Syracuse University

Master of Science in Applied Data Science (ADS)

Portfolio Milestone - Project

Venkata Sharat Sripada
SUID# 436036419
vssripad@syr.edu

Overview

I used to work at VMWare, Inc until recently, and a talk by Prof. Michael Jordan from Berkeley at one of VMware's R&D conferences in the Spring of 2019 inspired me to pursue the field of data-science and machine-learning. At the time I was applying machine-learning to niche problems in my role as a scale and performance Engineering Lead in the computer networking/virtualization technology, viz:

- sifting through hundreds of unit-tests, randomizing, and sequencing them in ways that would cover code paths eventually making software components resilient
- training a Random-forest model on traffic patterns in a datacenter and aiding software-defined networks (SDNs) to enforce dropping anonymous traffic patterns
- using NLP and Apriori algorithms in large scale log aggregation endpoints to correlate events and auto-heal based on a remediation table lookup

However, in quest to gain deeper knowledge I had applied in the Fall of 2019 to the program of Applied Data-Science at Syracuse. More recently I transitioned to a new role at Amazon LLC as a System Development Engineer working on retail *search engine [1]*, where I intend to not just use knowledge garnered through this course but eventually cut into areas of research moving the needle of innovation, science, and Engineering.

Learning Goals and Objectives

While largely the goal was to explore where academia was meeting the industry, there was keen interest about the curriculum that led to the following goals:

- dealing with data-sets – how to collect or obtain data in the real-world, munge or cleanse it and stage it so models can be deployed and consumed in production environments
- learning to make predictions using the right machine-learning algorithms
- guidance by Professors at regular cadence with sync-sessions, evaluation of assignments and projects
- understand how data analysis or interpretation can be communicated back to business or stake holders via meaningful and succinct visualization and messaging

Course Structure

The course was split into two parts:

- Asynchronous module
  - Recorded videos presented by the Dean or Professor teaching the course
  - Questions pertaining to topics presented
  - Simulated lab exercises
  - Homework and Quizzes
- Synchronous module
  - 90-min live session delivered by the Professor
  - Breakout discussions
  - Online Exams

Program catalog - Outline

The Applied Data Science program comprised 11 courses (33 credits) in all and was organized as:

- Primary Core - 6 courses (18 credits)
- Secondary Core in *Language Analytics Track [2]* - 2 courses (6 credits)
- Elective Courses - 3 courses (9 credits)
- Exit requirement (1 credit)

Below are details of all completed courses (ranked chronologically):

Course	Description	Track	Professor	Term	Credits
MBC-638 - Data Anls & Decisn Making	Introduction to statistics to perform analysis and interpret results in a meaningful way. Understand value of data collection and analysis in acquiring knowledge and making decisions in today's business environment	Elective	Luz Flores Lee	Fall 2019	3
IST-687 - Introduction to Data Science	Explore key concepts related to data-science, statistics, information visualization and text mining using R	Primary Core	G. Krudys	Spring 2020	3
IST-659 - Data Admin Concepts & Db Mgmt	Data administration and concepts and skills – data analysis techniques, data modelling and schema design	Primary Core	Chad Harper	Spring 2020	3

Program catalog - Outline

Course	Description	Track	Professor	Term	Credits
IST-707 - Data Analytics	Focus on machine-learning model building and optimization, real-world applications	Primary Core	Jeremy Bolton	Summer 2020	3
IST-652 - Scripting for Data Analysis	Skills of scripting (using Python) to solve problems of accessing and preparing data in a variety of formats – establishes skills essential to form data science pipelines	Elective	D. Landowski	Winter 2020	3
IST-718 - Big Data Analytics	Insights into logistic regression, decision trees and neural networks to make predictions – build Apache spark/python to build big data analytics pipelines	Primary Core	Jon Fox	Spring 2021	3
IST-664 - Natural Language Processing	Linguistic and computational aspect of natural language processing – used NLTK libraries in Python to solve various problems related to real-world applications of NLP	Secondary Core	Norma Polamino	Summer 2021	3
IST-736 - Text Mining	Advanced text mining algorithms for information extraction, text classification and clustering, opinion mining, and their applications in real-world problems	Secondary Core	Jeremy Bolton	Fall 2021	3
IST-719 - Information Visualization	Information visualization through the R programming language and Adobe illustrator - Data cleaning techniques, control of the R graphics environment, develop custom plots, visually explore data, use design concepts to visually communicate the story in the data, and discuss issues related to the ethics of data visualization	Elective	G. Krudys	Spring 2022	3

Program catalog - Outline

Course	Description	Track	Professor	Term	Credits
IST-772 - Quant Reasoning Data Science	Multiple strategies for inferential reasoning about quantitative data and methods for connecting data provenance to substantive analytical conclusions	Primary Core	Jason Anastasopoulos	Spring 2022	3
SCM-651 - Business Analytics	Developing a portfolio of skills related to: Data collection – use tools like Google Analytics to collect/ organize data Data analysis: identify patterns in the data via visualization, statistical analysis, and data mining Strategy and decisions: develop alternative strategies based on the data Implementation: develop a plan of action to implement the business decisions	Primary Core	Don Harter	Summer 2022	3

Projects and key learnings from coursework

Title - Server and network inventory mgmt at scale

Goal

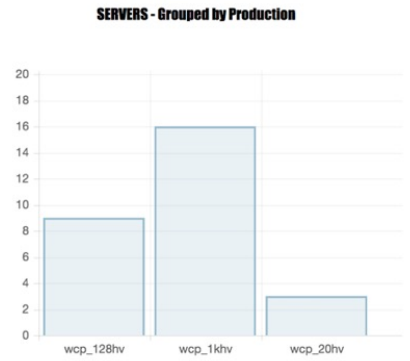
Build a dashboard to view/manage server and network inventory in an enterprise datacenter

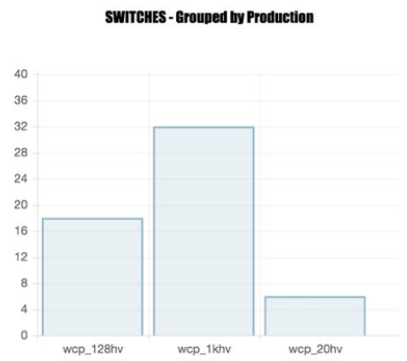
Source Code

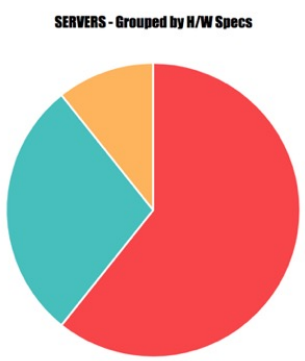
Python-Flask for webserver (SQL DB) and HTML/JavaScript for frontend/GUI:
https://github.com/sharatsv/MS-DataScience/tree/main/IST-659/Final-Project

Key-highlights

- Conceptualization to realization using techniques imbibed via IST-659 in designing databases
  - All tables organically went from ERD through logical-modeling and normalization process before being 3NF compliant
- All tables had real data from production environments which greatly established a workflow that would eventually roll out to manage larger datacenters
- A fully functional dashboard (below) showing resources consumed across environments







Title: Google image retrieval using Landmarks data (Kaggle competition)

Goal

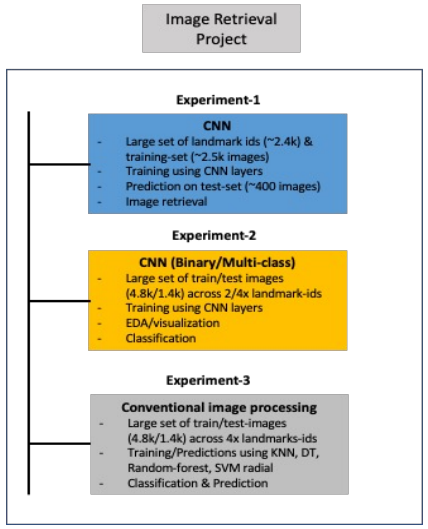
To retrieve corresponding landmarks given a test image

Source Code

In R using the Keras/Tensorflow libraries: https://github.com/sharatsv/MS-DataScience/tree/main/IST-707/Final-Project

Key-highlights

- For image classification, convolutional neural networks give much higher accuracy in making predictions than traditional classification algorithms, as shown in the table below
- CNNs can even work well with very small sample sets
- When dealing with large image data sets, like Google Landmark data, CNNs need much computing power to carry out their “learning”
- Keras library with TensorFlow backend provides an easy-to-use framework that allows fast prototyping
- Hyperparameter-tuning of CNNs can get very complex and time-consuming



Accuracy Comparison for All Models

Model	Accuracy (4 labels)	Accuracy (2 labels)
Decision Tree	53.71%	70%
Random Forest	63.54%	83.70%
KNN	56.73% (k=5)	80.34% (k=17)
SVM Radial	61.43%	78.86%
CNN	93.1%	97.7%

Title: Human protein Atlas Single cell classifier (Kaggle competition)

Goal

To develop models capable of classifying mixed patterns of proteins in microscopic images

Source Code

In Python leveraging Apriori algorithms and Tensorflow for deep learning

- https://github.com/sharatsv/MS-DataScience/tree/main/IST-718/Final-Project
- https://github.com/srihari-busam/hpa-deep-learning

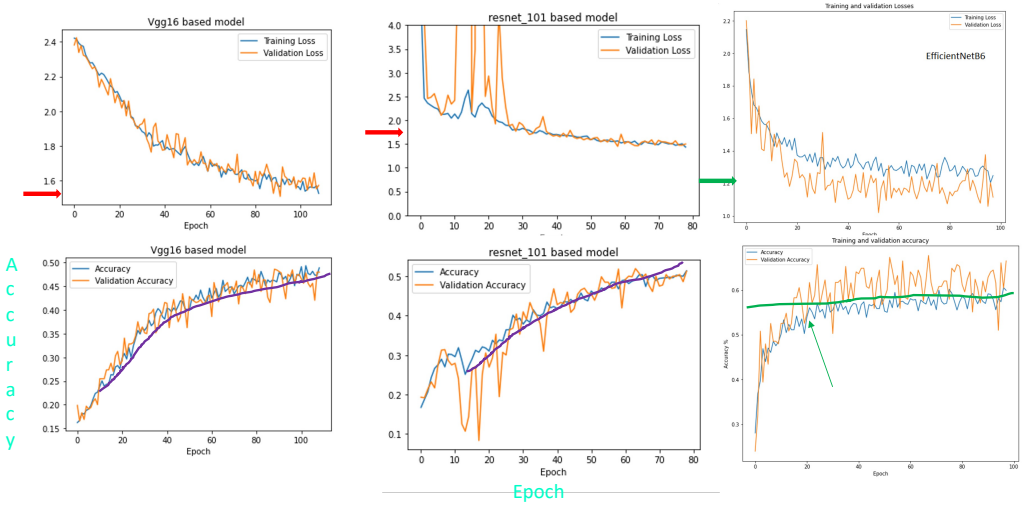
Key-highlights

At the outset, the leaderboard on the Kaggle competition was at **~44%** accuracy and beating this number became our immediate goal. We reduced the scope of the problem from 19 to 12 class detection but achieved ***~67%*** *accuracy*

ConvNet models - experiments & results

Model based on	Input Details	Trainable params	Optimizer	Training Time	Training Accuracy	Validation Accuracy
VGG16 16 layers Weights: None Weights locked : NO	Res:512 X 512 Batchsize: 32 Steps per epoch:100 Validation steps: 20	14,714,688	Adam (lr=0.0001)	~5hrs	**46.59%**	**48.59%**
Resnet101 101 layers Weights: imagenet Weights locked : NO	Res:512 X 512 Batchsize: 24 Steps per epoch:120 Validation steps: 30	48,844,300	NAdam (lr=0.0001)	~5hrs	**49.78%**	**51.94%**
EfficientNetB6 Weights: imagenet Weights locked : NO	Res:512 X 512 Batchsize: 7 Steps per epoch:150 Validation steps: 10	40,763,364	Nadam (lr=0.0001) (tried Ada, rmsprop,SGD)	~8hrs	**70.08%**	**67.14%**

Loss & Accuracy



Title: Datacenter Security Analytics/Recommendation

Goal

Gather flow data in the form of logs (un-structured data) across the distributed plane and build an analytics/recommendation system based on machine-learning

Source Code

In Python using matplotlib for visualization and sklearn's DecisionTreeClassifier for recommendation

- https://github.com/sharatsv/MS-DataScience/tree/main/IST-652/Final-Project

Key-highlights

Extracted a deep view of traffic patterns across a large datacenter – actions on flows at large, sources/protocols that seemed problematic

Used Decision-Tree (sklearn libraries) to train, test and eventually make recommendations about enforcing to permit/drop traffic given a packet flow tuple:

- Data-split: 80% train, 20% test (also used K-cross validation)
- Result: Accuracy of 78%

SOURCE OF DATA

```
2017-10-19T22:38:05.586Z 58734 INET match PASS domain-c8/1006 OUT 84 ICMP 172.18.8.121->172.18.8.119 RULE_TAG
2017-10-19T22:38:08.723Z 58734 INET match PASS domain-c8/1006 OUT 60 TCP 172.18.8.121/36485->172.18.8.119/22 S RULE_TAG
2017-10-19T22:38:18.785Z 58734 INET TERM domain-c8/1006 OUT ICMP 8 0 172.18.8.121->172.18.8.119 2/2 168/168 RULE_TAG
2017-10-19T22:38:20.789Z 58734 INET TERM domain-c8/1006 OUT TCP FIN 172.18.8.121/36484->172.18.8.119/22 44/33 4965/5009 RULE_TAG
```

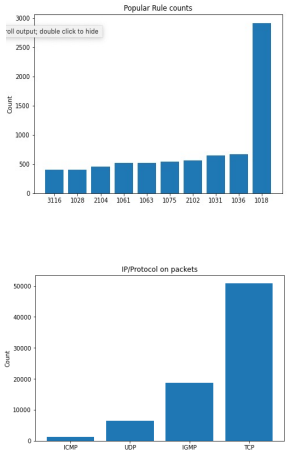
Log excerpt on distributed-plane (raw un-structured data)

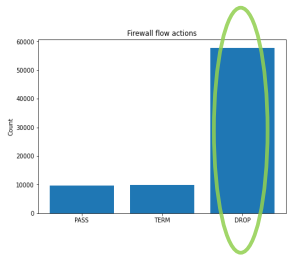
	time	reason	source	action	rule	dir	pktlen	proto	sip	dip	sport	dport
0	2020-12-05T20:51:02.220Z	match	20.20.177.78	DROP	2026	IN	36	IGMP	0.0.0.0	224.0.0.1	0	0
1	2020-12-05T20:51:02.220Z	match	20.20.177.78	DROP	2026	IN	36	IGMP	0.0.0.0	224.0.0.1	0	0
2	2020-12-05T20:51:02.220Z	match	20.20.177.78	DROP	2026	IN	76	ICMP	fe80::ffff:ffff:ffff:ffff	ff02::1	0	0
3	2020-12-05T20:51:02.220Z	match	20.20.177.78	DROP	2026	IN	36	IGMP	0.0.0.0	224.0.0.1	0	0
4	2020-12-05T20:51:02.220Z	match	20.20.177.78	DROP	2026	IN	76	ICMP	fe80::ffff:ffff:ffff:ffff	ff02::1	0	0

Data loaded from Mongo-DB into data-frame

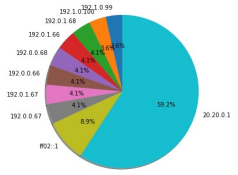
- The source of data will largely be logs written by software components enforcing network policies or rules. It captures traffic flows/tuples hitting an action (allow/deny).
- A traffic flow or tuple would comprise the following fields:
  - src ip-address
  - dst ip-address
  - src port
  - dst port
  - TCP/IP Protocol
  - Action – Deny/Permit

ANALYTICS-DASHBOARD

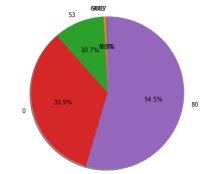




Zoom in on DROP flows



Drop flows categorized by source



Drop flows categorized by L4 ports

Title: Marvel Search Engine

Goal

Given a search query (monologue, dialogue or character), search a large corpus of text/documents comprising scripts from Marvel Universe and retrieve information like Marvel movie, which character said it etc.

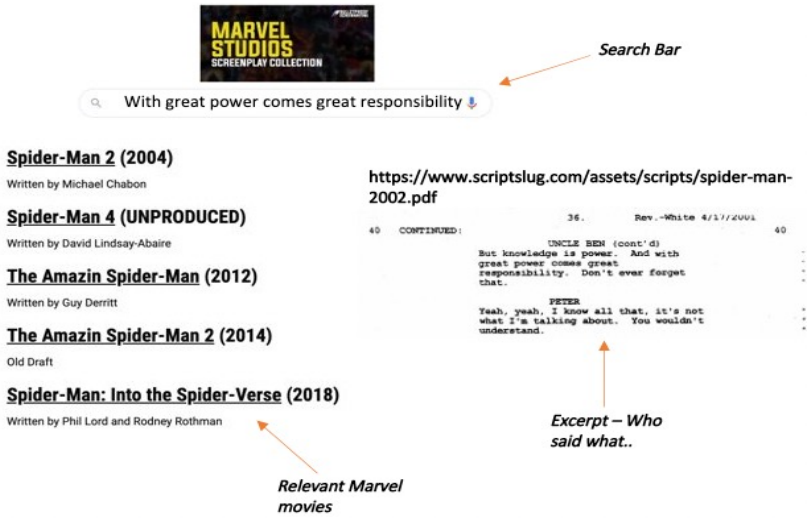
Source Code

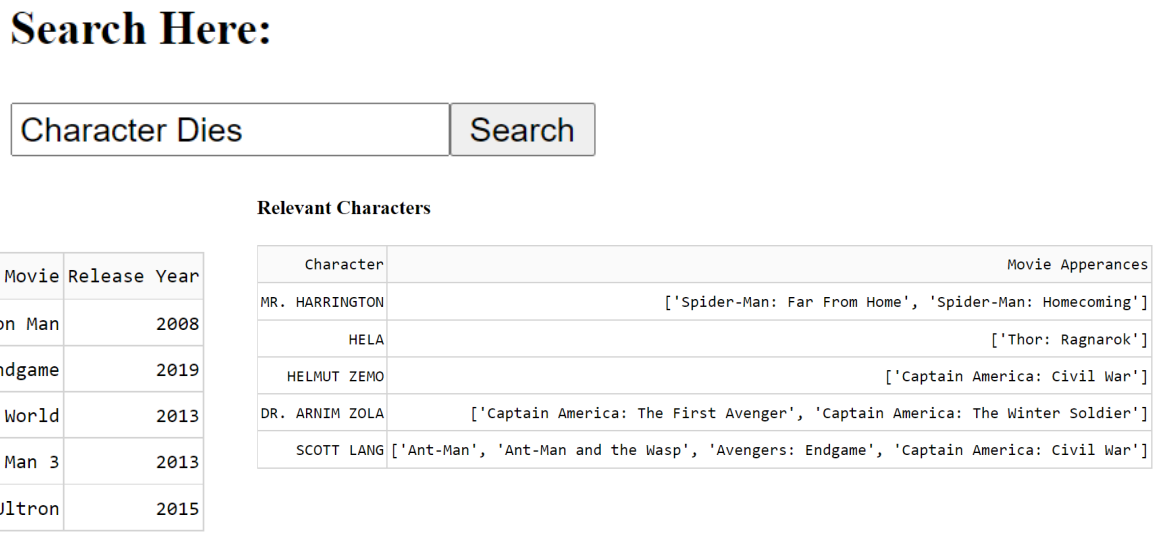
In Python using libraries related to BM25/Google Universal Sentence Encoder (USE)

- https://github.com/sharatsv/MS-DataScience/tree/main/IST-736/Final-Project
- https://github.com/ramosem97/mcu_marvel_search_engine

Key-highlights

- Built a search engine like experience using BM25 and Google USE retrieving most relevant information from a large corpus of text documents at the backend
- Used Python-Flask framework to build a simple frontend/GUI





Movie	Release Year
Iron Man	2008
Avengers: Endgame	2019
Thor: The Dark World	2013
Iron Man 3	2013
Avengers: Age of Ultron	2015

Character	Movie Apperances
MR. HARRINGTON	['Spider-Man: Far From Home', 'Spider-Man: Homecoming']
HELA	['Thor: Ragnarok']
HELMUT ZEMO	['Captain America: Civil War']
DR. ARNIM ZOLA	['Captain America: The First Avenger', 'Captain America: The Winter Soldier']
SCOTT LANG	['Ant-Man', 'Ant-Man and the Wasp', 'Avengers: Endgame', 'Captain America: Civil War']

References

Textbooks and Readings

- **Discovering Statistics** by Daniel T. Larose - 3rd edition
- **Understanding Variation** - The Key to Managing Chaos, 2nd edition By Donald J. Wheeler; SPC Press
- Hoffer, J. A, Ramesh, V., & Topi, H. (2016). **Modern database management** (12th ed.). New
- York, NY: Pearson.
- **Introduction to Data Science** (2017), by Jeffrey S. Saltz & Jeffrey M. Stanton.
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) **Introduction to Data Mining**.
- Tom Mitchell (1997) **Machine Learning**
- Brett Lantz (2015) **Machine Learning with R** (second edition).
- Stanton (2017), **Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R**
- Bird, S., Klein, E., & Loper, E. **Natural language processing** with Python
- Jurafsky, D., & Martin, J. H. **Speech, and language processing** (3rd ed. draft).
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) **Introduction to Data Mining**
- Yau, N. (2011). Visualize this: **The Flowing Data guide to design, visualization, and statistics**. Wiley Publishing.
- Yau, N. (2013). **Data points: Visualization that means something**. Wiley Publishing.
- **Python for Everybody**: https://www.py4e.com/book (Python version)
- Miller, Thomas W., **Modeling Techniques in Predictive Analytics with Python and R**, Pearson, 2015.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, **Deep Learning (DL)**, MIT Press, 2016
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, **An Introduction to Statistical Learning with Applications** in R, Springer, 2013
- [Marvel-Dialog-NLP] https://github.com/prestondunton/marvel-dialogue-nlp/tree/master/data
- [MCU-Script-PDF] https://bulletproofscreenwriting.tv/marvel-studios-screenplays-download/
- [1] Search Engine drives retail site amazon.com where given a query, the backend sifts through billions of products and ranks by relevance items customers are likely to buy
- [2] Aligning to a competency related to language analytics and had courses - Natural Language processing (NLP) and Text Mining

Source Repository

https://github.com/sharatsv/MS-DataScience

Resume

https://www.linkedin.com/in/sharat-s-6a14385/

Thank you!