Rijul Vohra

rvohra@usc.edu

in: https://www.linkedin.com/in/rijul-vohra-367764126 **Phone Number** : 213-725-8073

: https://github.com/rijulvohra **Portfolio** : https://rijulvohra.github.io/rijulvohra/

Education

- **University of Southern California** Los Angeles, CA
  Master of Science in Applied Data Science; GPA: 3.88 August 2019-May 2021
  - **Relevant Coursework**: Data Management, Machine Learning, Natural Language Processing, Building Knowledge Graphs, Algorithms, Data Mining, Probability and Statistics
- **Thapar Institute of Engineering and Technology** Patiala, India
  Bachelor of Engineering in Electronics and Communication; GPA: 9.52/10.0 July 2015-June 2019

Skills

- **Languages and Databases**: Python, C++, SQL, NoSQL, MongoDB, MySQL, RDF, Blazegraph, SPARQL
- **Data Science Skills**: Linear Regression, Logistic Regression, Random Forests, Clustering, k-Means, Boosting, Support Vector Machines, Natural Language Processing, RNN, LSTM, Word Embeddings, Seq2Seq models, Knowledge Graphs, Linked Data, Hypothesis Testing, MapReduce, Spark
- **Data Science Libraries**: Pytorch, Scikit-learn, Numpy, Matplotlib, nltk, Pandas, Seaborn, SciPy, gensim, XGBoost, Scrapy, Beautiful Soup, Spacy, Snorkel, RDFLib, PySpark
- **Technology**: Jupyter Notebook, AWS EC2, S3, Linux, Git, Docker, Flask, Jenkins

Experience

- **USC Information Sciences Institue - Center on Knowledge Graphs** Los Angeles, CA
  Graduate Research Assistant - Mentor: Prof. Pedro Szekely August 2020-Present
  - **Harmonize**: *Python, KGTK, Wikidata dump, RxNorm data dump, Blazegraph, ElasticSearch*
  - Created a linked data source for drugs with **Wikidata** and **RxNorm**
  - Developed a pipeline to generate triples and load them to Blazegraph
  - Indexed **92 million** data items from wikidata, wikipedia using ElasticSearch
  - Developing algorithm for candidate generation for entities in a table by linking those entities to wikidata
  - Contributing to open source projects: KGTK and table-linker
- **Novartis International AG** East Hanover, NJ
  Data Science Intern, Data Strategy Team June 2020-August 2020
  - **FAIRification of Data**: *Python, Wikidata, SPARQL, RDF, AWS EC2, S3, Git, Docker, Jenkins*
  - Cleaned dirty transaction data by developing Linked Master Data Management using Wikibase infrastructure
  - Optimized reconciliation using OpenRefine, speedup by **50%**
  - Linked bioportal entities to Qnodes in Wikidata, overall precision **85%**
  - Streamlined reconciliation process for data curators by integrating entity linking webservice with OpenRefine UI
- **Marshall School of Business - USC** Los Angeles, CA
  Graduate Research Assistant - Mentor: Professor Gerard Hoberg November 2019-June 2020
  - **Business Open Knowledge Network**: *Python, Snorkel, Scrapy, Spacy, BeautifulSoup, FLAIR NLP*
  - Developed broad crawler to crawl **100,000** company's webpage extracting Mergers and Acquisitions
  - Achieved recall of **69%** on extracting names of target and acquirer companies
  - Extracted customer relations from Capital IQ database for 20 years(2000 - 2020) using Snorkel with an F1 score of **68.1%**

Projects

- **Knowledge Graph for Video Games and System Requirements(Github)**: *Python, fastText, rltk, scikit-learn, Git*
  - Game Recommendation System based on user likes and system specification, built using Knowledge Graph
  - Knowledge Graph also has links for cheapest purchase source
- **Machine Translation from German to English using Seq2Seq models(Github)**: *Pytorch, torchText*
  - Experimented with vanilla Seq2Seq model using LSTM
  - Used Attention as well for the Machine Translation task
- **Helping Robots Navigate(Github)**: *Python, scikit-learn, fastai, pandas, matplotlib, Jupyter, Git*
  - Multi Label Classification to help robot navigate the surface on which it is placed based on sensor data - acceleration, velocity. Achieved accuracy of **90.9%**.
- **Presentation Template Recommendation System(Github)**: *Python, Pytorch, Scrapy, BeautifulSoup, Git*
  - System that recommends presentation templates based on textual content with **MRR of 0.693**