Umang Mehta

Location: Chicago, IL
umangrmehta.github.io

Email: umangrmehta@gmail.com
linkedin.com/in/umangrmehta

Cell: +1 (812) 955-9507
github.com/umangrmehta

Objective

I aspire to work in the field of **Machine Learning**, **Artificial Intelligence,** and **Cognitive Science & Engineering** with applications involving **Speech Recognition**, **Natural Language Processing,** and **Conversational AI**.

Professional Profile

Data Scientist & **AI Research Engineer** with **2.5 years** of active research & development experience in **Machine Learning(ML), Neural Networks, Natural Language Processing (NLP)**, **Social Media Mining, Conversational AI**, **Information Retrieval (IR), Knowledge Graph**, **Predictive Analytics, Big Data Engineering, Python & Scala** and **5 years** of professional experience as **Full-Stack Software Engineer** & **Team Mentor** with expertise in **Database, Java** & **Web Development**.

Education

Master's Degree in Data Science **MAY 2019**
Luddy School of Informatics Computing and Engineering (SICE), Indiana University, Bloomington
Major Courses: Advanced NLP, Machine Learning, Deep Learning Systems, Data Mining, Elements of AI, Algorithms Design and Analysis, Statistical Inference, Advanced Database Concepts, High-Performance Big Data Systems. GPA: 3.6

Bachelor's Degree in Computer Engineering **MAY 2013**
K J Somaiya College of Engineering (KJSCE), Mumbai University, India
Graduated with SECOND Class, aggregate 59.77% equivalent to GPA 3.21.

Technical Skills

Languages:	Java (5 yrs), Scala (1 yr), Python (2 yrs), Javascript (5 yrs), C/C++, GraphQL, Gremlin, R, MATLAB, Groovy (1 yr), PHP(3 yrs)
Database:	**RDBMS-** MySQL(5 yrs), PostgreSQL, Oracle; **NoSQL-** MongoDB(1 yr), Apache Solr(2 yrs), Elasticsearch, Redis, DynamoDB; **Graph -** Neo4J, JanusGraph, Dgraph
Web Design:	HTML5-CSS3(4 yrs), JQuery(4 yrs), AngularJS(3 yrs), UnderscoreJS(3 yrs), AJAX(4 yrs), JSON(4 yrs), Bootstrap(4 yrs), Polymer
Frameworks:	Spring Framework(3 yrs), Hibernate ORM(3 yrs), Play Framework(1 yr), Morphia, Smack(XMPP), JSSC, Apache HTTP(3 yrs), PHP-MySQL(1 yr), Symfony Framework, Doctrine ORM, NumPy, SciPy, Pandas, NetworkX, Django, Flask, ND4J
Machine Learning:	TensorFlow(1 yr), Keras, Deeplearning4J, Scikit-Learn(1 yr)
NLP Toolkits:	Rasa, NLTK, Stanford CoreNLP, Spacy, OpenNLP, Spark NLP, Polyglot, Gensim
Big Data:	Hadoop, Spark, Kafka, Snowflake, One Lake, Nebula, Teradata
Project Utility:	**VCS -** Git, SVN; **Dependency -** Maven, Gradle, Composer, NPM; **SCRUM -** Jira
Servers:	Apache 2, Apache Tomcat 7-8, NodeJS, Netty, Openfire, Nginx, WildFly
Deployment Tools:	Docker, Kubernetes, AWS Lambda, AWS EC2, AWS Sagemaker, AWS EMR, Airflow
OS:	Ubuntu(Linux), Windows, macOS
Cloud:	Amazon Web Services(AWS), Google Cloud Services(GCS)
Certification:	Oracle Certified Professional JavaSE6 Programmer

Work Experience

Capital One, Chicago, IL **JAN 2019 - PRESENT**
Data Engineer

- Developing Data Processing Pipeline using **Apache Spark** and **Scala** to process Credit Card Requests in batches with joining data from other sources and APIs and make the final output available for Card Embossing Process which will replace the existing mainframe system making the process more efficient by 70%.
- Integrating Streaming Data Pipeline using **Apache Kafka** in various data pipelines.
- Building data pipelines for data transfer and warehousing using **Enterprise File Gateway, Snowflake and Apache Spark** for incoming data from external sources to be used by analytics intents like Anti-Money Laundering and Fraud Detection.
- Building serverless functions using **Python** to spin up a transient **AWS EMR** to run the data pipelines.
- Using **AWS Lambda**, **AWS EMR** AWS **CloudFormation** and **AWS S3** for production deployment.
- Maintaining code and development pipeline using **Git** on **GitHub**.
- Following **SCRUM Agile Development** processes using **Jira.**
- **Technologies & Tools:** Scala, Java, Apache Spark, Apache Kafka, Git, AWS EC2, AWS EMR, AWS Lambda, AWS S3, AWS CloudFormation, Linux, Spring

Hello Nesh Inc., Houston, TX **JUNE 2019 - OCT 2019**
Data Scientist - NLP & AI

- Implemented **Conversational AI** Assistant backend using **Rasa** framework in **Python** with a pipeline consisting of **Intent Classifier**, **Spacy**, **CRF Entity Extractor**, **POS Tagging** and **Constituent Parsing**.
- Trained and deployed **Deep Learning Neural Network** models with **Dense Neural Networks** and **LSTM** for Text Classification using **TensorFlow**, **Keras** and **SciKit-Learn** through **AWS Sagemaker** in **Python**.
- Built a Question Answer System based on **BERT** in **Python** served using **Flask** on **AWS EC2.**
- Developed a **Knowledge Extraction** pipeline for **100k+** public documents with stages and components involving **Web Scraping** and **HTML Parsing** with **BeautifulSoup**; Semantic Extraction using **Named Entity Recognition(NER)**, **Constituent Parsing** & **Anaphora Resolution** with **Spacy**, **Duckling** and **Conditional Random Fields**; **Topic Modeling** with **TextRazor** and **Gensim** in **Python** deployed through **AWS EC2, AWS DynamoDB** and **AWS S3**.
- Developed a **Knowledge Graph** using **Dgraph** as graph database and **GraphQL** as query language to represent the extracted knowledge with **1M+** entities, deployed on **Linux** using **Docker** and **AWS EC2.**
- Programmed a **Data Access Layer** for Knowledge Graph in **NodeJS** and **Javascript** deployed through **AWS Lambda**.
- Conceptualized and developed PoC of **Diagnostic Analysis and Predictive Analytics** of Oil Well Failure with **Maximum-Likelihood(MLE)** and **Maximum-A-Priori(MAP)** estimations and **Expectation-Maximization(EM) Algorithm** with basic **feature engineering** using **NumPy, SciPy, Pandas** and **SciKit-Learn** in **Python**.
- Maintained code and development pipeline using **Git** on **Bitbucket**.
- Followed **SCRUM Agile Development** processes using **Jira.**
- **Technologies & Tools:** Python, Java, Rasa, TensorFlow, Keras, SciKit-Learn, BERT, Stanford CoreNLP, Spacy, TextRazor, Gensim, Dash by Plotly, NumPy, SciPy, Pandas, Flask, Dgraph, Docker, Kubernetes, Nginx, Git, AWS EC2, AWS S3, AWS RDS, AWS Lambda, AWS Sagemaker, AWS DynamoDB, Gremlin, JanusGraph, Linux, Javascript, NodeJS

Kelley School of Business, Bloomington, IN **JUNE 2018 - MAY 2019**
Graduate Research Assistant under Prof. Matthew Josefy

- Worked under Prof. Matthew Josefy utilizing **ML** & **NLP** for research on Strategy & Entrepreneurship.
- Researched and implemented NLP methods to extract relevant information from **50k+** SEC Filings.
- Developed **predictive models** to analyze the business model and board leadership structure of companies and used them for **diagnostic analysis** as well.
- Built Text Classifiers using NLTK & SciKit-Learn with 90% accuracy measured with 10-Fold Cross Validation.

- Built ML models using **NumPy**, **Pandas** and **SciKit-Learn** in **Python**.
- Implemented **Semantic Extraction** with **NER** using **Stanford CoreNLP** and **Spacy** in **Java** and **Python**.
- Integrated a download and parsing pipeline for SEC Filings using **BeautifulSoup** in **Python**.
- Maintained code and development pipeline using **Git** on **GitHub**.
- Deployed the projects on an on-premise **Windows** system.
- **Technologies & Tools:** Python, Java, NumPy, SciKit-Learn, Pandas, Stanford CoreNLP, Spacy, Git, TensorFlow, Keras, BeautifulSoup, Windows

AriaData Inc.(Aridat), Bloomington, IN **FEB 2018 - MAY 2019**
Chief NLP Research Engineer (Unpaid Position)

- Built an analytics engine to determine the critical reception of an artist's work based on chatter on social media.
- Lead and advised on the research and implementations of advanced NLP methods to improve the efficiency of Sentiment Classifier and add new functionality to improvise analytics provided to artists.
- Developed a **Data Preprocessing** and **Feature Engineering** module using **NLTK** with steps involving **POS Tagging**, **Stop Words Removal**, **Stemming** and **Lemmatization**, **Negation Handling**, **N-gram** and **Sentiment Scores** using **Afinn** and **TextBlob**.
- Built Sentiment Classifier using an **ensemble** of **Naïve Bayes** and **Multiclass Logistic Regression** to classify tweets from artists as +1, 0 and -1 and implemented metrics to analyze sentiment distribution over different demographics.
- Deployed the project on an on-premise **Linux** system.
- **Technologies & Tools:** Python, Java, Numpy, Scikit-Learn, Pandas, Stanford CoreNLP, NLTK, Afinn, TextBlob, MongoDB, Git, Matplotlib, Plotly, TensorFlow, Keras, Linux

Vitruvian Technologies Pvt. Ltd.(VTPL), Mumbai, India **JUNE 2014 - JULY 2017**
Senior Developer & Team Mentor

- Developed functionalities for **web-based ERP** and **CRM** systems in the domain of **Real Estate**.
- Collaborated in a team of 12 for project development including **Object Oriented Programming** using **Spring** in **Java** & **Groovy** and **Reactive Programming** using **Play** in **Scala**
- Handled **Data Structure** & **Database Design** in **Java, Hibernate** & **MySQL**
- Designed and developed **Web Applications** with **HTML 5, CSS 3, Javascript, Jquery, AngularJS, UnderscoreJS, Spring MVC** in **Java** & **Laravel** in **PHP**
- Built **Full-Text Search** functionalities with **Apache Solr** and **ElasticSearch**
- Developed a routing software to redirect users to different tenant clusters based on the user logging in through a central login page using **Java**.
- Deployed projects using **AWS EC2, AWS S3, AWS RDS, AWS Route53, AWS Cloud CDN** and **Linux**
- Used **Git** and **SVN** as version control systems for different projects
- Lead, trained and mentored a sub-team of 5 throughout the development of the projects
- Contributed to the Core Framework, proprietarily used by the firm for project development.
- **Technologies & Tools:** Java, Scala, Groovy, Spring Framework, Hibernate ORM, MySQL, Apache Solr, ElasticSearch, HTML 5, CSS 3, Javascript, AngularJS, UnderscoreJS, Bootstrap, AJAX, Jquery, PHP, Laravel Framework, Play Framework, Git, SVN, AWS EC2, AWS S3, AWS RDS, AWS Route53, AWS Cloud CDN, Linux

Algonation, Mumbai, India **JUNE 2013 - MAY 2014**
Co-founder, Developer

- Developed websites, web portals and mobile apps using **HTML 5, CSS 3, Bootstrap, Javascript, AJAX, Jquery, PHP** and **MySQL** for small and medium scale enterprises.
- Built IM Chat Apps using **Java, XMPP, Smack** and **Openfire** on **Android OS.**
- Built server software for TCP Layer Protocols customized for cloud-based industrial requirements with **Java, Netty, RabbitMQ** and **MySQL** deployed on **Linux** using **Google Cloud Services.**
- Developed standalone and distributed software for some stages of manufacturing production lines using **Java** and **JavaFX** for **Windows**.
- Mentored and trained groups of 3-4 undergraduate interns for developing industry level projects.

- **Technologies & Tools:** Java, PHP, HTML 5, CSS 3, Bootstrap, Javascript, AJAX, Jquery, PHP, RabbitMQ, Netty, JavaFX, Openfire, Smack, XMPP, MySQL, Linux, Windows, Google Cloud Services

Research Innovation Incubation Design Laboratory (Riidl), Mumbai, India **JUNE 2012 - MAY 2013**
Software Engineer Intern

- Developed a web-based ERP application for educational institutes using **HTML, CSS, Bootstrap, Javascript, PHP**.
- Designed the **data structures** & **schema** and managed the database transactions using **MySQL**
- Built mobile App for the ERP using **Android SDK** for **Android** phones and **Java ME** for feature phones
- Deployed the ERP on a hosting service using cPanel.
- **Technologies & Tools:** Java, PHP, HTML, CSS, Bootstrap, Javascript, PHP, MySQL, Java ME, Android SDK, cPanel

Academic Projects

Neural Conversation Model — *Seq2Seq Learning using LSTM/RNN* **AUG 2018 - DEC 2018**

- Built a Proof of Concept(PoC) for Sequence-to-Sequence (Seq2Seq) Learning for a Neural Conversation Model with Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) using Deeplearning4J (DL4J) library.
- Trained Model on Cornell Movie-Dialog Corpus.
- Experimented with High-Performance Computing (HPC) optimization options in DL4J.
- Experimented with Big Data System coupling with DL4J for Hadoop and Spark.
- **Technologies & Tools**: Java, DL4J, ND4J, Hadoop, Spark, Intel DAAL, Git, Linux

OpenIE — *Open-Domain Information Extraction* **MAY 2018 - PRESENT**

- Worked for a research project under computational linguistics faculty at IU Prof. Damir Cavar for processing unstructured text to extract data, knowledge, entities and relations.
- Enabling semantic search with concept abstraction and linking concepts to concepts in large knowledge graphs like YAGO, DBPedia and Microsoft Concept Graph.
- **Technologies & Tools**: Python, Java, Numpy, Scikit-Learn, Pandas, Stanford CoreNLP, Spacy, OpenNLP, NLTK, Neo4J, Django, Git, WildFly, Linux

Speech Prosody and Pragmatics — *Detecting Prosody and Pragmatics of spoken language* **MAY 2018 - PRESENT**

- Working under computational linguistics faculty at IU Prof. Damir Cavar for this research project focusing on prosody, intonation contour detection, focus and stress pattern analysis for the processing of semantic and pragmatic aspects of spoken language.
- **Technologies & Tools:** Python, Java, Numpy, Scikit-Learn, Pandas, Google Cloud Speech API, Git

Image Orientation — *Classifier using Neural Network and Adaboost* **DEC 2017**

- Developed a Multi-Layer Feed-Forward Neural Network and an AdaBoost classifier in collaboration with 2 people to identify the orientation of an image.
- Achieved accuracy of over 71% in Neural Network and over 67% in Adaboost on a test set of 1000 images trained on a corpus of over 36000 images.
- **Technologies & Tools:** Python, Numpy, Pandas, Git, Linux

Part of Speech — *Tagging using Hidden Markov Model* **NOV 2017**

- Developed a program in collaboration with 2 people to identify parts of speech for English words in a sentence using Hidden Markov Model and Viterbi Algorithm.
- Achieved a word accuracy of over 95% and sentence accuracy of over 54% on a test set of 2000 sentences consisting of over 29000 words trained on a corpus of over 44000 sentences.
- **Technologies & Tools:** Python, Numpy, Pandas, Git, Linux

Pichu — *Simplified Version of Chess* **OCT 2017**

- Collaborated in a group of 3 people to develop a program in Python using the Minimax Algorithm to recommend a chess move for a given board state.
- Used Object Oriented Programming in Python to maintain the game states.
- **Technologies & Tools:** Python, Git, Linux

Leadership

Data Science Club at Indiana University **FEB 2018 - FEB 2019**
President and Treasurer

- Perfected the foundational structure of the Club Leadership.
- Organized hackathon for analyzing the opioid crisis in collaboration with SPEA at Indiana University.
- Spearheaded the initiative of monthly Newsletter and a semester wise e-Magazine.

Students' Council, K J Somaiya College of Engineering, Mumbai **JUNE 2011 - MAY 2012**
Creative Head

- Organized Technical and Cultural Festivals in KJSCE as Creative Head of the Organizing Committee.
- Headed the Creative Team for Kshitij, the Annual College Magazine of KJSCE.

Extra-Curricular

- Author and Editor of a Entertainment review blog "The Minimalist Critic" on Medium.
- Won 2 consecutive Annual Chili Cook-offs in 2017 and 2018 organized by SICE at Indiana University.
- Organized Bollywood Quiz as Event Head in Symphony, the Annual Cultural Festival of KJSCE in 2011.
- Volunteered in the Infra and Logistics Team for Technical and Cultural Festivals in KJSCE from 2009 to 2011.
- As a hobby writer, got my poetic verses published in Kshitij, the Annual College Magazine of KJSCE and articles on topics of technology published in the Technical Magazine of the Computer Society of India, KJSCE Chapter.