2andrad 20 0 13.9G 10.9G 9284 R 99.8 6.3 3:13.25 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
38013 a2andrad 20 0 15.0G 12.0G 7784 R 99.8 6.9 2:56.65 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37930 a2andrad 20 0 5440M 2549M 9544 R 99.8 1.4 6:09.19 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37983 a2andrad 20 0 15.0G 11.9G 9284 R 99.8 6.9 3:09.52 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
37928 a2andrad 20 0 5503M 2612M 9544 R 99.8 1.5 6:09.62 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andra
F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9Kill F10Quit

GridSearch


data scientist's
THE #1 PROGRAMMER EXCUSE
FOR LEGITIMATELY SLACKING OFF:
model's
training
"MY CODE'S COMPILING."
HEY! GET BACK TO WORK!
training
COMPILING!
OH. CARRY ON.

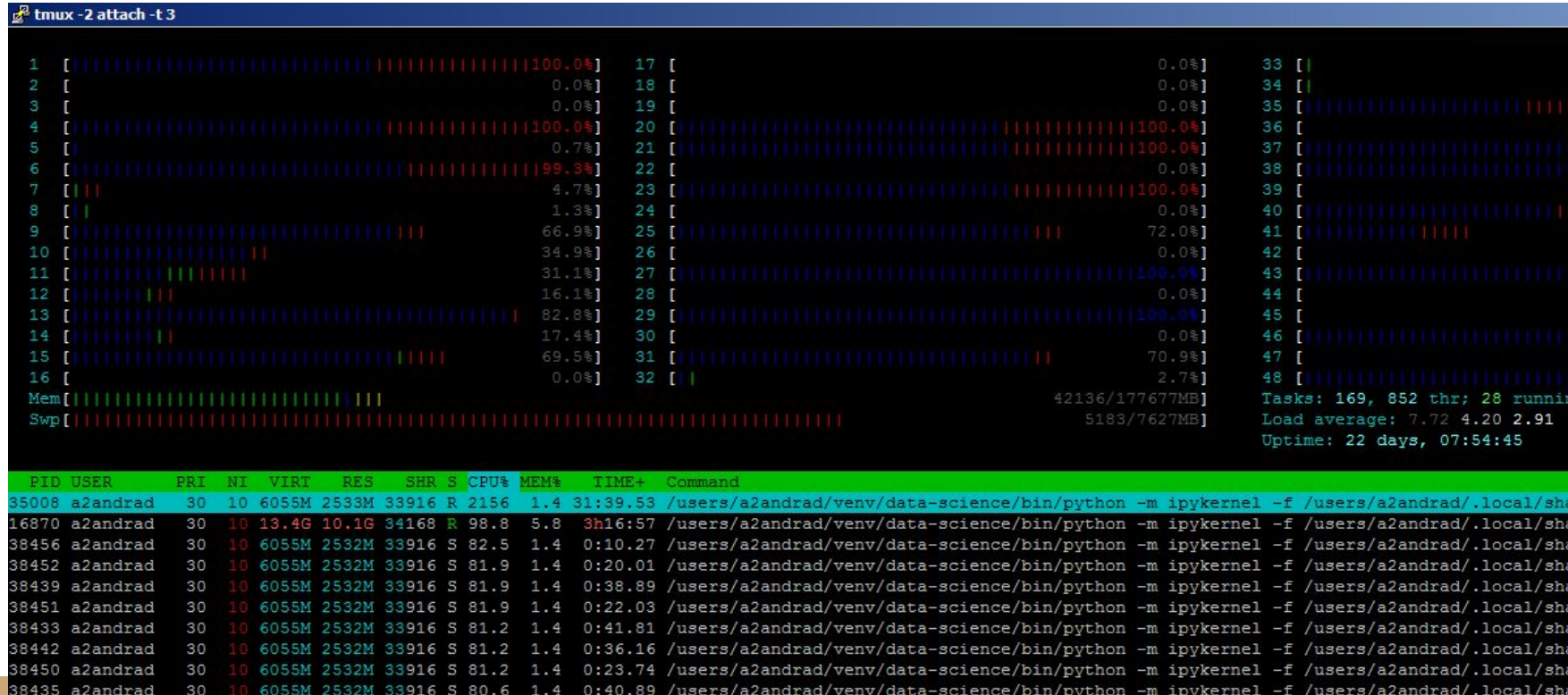
GridSearch

Hello Andrew,

This morning (Wednesday, November 30) your python processes running on HFCS consumed all available RAM and CPU. Your processes have been killed by the Systems Committee as they were in violation of our Machine Usage Agreement (http://csclub.uwaterloo.ca/services/machine_usage), specifically the section titled "User Responsibilities".

If you restart the processes, please ensure that they do not consume all available resources on the machine.

GridSearch


tmux -2 attach -t 3
Mem[42136/177677MB]
Swp[5183/7627MB]
Tasks: 169, 852 thr; 28 runni
Load average: 7.72 4.20 2.91
Uptime: 22 days, 07:54:45
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
35008 a2andrad 30 10 6055M 2533M 33916 R 2156 1.4 31:39.53 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
16870 a2andrad 30 10 13.4G 10.1G 34168 R 98.8 5.8 3h16:57 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38456 a2andrad 30 10 6055M 2532M 33916 S 82.5 1.4 0:10.27 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38452 a2andrad 30 10 6055M 2532M 33916 S 81.9 1.4 0:20.01 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38439 a2andrad 30 10 6055M 2532M 33916 S 81.9 1.4 0:38.89 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38451 a2andrad 30 10 6055M 2532M 33916 S 81.9 1.4 0:22.03 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38433 a2andrad 30 10 6055M 2532M 33916 S 81.2 1.4 0:41.81 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38442 a2andrad 30 10 6055M 2532M 33916 S 81.2 1.4 0:36.16 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh
38450 a2andrad 30 10 6055M 2532M 33916 S 81.2 1.4 0:23.74 /users/a2andrad/venv/data-science/bin/python -m ipykernel -f /users/a2andrad/.local/sh

Approach

1. Data Collection and Storage
    a. Scraping
    b. SQL database
    c. Pandas (because we can)
2. Exploratory Data Analysis
    a. Data Visualization
    b. Determine papers using AI techniques
    c. K-means clustering
3. Multi Label Document Classification
    a. Linear Support Vector Machine
    b. Grid search + SGD+ Dimensionality Reduction

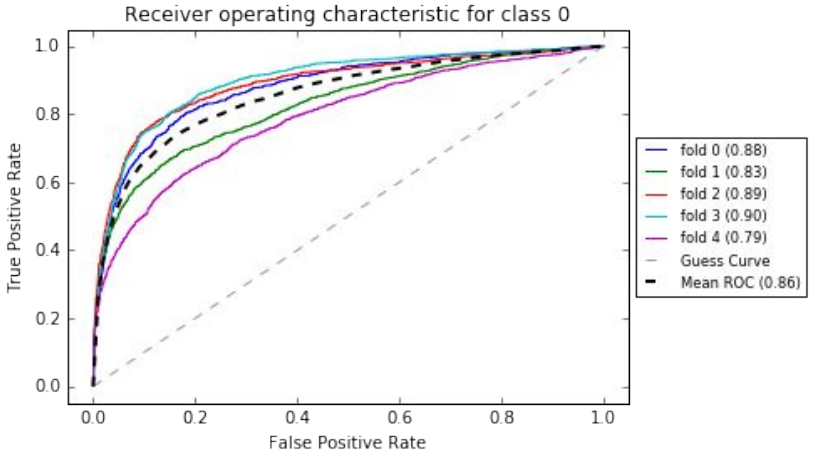
Reduce Dimensionality

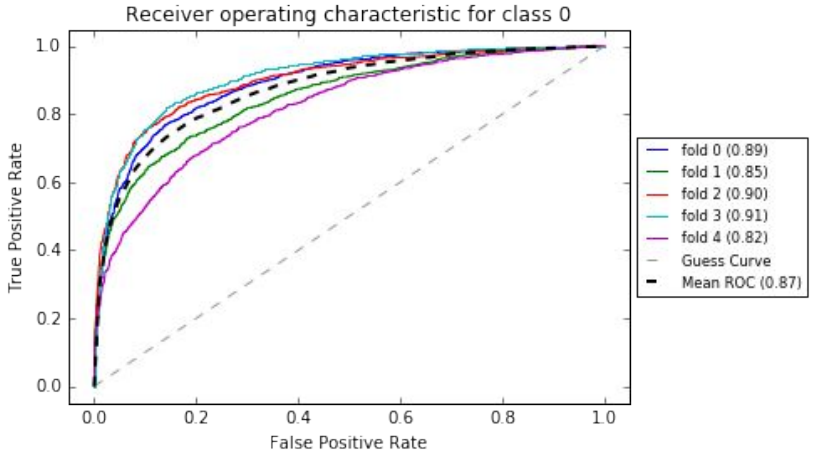
Disciplines have text!

Build dict from the discipline text

Use document term vector from discipline

Full Abstract vs. Custom Dictionary

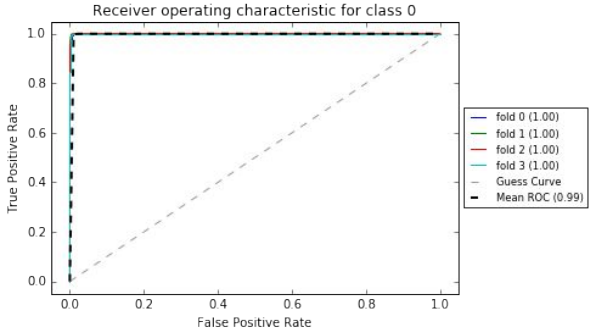


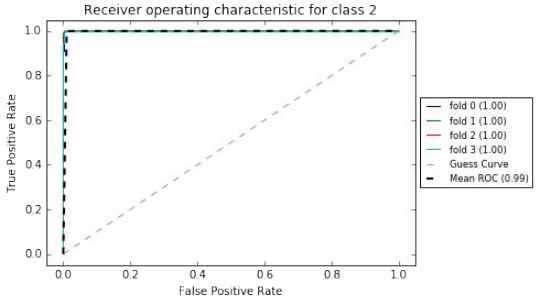


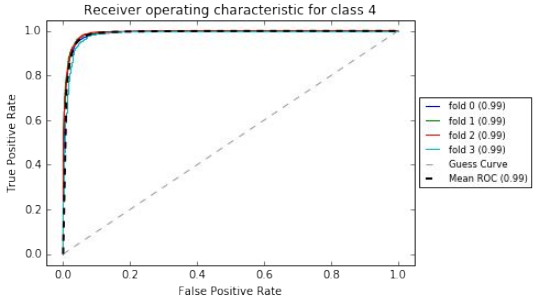
Approach

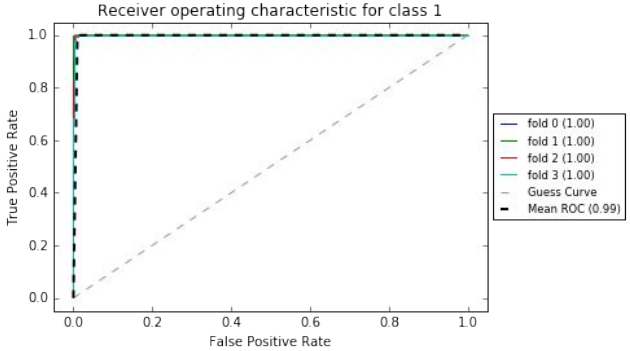
1. Data Collection and Storage
    a. Scraping
    b. SQL database
    c. Pandas (because we can)
2. Exploratory Data Analysis
    a. Data Visualization
    b. Determine papers using AI techniques
    c. K-means clustering
3. Multi Label Document Classification
    a. Linear Support Vector Machine
    b. Grid search + SGD + Dimensionality Reduction
    c. **Random Forest**

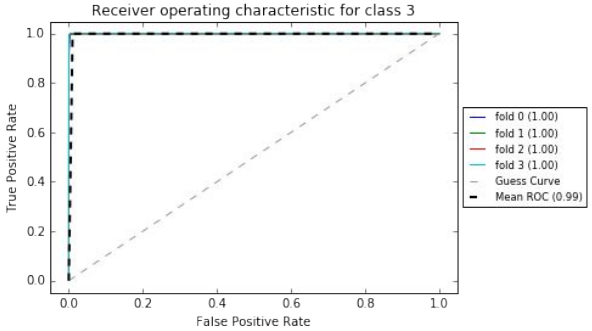
Random Forest! (severely overfitting)

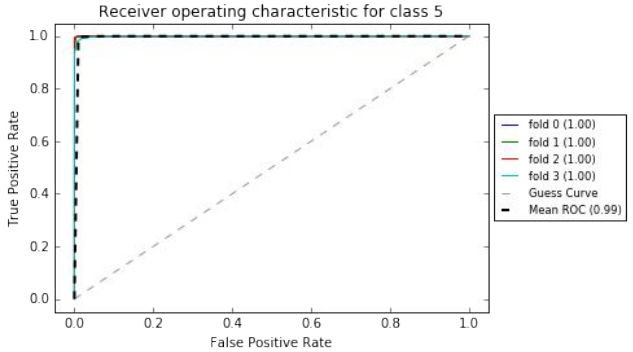




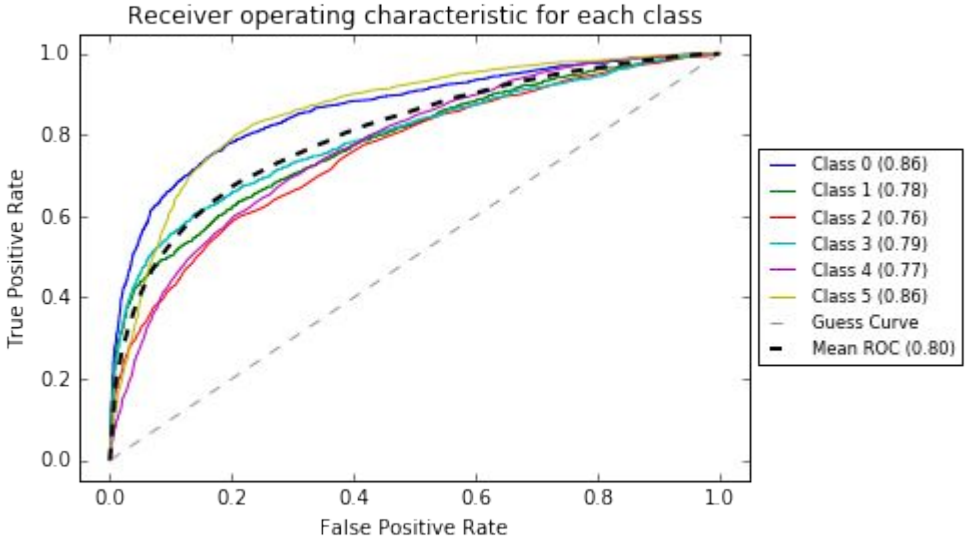








30% hold out (model on 70% k folds)



Top 10 features: Random Forest

Drilling and Completions
wellbor
drill
string
bit
reservoir
cement
complet
format
mud
design

Health, Safety, Security, Environment and Social Responsibility
health
manag
environment
assess
emiss
hse
inject
model
train
impact

Management and Information
develop
integr
decis
oper
project
pressur
engin
time
safeti
drill

Top 10 features: Random Forest

Project Facilities and Construction
offshor
instal
facil
riser
design
moor
reservoir
float
use
format

Production and Operations
pump
drill
product
hydraul
fractur
reservoir
water
enhanc
use
flow

Reservoir Description and Dynamics
simul
interpret
reservoir
recoveri
log
inject
miscibl
model
drill

Approach

1. Data Collection and Storage
   a. Scraping
   b. SQL database
   c. Pandas (because we can)
2. Exploratory Data Analysis
   a. Data Visualization
   b. Determine papers using AI techniques
   c. K-means clustering
3. Multi Label Document Classification
   a. Linear Support Vector Machine
   b. Grid search + SGD + Dimensionality Reduction
   c. Random Forest
   d. Final SGD Model

Final SGD Model

```
final_pipeline = Pipeline([
    ('vect', CountVectorizer(ngram_range=(1,3), max_df = 0.4, min_df=0001)),
    ('tfidf', TfidfTransformer()),
    ('clf', SGDClassifier(loss='modified_huber', n_iter=10, penalty='elasticnet')),
])
```

Final SGD Model

MCC 70% --> 30% holdout

Class 0: 0.60 --> 0.**62**

Class 1: 0.46 --> 0.46

Class 2: 0.33 --> 0.**36**

Class 3: 0.50 --> 0.**51**

Class 4: 0.41 --> 0.40

Class 5: 0.58 --> 0.56

Predictions

89170 predictions

Class 0: 13140 (15%)

Class 1: 4230 (5%)

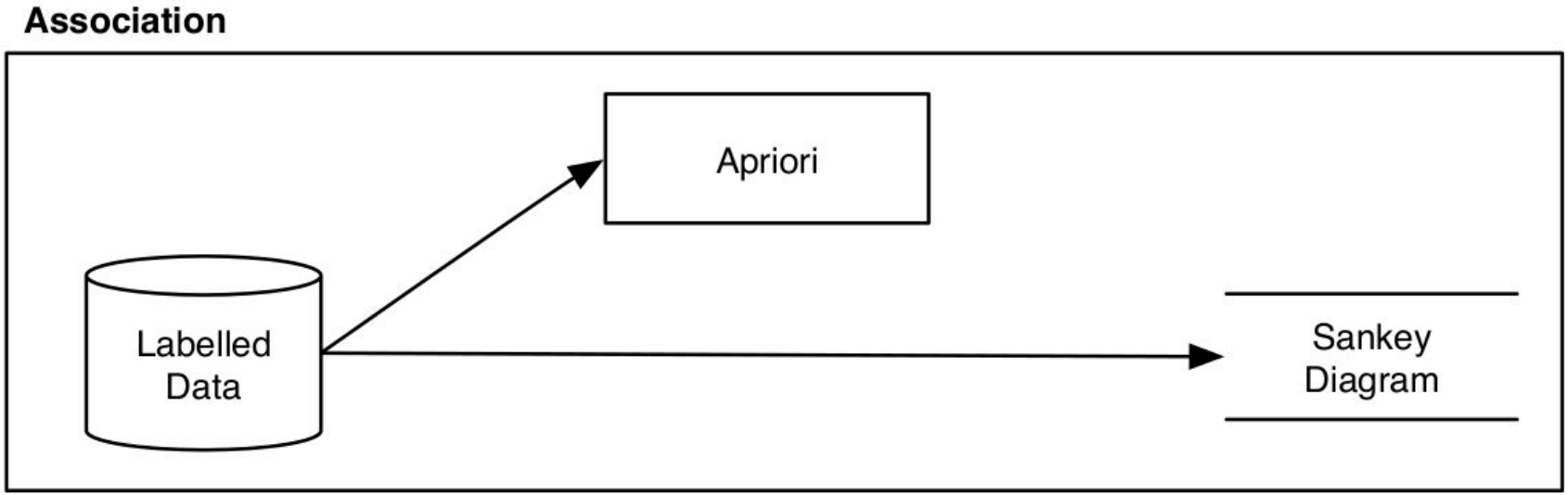
Class 2: 3050 (3%)

Class 3: 15200 (17%)

Class 4: 15300 (17%)

Class 5: 18560 (21%)

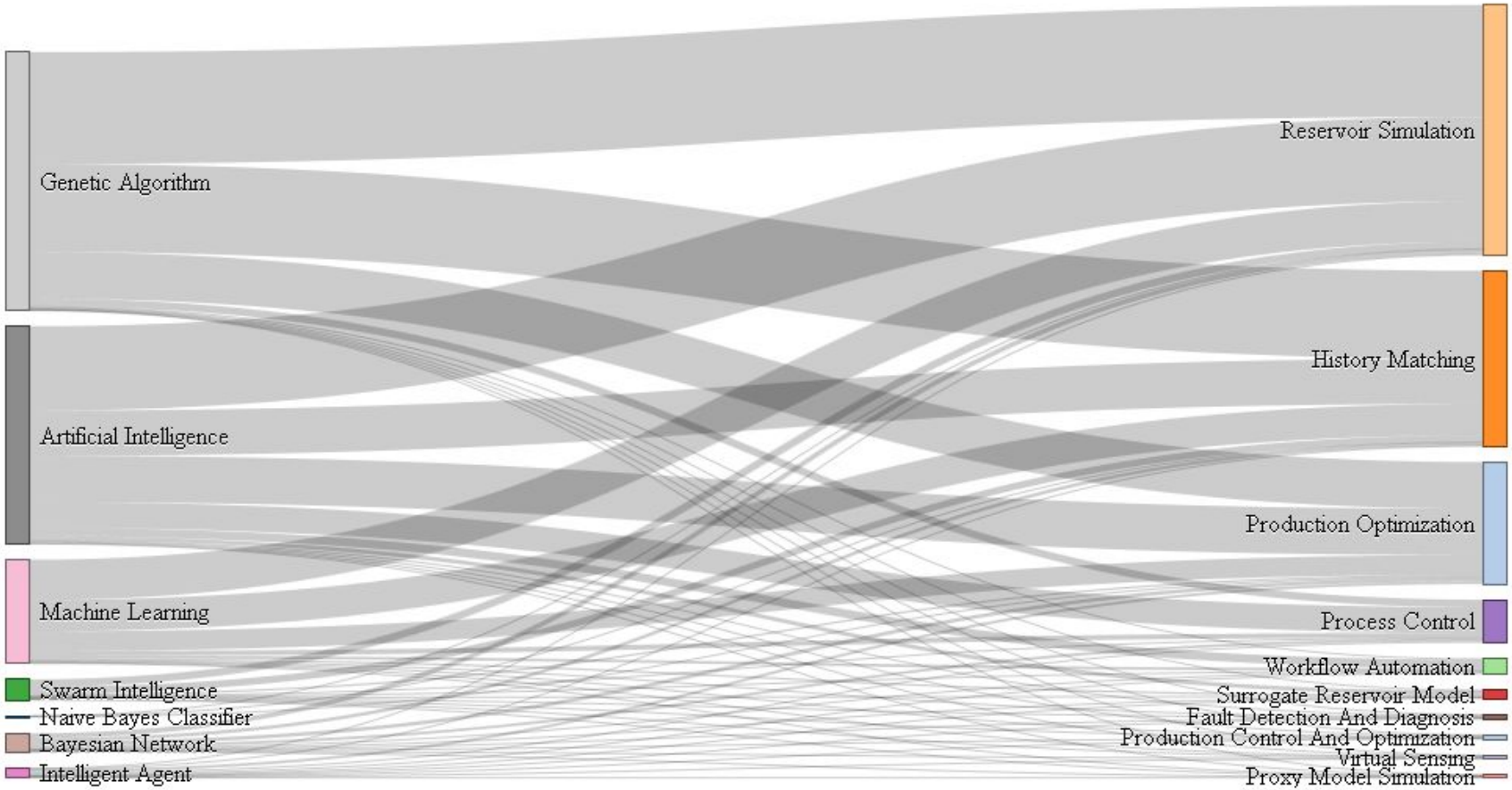
Association



Apriori

rules	support	confidence	lift
{} => {class_0_prediction=1}	0.1190476190	0.119047619	1.00000000
{regress=1} => {class_0_prediction=1}	0.0430597771	0.144680851	1.21531915
{neural_network=1} => {class_0_prediction=1}	0.0215298886	0.122478386	1.02881844
{mont_carlo=1} => {class_0_prediction=1}	0.0197568389	0.100905563	0.84760673
{regress_analysis=1} => {class_0_prediction=1}	0.0141843972	0.200000000	1.68000000
{regress_analysis=1,regress=1} => {class_0_prediction=1}	0.0141843972	0.200000000	1.68000000
{data_manag=1} => {class_0_prediction=1}	0.0134245187	0.187279152	1.57314488
{expert_system=1} => {class_0_prediction=1}	0.0116514691	0.211981567	1.78064516
{genet=1} => {class_0_prediction=1}	0.0078520770	0.064049587	0.53801653
{bayesian=1} => {class_0_prediction=1}	0.0058257345	0.076411960	0.64186047


artifici_intellig;expert_system;neural_network;mont_carlo(1)
artifici_intellig;expert_system;bayesian(1)
genet;neural_network;mont_carlo(1)
artifici_intellig;data_manag(1)
expert_system;data_manag(1)
artifici_intellig;expert_system;neural_network(1)
artifici_intellig;expert_system(1)
data_manag(1)
expert_system;bayesian(1)
expert_system;genet;neural_network(1)
regress;gaussian;mont_carlo(1)
artifici_intellig;data_manag;neural_network(1)
artifici_intellig;expert_system;genet;neural_network(1)
expert_system(1)
artifici_intellig;bayesian(1)
artifici_intellig;regress;data_manag;genet;neural_network(1)
expert_system;neural_network(1)
artifici_intellig;neural_network(1)
bayesian;genet;mont_carlo(1)
artifici_intellig;regress(1)
bayesian;neural_network(1)
artifici_intellig;genet;neural_network(1)
artifici_intellig;genet;mont_carlo(1)
genet;neural_network(1)
mont_carlo(1)
artifici_intellig;expert_system;regress(1)
data_manag;neural_network(1)
expert_system;genet(1)
expert_system;regress;neural_network(1)
regress;genet(1)
regress;mont_carlo(1)
regress(1)
bayesian;mont_carlo(1)
neural_network(1)
bayesian(1)
regress;bayesian;neural_network(1)
neural_network;mont_carlo(1)
genet(1)
regress;bayesian(1)
regress;neural_network(1)
regress;gaussian(1)
regress;genet;neural_network(1)
genet;mont_carlo(1)
gaussian(1)
artifici_intellig;genet(1)
expert_system;regress(1)
bayesian;gaussian(1)
gaussian;mont_carlo(1)
genet;gaussian(1)
gaussian;neural_network(1)
artifici_intellig;regress;neural_network(1)
bayesian;gaussian;mont_carlo(1)
bayesian;genet(1)
bayesian;genet;gaussian(1)
bayesian;neural_network;mont_carlo(1)
data_manag;gaussian(1)
expert_system;bayesian;mont_carlo(1)
gaussian;neural_network;mont_carlo(1)
genet;gaussian;mont_carlo(1)
regress;bayesian;mont_carlo(1)
regress;bayesian;neural_network;mont_carlo(1)
regress;data_manag(1)
regress;neural_network;mont_carlo(1)
regress;gaussian;neural_network(1)
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Production and Operations;Reservoir Descripti
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Productio
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities
Drilling and Completions;Management and Information;Management and Information;Reservoir Descripti
Drilling and Completions;Management and Information;Management and Information;Project Facilities and Construction;Productio
Management and Information;Managemen
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Managemen
Drilling and Completions;Management and Information;Managemen
Management and Information;Management and Information;Productio
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Project Facilities and Construction;Reservoir Descripti
Management and Information;Management and Information;Project Facilities
Management and Information;Management and Information;Production and Operations;Reservoir Descripti
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Management and Information;Management and Information;Project Facilities and Construction;Productio
Drilling and Completions;Productio
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Production and Operations;Reservoir Descripti
Health, Safety, Security, Environment and So
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities and Construction;Productio
Drilling
Drilling and Completions;Management and Information;Management and Information;Productio
Management and Information;Management and Information;Reservoir Descripti
Drilling and Completions;Management and Information;Management and Information;Project Facilities
Drilling and Completions;Health, Safety, Security, Environment and So
Drilling and Completions;Reservoir Descripti
Productio
Project Facilities and Construction;Productio
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Project Facilities
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Health, Safety, Security, Environment and Social Responsibility;Project Facilities and Construction;Productio
Drilling and Completions;Management and Information;Management and Information;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities and Construction;Productio
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Project Facilities and Construction;Reservoir Descripti
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Project Facilities and Construction;Productio
Production and Operations;Reservoir Descripti
Drilling and Completions;Project Facilities
Drilling and Completions;Project Facilities and Construction;Productio
Drilling and Completions;Production and Operations;Reservoir Descripti
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Production and Operations;Reservoir Descripti
Health, Safety, Security, Environment and Social Responsibility;Productio
Drilling and Completions;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Health, Safety, Security, Environment and Social Responsibility;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Productio
Management and Information;Management and Information;Project Facilities and Construction;Production and Operations;Reservoir Descripti
Drilling and Completions;Management and Information;Management and Information;Production and Operations;Reservoir Descripti
Project Facilities
Reservoir Descripti
Project Facilities and Construction;Production and Operations;Reservoir Descripti
Drilling and Completions;Health, Safety, Security, Environment and Social Responsibility;Management and Information;Management and Information;Productio


Genetic Algorithm
Artificial Intelligence
Machine Learning
Swarm Intelligence
Naive Bayes Classifier
Bayesian Network
Intelligent Agent
Reservoir Simulation
History Matching
Production Optimization
Process Control
Workflow Automation
Surrogate Reservoir Model
Fault Detection And Diagnosis
Production Control And Optimization
Virtual Sensing
Proxy Model Simulation

Thanks!