HPCC SYSTEMS®

HPCC Systems Intern Program

Lorraine Chapman
Manager Business Analyst
HPCC Systems Intern Program Manager
LexisNexis Risk Solutions

What is HPCC Systems?

- End to end big data analytics solution – but it doesn't have to be
- First created in 2000, went open source in 2011 – hpccsystems.com
- Underpins a number of our own data driven business initiatives
- Used in a wide variety of ways – IoT, academic research, business
- Thor Cluster – Data Refinery
- ROXIE Cluster – Data Delivery
- Queries coded in ECL (Enterprise Control Language) training available
- Contribute to the platform / machine learning library /use case
- Cloud native platform - https://wiki.hpccsystems.com/x/FYD0B

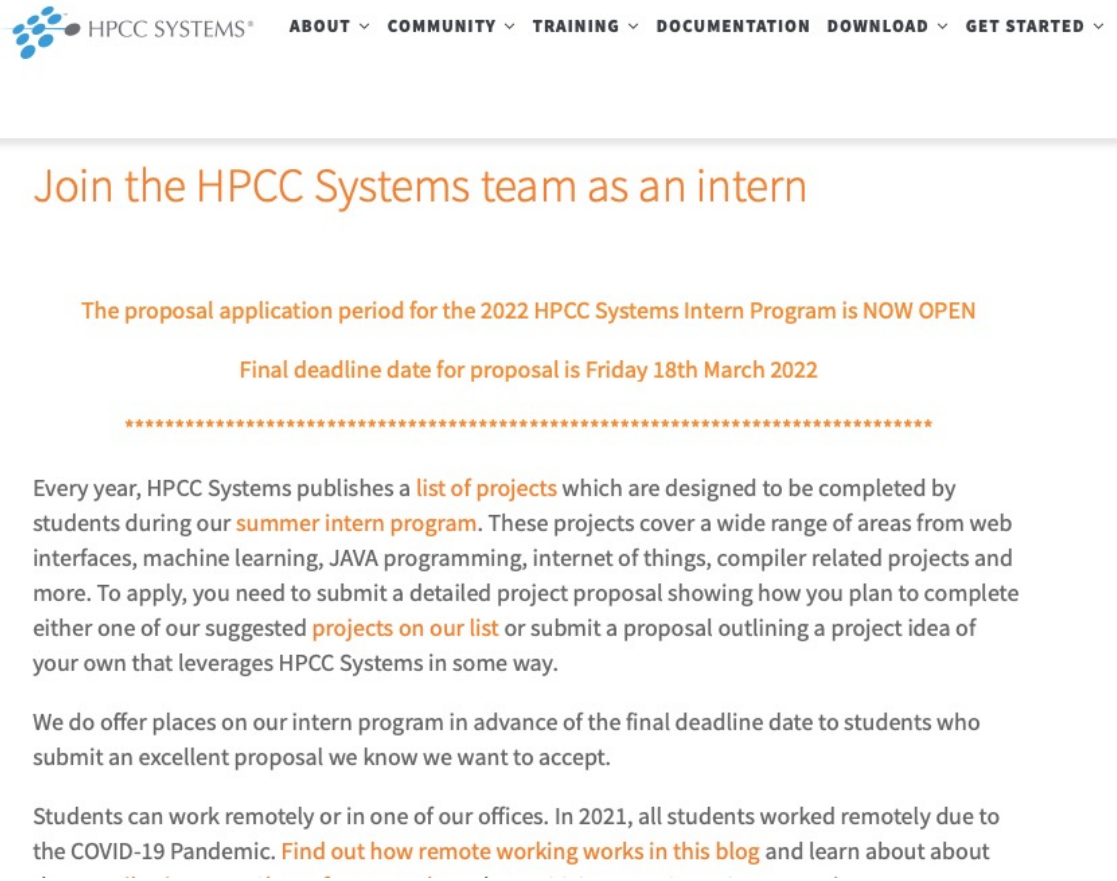
HPCC SYSTEMS®

HPCC Systems Intern Program

- 12 weeks paid program
- High school through to PhD
- Global – Asia, US and Europe
- May through August
- Remote or LN office based (Alpharetta, GA or Boca Raton, FL)
- Proposal period deadline: 18th March 2022
- Review panel decides
- Mentoring
- Community involvement

Read my blog

hpccsystems.com/intern



Timeline - NCF Perspective

2022 internships

- 12 weeks
- Deadline - Friday 18th March
- Looks for projects now: hpccsystems.com/ideas-list
- Suggest your own using HPCC Systems

2023 Internships as practicum

- 16 weeks
- Apply December 2022 for an early start (Jan/Feb)

HPCC SYSTEMS®

Class of 2021 Projects - Use Cases

- Ingestion/Analysis of collegiate women's basketball GPS data using HPCC Systems and RealBI
- COVID-19 Tracker and Global Map Improvements
- Processing Robotics Data with and HPCC Systems Cluster on Kubernetes
- Improvements to the HPCC Systems Structured Query Language (HSQL)

Class of 2021 Projects – Cloud Related

- Using Azure Spot Instances
- Ingress Configuration
- Apply Docker Image Build and Kubernetes Security Principles

HPCC SYSTEMS®

Class of 2021 Projects – Machine Learning

- Implement a PMML Processor
- Toxicity Detection Machine Learning Project
- Causality – Probabilities and Conditional Probabilities
- Causality– Independence, Conditional Independence and Directionality
- Causality – Counterfactual and Interventional Layers
- Causality 2021 Project Details
  https://hpccsystems.com/blog/causality-2021
- Learn more: https://hpccsystems.com/blog/intern-intro-2021

Project Proposal

- Find out more:
  hpccsystems.com/student-wiki
- Available Projects List
  hpccsystems.com/ideas-list
- Highlight the main tasks
- Timeline of work for each week
- Challenges and solutions
- Liaise with the project mentor

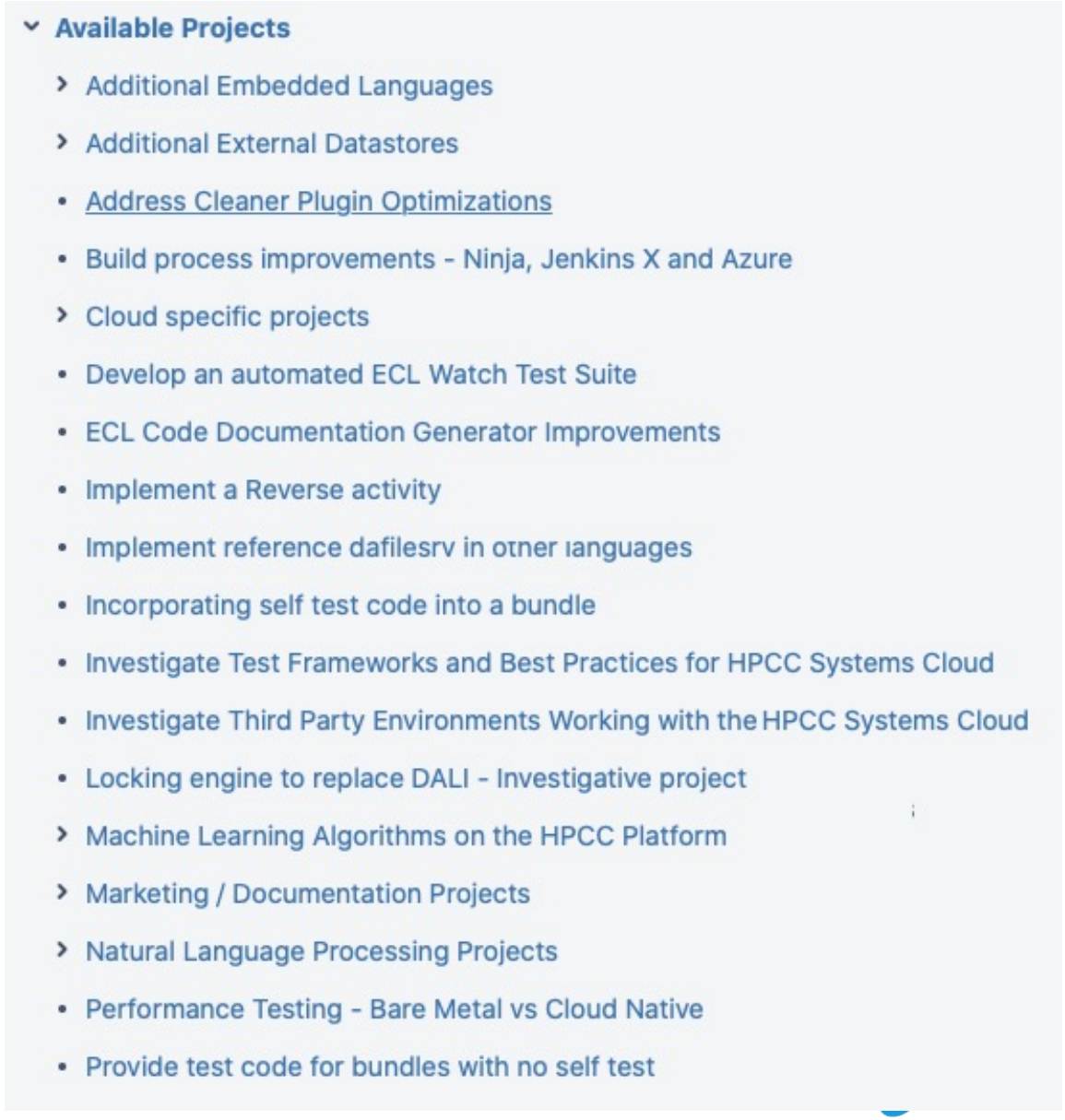
Project Proposal
Title : Implement Latent Semantic Analysis in ECL-ML **Deliverables : Will be implemented** 1) FUNCTION to convert Document Corpus into term-document Matrix efficiently 2) FUNCTION to perform SVD on constructed term-document Matrix 3) FUNCTION to reduce Components of SVD by given Rank 4) Transform initial Document Term Vectors into Reduced Representation 5) Implementation of "Folding-In" method of LSA to make addition of new Documents in pre-computed LSA results efficiently. 6) Checks to determine when LSA needs to be re-performed due to repeated "Folding-In" 7) FUNCTION to compute query representation in reduced dimension 8) FUNCTION to calculate Document-Query Similarity and return best matched documents 9) Tests and Documentation **Wishlist :** 10) Implementation of SVD for Dense Matrix 11) Checks for Performance in both Sparse and Dense Matrix format. 12) Improving accuracy of LSA by implementing Locality Sensitive Hashing **13)** Including other Information Retrieval Measures like Latent Dirichlet Model and Topic Modelling based on LSA

Timeline :	
Design of Workflow from Data Input till result production Collection of Test Documents. Best Sources include datasets found in TREC IR competitions as well as from Wikipedia for benchmarking. Preprocessing and dataset-specific cleaning of above documents	25th May – 5th June
Convert text documents in RECORDs using Enumerate Function in Docs Module	6th June – 15th June
Improve methods for cleaning and splitting Text Documents into words. Specifically, : Include implementation for SnowBall Stemmer, which performs universally better than Porter Stemmer Include implementation for Lemmatization using WordNet	

HPCC SYSTEMS®

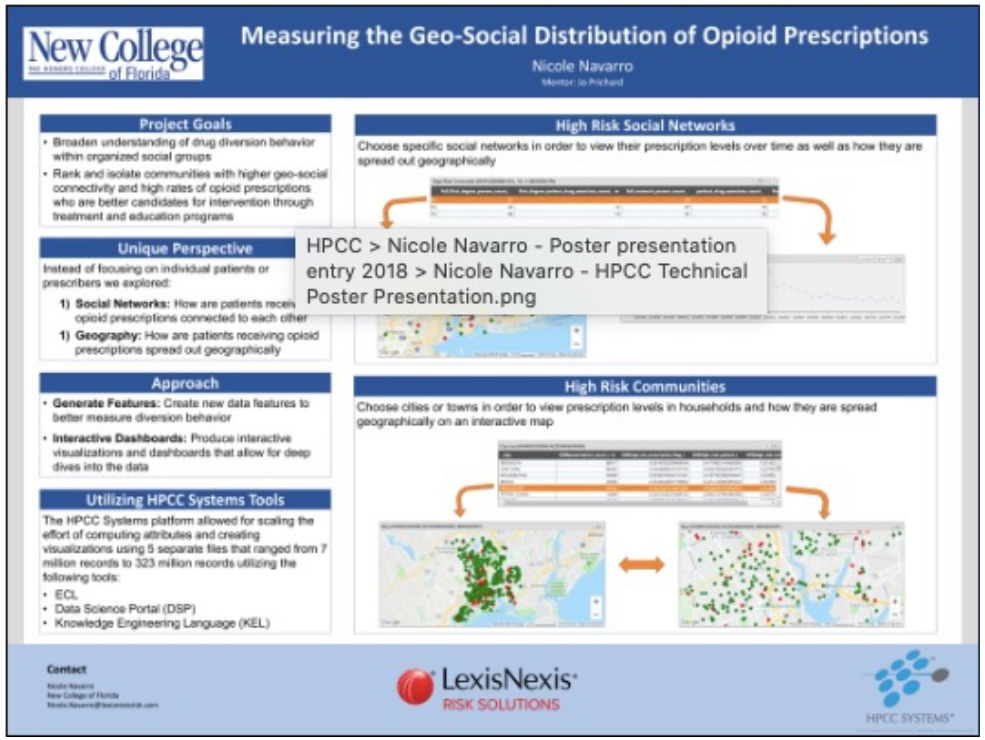
Mentoring and Support

- RELX Colleagues, School Teachers, Business Contacts, Professors
- Subject Matter Experts and Wellbeing
- Evaluations – Mid Term and Final
- Progress Reports - Weekly
- Daily stand-up call
- Email, Gitter, Teams/Zoom etc
- Intern Chat and Share



Opportunities to Share Your Work

- Personal Blog Journal
- HPCC Systems Blog
- Presentations
- Poster Contest
- Community Day
- Social Media Channels
- HPCC Systems Website
- GitHub Repository



HPCC > Nicole Navarro - Poster presentation entry 2018 > Nicole Navarro - HPCC Technical Poster Presentation.png

Applying HPCC Systems TextVectors to SEC Filings

Lorraine Chapman on 07/30/2020

This blog features a student from our 2020 HPCC Systems intern program. Matthias Murray, joined us from New College of Florida, to work on a machine learning related project which will be a significant and popular contribution to our machine learning library. Read on to find out about Matthias's achievements as he introduces his project and demonstrates its potential.

Personal Development Experience

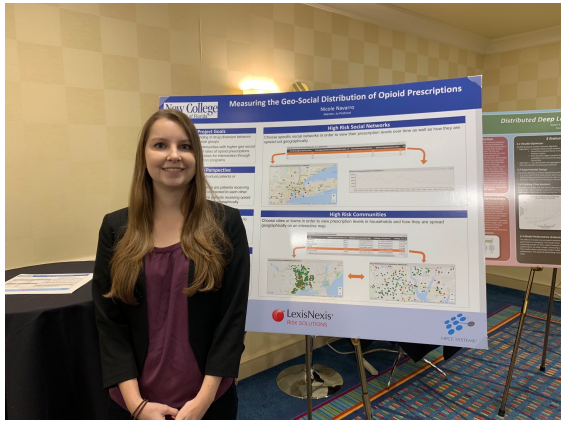
- Office/teamworking experience
- Project management
- Communication skills
- Research and development real world experience
- Bleeding edge development
- Confidence in decision making
- Use of developer tools and processes
- Contribute to an active open-source community

Be Part of the Team and Make Your Mark

Nicola Navarro

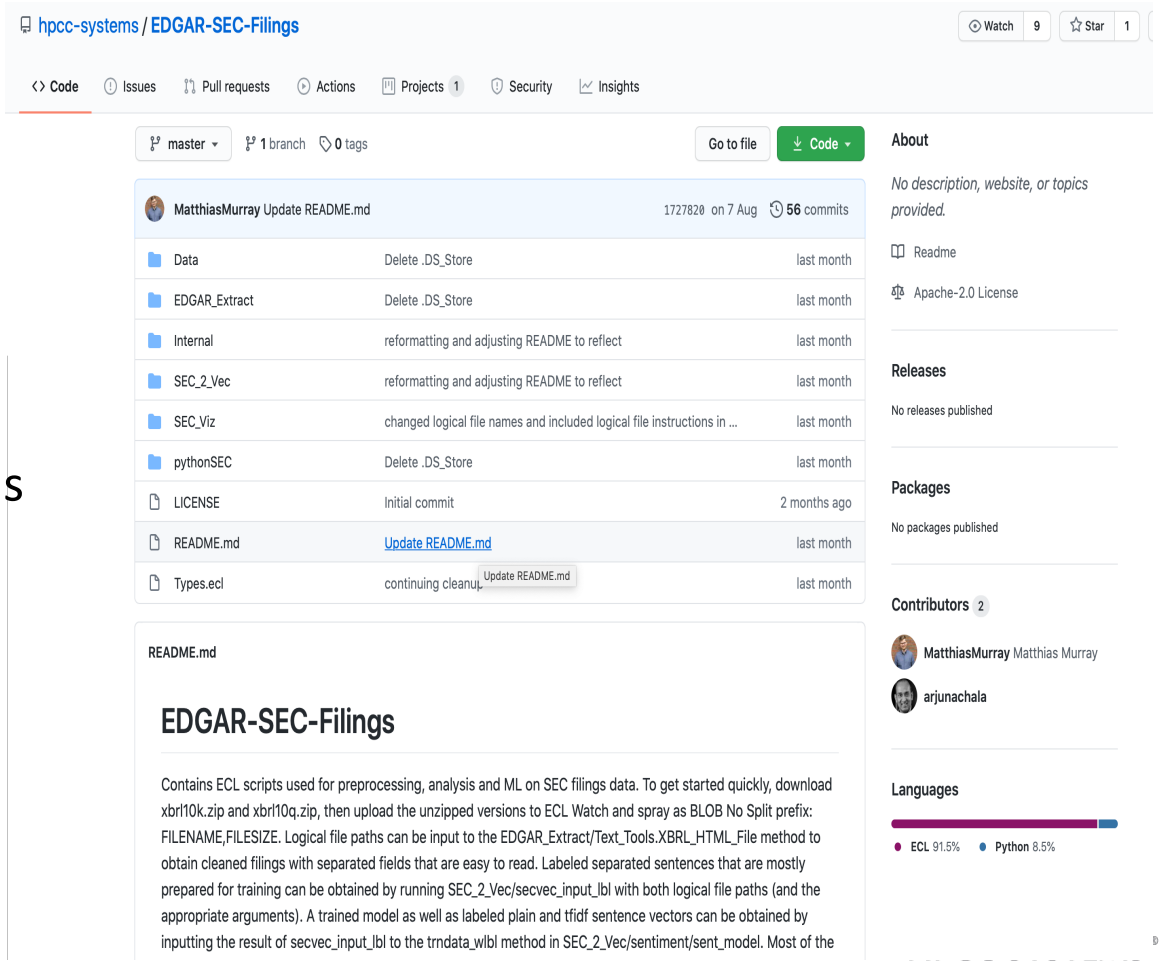
2nd Place Poster Winner 2018

Measuring the geo-social distribution of opioid prescriptions



Matthias Murray

Applying HPCC Systems word vectors to SEC filings



Future Employment Opportunities

- Open positions at LexisNexis Risk Solutions Group: https://risk.lexisnexis.com/group/careers
- Previous interns have a head start and we can vouch for you
- Hired 2 interns from the 2021 program
- Not all interns are ready for immediate employment
- Intern for a second time or a third…

Learn more…

- Student wiki - hpccsystems.com/student-wiki
- Available Projects List - hpccsystems.com/ideas-list
- Student journals and presentations https://wiki.hpccsystems.com/x/GQFdB
- Read about the program - hpccsystems.com/interns/
- Student Testimonials https://wiki.hpccsystems.com/x/KoNc
- HPCC Systems GitHub Repository https://github.com/hpcc-systems
- Student Posters https://wiki.hpccsystems.com/display/hpcc/HPCC+Systems+Technical+Presentations

Get in Touch

Lorraine.Chapman@lexisnexisrisk.com

HPCC SYSTEMS®