IST-782

Applied Data Science Portfolio

By:

Pankaj Yadav
Email: pyadav05@syr.edu
SUID: 429410363

Table of Contents

Learning Outcomes:

1. Advanced Data Analysis
    a. Objective: Elevate my expertise in statistical analysis for insightful decision-making.
    b. Plan of Action:
        i. Deepen theoretical understanding of statistics.
        ii. Gain proficiency in advanced data manipulation techniques.
        iii. Learn to implement complex statistical models for diverse datasets.
    c. Outcome: Skilled in transforming raw data into actionable insights through sophisticated analytical methods.

2. In-Depth Machine Learning
    a. Objective: Become adept in building and applying intricate machine learning models.
    b. Plan of Action:
        i. Study advanced machine learning concepts, including deep learning and AI ethics.
        ii. Engage in hands-on projects to apply theories in real-world situations.
        iii. Develop skills to critically evaluate and refine model performance.
    c. Outcome: Capable of creating robust, ethical machine learning solutions for complex problems.

3. Programming Expertise
    a. Objective: Attain advanced proficiency in essential programming languages.
    b. Plan of Action:
        i. Enhance skills in Python, R, and SQL through complex projects.
        ii. Explore additional languages and frameworks as needed.
        iii. Focus on writing efficient, clean, and reusable code.
    c. Outcome: Highly competent in using programming as a powerful tool for data analysis, machine learning, and beyond.

4. Comprehensive Data Management
    a. Objective: Master the management of large and intricate data sets.
    b. Plan of Action:
        i. Develop advanced skills in data cleaning, storage, and retrieval.
        ii. Learn to design and implement scalable data architectures.
        iii. Explore big data technologies and their applications.
    c. Outcome: Ability to efficiently manage and manipulate vast datasets, ensuring data integrity and accessibility.

5. Mastering Data Visualization
    a. Objective: Excel in creating visually striking and informative data representations.

b. Plan of Action:
    i. Master advanced visual design principles.
    ii. Learn to utilize cutting-edge visualization tools and software.
    iii. Develop techniques to tailor visualizations for varied audience needs.
c. Outcome: Proficient in crafting compelling visual narratives that make data easily comprehensible and engaging.

6. Integrated Business Acumen
    a. Objective: Merge technical skills with a deep understanding of business strategy.
    b. Plan of Action:
        i. Study business strategy and organizational behavior.
        ii. Engage in cross-functional projects to apply data insights in business contexts.
        iii. Develop communication skills for effective stakeholder engagement.
    c. Outcome: Adept at aligning data-driven strategies with business goals, effectively communicating insights to drive organizational success.

These enhanced learning pathways are meticulously crafted to foster a comprehensive skill set, positioning me at the forefront of the evolving field of data science. They represent a commitment to continual learning and adaptation, ensuring relevance and impact in my professional endeavors.

Program Learning Goals

1. **Mastery in Data Collection, Storage, and Accessibility**: This program is structured to empower students with the capabilities to efficiently handle data through advanced technologies. It involves imparting essential skills in identifying, selecting, and employing the best technologies for data collection and storage, including learning to gather data from diverse sources and manage it for optimal use and accessibility. The key outcome is for students to become adept at collecting, storing, and accessing data using state-of-the-art technologies, enabling them to make informed, data-driven decisions and contribute significantly to the data science field.

2. **Generating Actionable Insights in Diverse Contexts**: Students will navigate through the entire data science life cycle to create relevant insights. This includes a comprehensive understanding of the data science process, from collection to cleaning, analysis, and visualization. The course explores various domains like statistical modeling, machine learning, and natural language processing to derive meaningful insights applicable across societal, business, and political contexts. The aim is for students to discern patterns and trends, shaping decisions that impact various sectors effectively.

3. **Visualization and Predictive Modelling Expertise**: The program emphasizes the development of skills in data visualization for easier interpretation of complex data and predictive modeling to forecast future trends based on historical data. The intended outcome is

for students to master the art of presenting data visually and predicting future events, thus enhancing decision-making processes.

4. **Proficiency in Programming for Data Science**: Developing technical expertise in programming languages like R and Python is a crucial component. The focus is on training in data manipulation, analysis, visualization, and model building using these languages, preparing students for varied analytical challenges. Graduates will be proficient in leveraging these programming languages to generate actionable insights, influencing decision-making in various scenarios.

5. **Effective Communication of Data Insights**: This aspect involves skillful articulation of data analysis results to diverse audiences. Training includes creating engaging visualizations and contextualizing data analysis through storytelling, aimed at both technical and non-technical stakeholders. The program aims to make students excel in presenting their data findings compellingly and understandably, ensuring effective communication with various audiences.

6. **Ethical Considerations in Data Science**: Embedding ethical practices in data and predictive model development and use is a critical part of the curriculum. It instills an understanding of ethical challenges in data science, focusing on fairness, bias, transparency, and privacy, and strategies to address these issues responsibly. Graduates will be equipped to recognize and navigate ethical dilemmas in their work, ensuring responsible and ethical data science practices.

Through these learning outcomes, the program is designed to shape students into well-rounded data science professionals, equipped with the necessary skills to navigate and excel in the rapidly evolving field of data science.

Projects

Yelp Recommendation System using PySpark

About the Course:

IST718 Big Data Analytics is an engaging and comprehensive course that equips students with the skills necessary for the ever-evolving field of data science. In this course, students actively learn to translate complex business challenges into analytics problems, using tools like linear and logistic regression, decision trees, and neural networks for predictive analysis. They delve into data science to extract actionable insights and construct robust big data analytics pipelines using Python and Apache Spark. The course also immerses students in both classic and cutting-edge machine learning techniques, emphasizing how these advanced analytics create competitive

advantages in various industries. It's an ideal learning journey for those seeking to apply their knowledge in real-world data-driven scenarios.

Project Goals:

The project sets out to harness the vast potential of Yelp's big data, focusing specifically on restaurants in Pennsylvania. The aim is to develop robust recommendation engines using sophisticated algorithms. Key objectives include translating user preferences and business characteristics into personalized recommendations and refining the overall user experience on online platforms.

Technology:

This project employs PySpark, a powerful framework for big data processing, to manage the extensive Yelp dataset, which includes millions of user reviews, business profiles, and user information. The specific technologies and methodologies used are:

- K-means clustering for segmenting restaurants and users based on common attributes.
- ALS (Alternating Least Squares) Collaborative Filtering for predicting user preferences and recommending restaurants.
- Extensive data preprocessing and feature engineering to extract meaningful insights from the dataset.

Insights:

Throughout the project, several key insights emerged, including:

- A significant concentration of restaurant-related businesses as well as high user activity within the dataset was particularly observed in Pennsylvania.
- There was also a prevalence of diverse restaurant categories, revealing a varied culinary landscape.
- Patterns in user preferences and behaviors was observed which was essential for tailoring personalized recommendations.
- The distribution of restaurant ratings, user reviews over time, and business status (open or closed) provided crucial context for the recommendation engines.

Project Outcomes:

The culmination of this project is the successful development of two distinct recommendation engines. The K-means model effectively segments restaurants into six clusters, each representing different culinary themes, from Asian and Mexican cuisines to Italian eateries and vibrant nightlife spots. This model proved its merit in generating relevant recommendations for diverse users, as exemplified in the personalized suggestions for users like Desirae and Brett. Similarly, the ALS model demonstrated its predictive accuracy, with an RMSE value of about 1.30, indicating its effectiveness in anticipating user preferences. Improvement in the accuracy of recommendations as evidenced by the project's evaluation metrics. This project not only showcases the practical

application of advanced analytics in creating customized recommendations but also contributes to the evolving narrative of data-driven decision-making in the business realm.

This project exemplifies the practical application of the Program Learning Outcomes. It demonstrates adept handling of large-scale data through PySpark, aligning with the skills in data collection, storage, and accessibility. The development of K-means and ALS Collaborative Filtering models showcases the ability to generate actionable insights and apply predictive modeling, reflecting proficiency in programming languages like Python. Furthermore, the project's focus on delivering personalized restaurant recommendations aligns with effective communication of data insights, and implicitly adheres to ethical considerations in data science. Thus, it serves as a practical illustration of the comprehensive skills developed through the program.

SkillSpotter: Mining and Skills from Job Descriptions using NER

About the Course

IST736 Text Mining is a comprehensive course that delves into the field of text mining, teaching students to apply various techniques for real-world problem-solving across domains like social media and scientific literature. Emphasizing practical case studies, the course focuses on understanding and implementing text mining concepts, developing technical solutions, and effectively communicating results. Students are encouraged to be curious, think critically, and utilize their math and Python programming skills. The course primarily utilizes the SU BlackBoard System for communication and requires active participation and engagement in exploring the complexities of text mining.

About the Project

The project ”SkillSpotter” is a novel approach to extracting and categorizing skills from job descriptions using advanced text mining techniques. Utilizing Named Entity Recognition (NER), the study leverages a fine-tuned BERT model that accurately identifies both general and specific technological and soft skills from a vast dataset of web-scraped job descriptions. The project addresses challenges in data sampling, annotation, and model optimization while considering the ethical implications of data usage and technology in human resource contexts. The findings offer valuable insights into job market trends and skill demands, highlighting the potential of NER in job recommendation systems.

The objective is to develop a model capable of recognizing key entities, specifically focusing on two categories: general skills like "writing, critical thinking, persuasion," and technology skills, including "Python, R, SAP," among others. This model aims to execute comprehensive matching, ensuring it accurately identifies various iterations and abbreviations of a term, such as recognizing "machine learning," "machine-learning," "ml," "machine lerning,"(misspelled) and "ML" as equivalent expressions.

Technology/Methods Used:

SkillSpotter, a sophisticated text mining tool, leverages Named Entity Recognition (NER) and a fine-tuned BERT model to analyze job requirements from web-scraped descriptions. We used BeautifulSoup 4 for data cleaning and DistilBERT for model training, chosen for its efficiency in NER tasks. The methodology involved data transformation through tokenization, BIO tagging for preparing text for NER, and handling misalignments due to BERT's sub-word tokenization.

Insights:

From a dataset of 180K job descriptions, we found a vast variety of job titles, with a significant number from the healthcare industry. Notably, our analysis highlighted an imbalance in the dataset, with roles like Software Developer being more prevalent. The skills taxonomy creation process, involving O*Net OnLine and job description data, facilitated the identification of explicit and implicit skills. This process was integral in developing a model capable of capturing the nuanced requirements of different tech roles.

This model can be used to extract skills from any text tokenized by the DistilBERT tokenizer. During the model testing and inferencing, we tried using this model to extract skills from 2 sample resumes and calculated the cosine similarity between all skills required for each job title. This model can also be used as a recommendation system, recommending jobs to people with a high similarity score for skills present in the job description.

Outcomes:

SkillSpotter achieved an impressive accuracy of 98.98% and an F1 score of 93.78%. It effectively identified skills from job descriptions and unseen ChatGPT-generated sentences, including known skills, new phrases, some abbreviations, and minor misspellings. Unique skills relevant to specific job titles were also identified, illustrating the model's nuanced understanding of different roles. The model performed well and in about 15% of the cases it was able to find new skills in job descriptions that were not present in our skill taxonomy. Thus we can say that the best technique for a recommendation system would be to maintain an adaptable skill taxonomy along with a BERT model to identify all the skills present in job descriptions and update the taxonomy whenever required. The model's ability to identify new skills in job descriptions suggests its potential in maintaining an adaptable skills taxonomy, which can be updated as required.

Ethical Implications:

We were mindful of the ethical dimensions of our project. The open-source data used was in compliance with the terms of data providers, ensuring responsible use. However, we recognize the potential risks associated with over-reliance on automated systems like SkillSpotter. In fields such as human resources, where the subtleties of human judgment are vital, relying solely on an NER model for candidate screening or job matching can be problematic. It's imperative to have human intervention to interpret and contextualize the findings of such models. Another significant ethical consideration is the potential bias from training our model exclusively on US-based job descriptions and technology roles. This focus might limit the model's applicability and accuracy

when dealing with job descriptions from other regions or non-tech roles, raising concerns about its generalizability and fairness.

Sign-Vision: An Image Classification System For Sign Language Detection

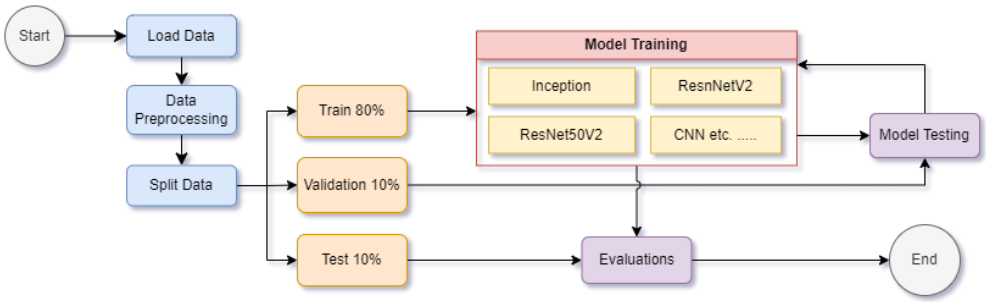
About the Course:

In IST 707, Applied Machine Learning, we are introduced to data analytics techniques, emphasizing both theoretical foundations and practical applications. The course covers essential tasks like data preparation, concept description, association rule analytics, classification, and clustering. The format includes lectures and lab sessions, giving us hands-on experience with open-source software to tackle real-world problems. We learn to document and analyze data analytics needs, apply relevant algorithms, and effectively communicate findings, preparing us to handle challenges in business, science, or other organizational contexts.

Project Goals:

The Sign-Vision project aimed to create an image classification model capable of accurately identifying signs in real-time. The primary objective was to develop a deep learning-based image classification model specifically designed to detect sign language. Focused on advancing a system for recognizing sign language images in real-time, the main goal was to develop image classification models that could accurately detect and identify different signs used in sign language. Utilizing advanced machine learning and deep learning techniques, the project achieved a high level of accuracy and precision in sign recognition. This initiative significantly enhanced accessibility for individuals with hearing impairments and served as a powerful communication tool in various settings.

Methodology:



To achieve accurate image classification, we sourced image data from open-source platforms like Kaggle. After collecting sufficient data, we preprocessed it to enhance its quality and ease of processing. We then split the data into three parts: a training set, a validation set, and a testing set. We allocated 80% of the data to the training set, 10% to the validation set, and the remaining 10%

to the testing set. This ensured that our models were trained on a substantial amount of data while also allowing for performance evaluation on unseen data. We trained multiple models using various architectures. To ensure optimal performance, we trained these models through multiple iterations and evaluated them on the validation dataset. This allowed us to fine-tune their hyperparameters and adjust their architectures to improve accuracy. Finally, once we obtained acceptable results on the validation dataset, we tested our models on the testing set for final evaluation. Based on these results, we selected the best-performing model for deployment.

To obtain the best-performing model, we performed this procedures on various CNN architectures ranging from small VGG architectures to large Resnet50 architectures. We also performed the same on Standard ML models like Random Forests and SVM for comparison.

Insights:

Classification Models	Accuracy
InceptionResNetv2 with ImageNet	92.41%
ResNetV2	99.35%
ResNet50V2	96%
CNN (1 layer)	98.26%
CNN (3 layers)	98.57%
Decision Tree	96.78%
Naive Bayes	42%
Multilevel Perceptron	98%

IsSign Models	Accuracy
Multilevel Perceptron	100%
Decision Tree	100%

The InceptionResnet V2 model has been found to be the best model for image classification of American Sign Language (ASL), achieving an impressive accuracy of 0.9935 and a validation accuracy of 0.9946. This model is based on the ResNetV2 architecture, which is designed to capture features at multiple scales and depths, making it ideal for image classification tasks.

Moreover, this architecture uses residual connections to mitigate vanishing gradients and improve the training of deep neural networks. It's important to note that other models, such as the custom sequential model with three convolutional layers and the decision tree model, also achieved high accuracy scores. Therefore, the selection of the appropriate model for a given project will depend on the specific requirements and constraints of that project.

In conclusion, the InceptionResnet V2 model is a highly effective model for image classification of American Sign Language. However, the choice of the most suitable model will depend on the specific needs of each project, taking into account factors such as accuracy, complexity, and speed of training and inference.

Challenges and resolutions:

Data quality: One of the primary challenges in this project would be ensuring the quality and quantity of the data. The model is only as good as the data it's trained on, so it's important to ensure that the dataset is diverse, representative, and contains enough samples of each class.

Resolution: Out initial dataset contained imbalanced data and standard sign language categories, so, to resolve this data quality issue, we loaded another dataset containing more diverse images and additional categories such as ‘nothing’, ‘del’, and ‘space’.

Data augmentation: In this project, the data is augmented using various techniques such as rescaling, shearing, zooming, and flipping. However, selecting the appropriate data augmentation techniques and their parameters can be challenging, as it can impact the quality of the model.

Resolution: For data augmentation, we had 2 choices. The first choice was to read the image data and apply preprocessing techniques and save the data. We used that data to train our conventional machine learning models. The second choice was to use an image generator function from tensorflow keras which would load the images from the directory and apply the augmentation techniques to those images. This was used when training our CNN models.

Model complexity: The ResNet50V2 model used in this project is a deep neural network with over 20 layers. While this makes it a powerful tool for image classification, it also makes it computationally expensive to train and can lead to overfitting if not properly regularized.

Resolution: We tried to train multiple models, however, some of them were highly complex and to resolve them, we tried to adjust our hyperparameters and train low number of epochs.

Hyperparameter tuning: Tuning the hyperparameters of the model and the data augmentation techniques can be a challenging task, as there is no single set of hyperparameters that work well for all problems.

Resolution: Initially, due to large dataset and highly complex models, the models trained took too long to train with multiple kernel crashes. To tackle this, we reduced the image size from 128x128 to 32x32, which reduced the number of trainable parameters. We also increased the batch size to 32 which reduced the training time and we trained the model in iterations with small epochs.

Hardware limitations: Training deep neural networks such as ResNet50V2 requires significant computational resources. The training time of the model could be hours or even days, and thus, hardware limitations such as the availability of GPUs can pose a challenge.

Resolution: To tackle this, we tried to train models after hyperparameter tuning on multiple google colab sessions and notebooks to distribute the processing load.

Conclusion

To Conclude, after building and testing multiple image classification models, we found that the ResNet50V2 model worked best when classifying sign language images, while the Multilayer Perceptron model performed better for binary classification of sign language images versus non-

sign language images. Our system showed promising results and can be used in a variety of applications, such as developing sign language translation tools, enhancing accessibility in public spaces, and aiding in sign language education. The use of multiple models and binary classification has proven to be an effective approach for improving the accuracy and precision of our system. However, there is still room for further improvement, such as fine-tuning the models for better performance on specific sign language dialects or including data augmentation techniques to improve the robustness of the system.

While developing the system, we encountered some challenges, but were able to overcome them with persistence and innovation. With this system, we believe that individuals with hearing impairments will have improved access to communication, education, and other essential services.

Neural Artistry: A Journey through Style Transfer

About the Course:

IST 691: Deep Learning in Practice offers an immersive introduction to deep learning, focusing on its principles and practical applications. In this course, we delved into deep learning methods for classifying data and predicting outcomes, understanding both the background and application of these techniques in real-world challenges. We gained an intuitive grasp of how deep learning functions and develop the ability to create solutions using state-of-the-art software. By the end of this course, we were able to translate data modeling needs into deep learning solutions, build models for specific predictive analytics challenges, and apply these concepts to real-world problems.

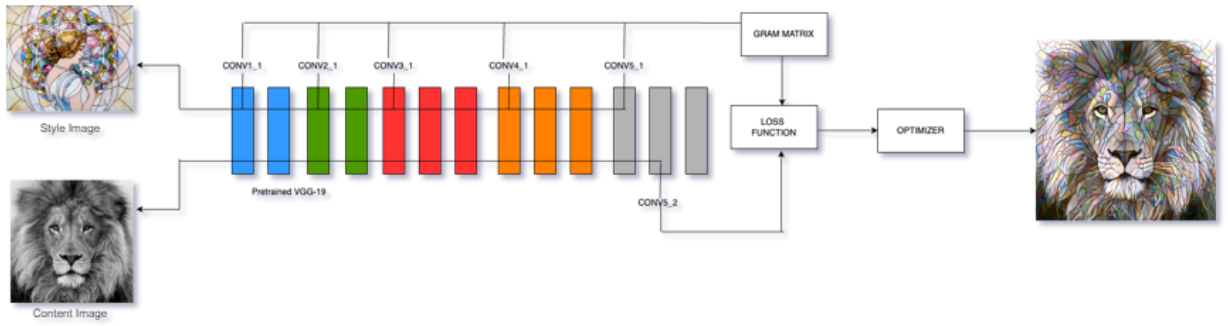
Project Goals:

This project delves into the intersection of AI and art through the lens of Neural Style Transfer (NST). My primary aim was to blend the essence of one image with the artistic flair of another, forging unique artworks. This exploration involved a detailed study of various optimizers, such as Adam and L-BFGS, alongside pre-trained neural network models like VGG19 and AlexNet. The endeavor was rooted in manipulating NST's loss function and experimenting with various hyperparameters, striving to create a diverse array of stylized images.

Technology Employed and Methodology:

The methodology unfolded in several structured phases:

- **Image Processing and Model Utilization**: Using VGG19 and AlexNet for their ability to capture complex image features effectively.
- **Layer Selection and Feature Map Analysis**: This involved selecting specific layers from these networks to extract content and style features, further encapsulated in a Gram matrix to emphasize patterns and textures.
- **Optimization and Style Integration**: Both Adam and L-BFGS were employed to optimize the NST process, adjusting pixel values to blend styles seamlessly.



In the neural style transfer process as shown in the above image, first the content image and style image are chosen as the inputs, which are then fed into a robust pre-trained Convolutional Neural Network (CNN) model like VGG19. This well-known pre-trained model captures complex features of both content and style images. Next, different layers in the CNN architecture are selected to capture different levels of abstraction. Typically, style is extracted from shallow layers (for our experimentation, we chose 5 layers), and content is extracted from the deeper layers (for our experimentation, we chose the last layer). The selection of layers is a balance between capturing intricate style patterns and preserving content details. We aim to study this balance. After the selection of different layers, the next step is Feature Map Extraction. The chosen layers are used to extract feature maps for both the content and style images once passed through the model, and these feature maps (representing the activation of neurons in the network) are stored in the form of a Gram Matrix G(x). This Gram matrix measures the correlations between different features, obtained by taking the dot product of the feature matrix with its transpose, emphasizing the patterns and textures in the style.

Following Feature Map Extraction, a style loss, and a content loss is calculated. The final loss is the sum of the two individual losses.

$$L = \alpha L_{content}(x) + \beta \sum L_{style}(G(x))$$

The final step is Optimization. Optimization algorithms such as Adam or L-BFGS are employed to iteratively adjust the pixel values of the generated image to minimize the overall loss. The objective is to converge to an image that matches the content of one image and the style of another. Throughout this process we can generate the image using the pixel value at each iteration, visualizing the change in the loss in our target image.

Experimental Insights and Data Representation:

For this project, multiple experiments were performed to in identify how changes in the NST Process such as the Neural Network, Loss Functions and other hyperparameters impact the resulting stylized images. The following are the observations from those experiments:

- **Optimizer Comparison**: A direct comparison between Adam and L-BFGS showed Adam's superior performance in achieving pronounced stylization. L-BFGS, on the other hand, produced slightly blurred images, indicating less style capture. This was reflected in the distinct visual outputs of each optimizer.

- **Layer Impact Analysis**: VGG19's middle layers, as opposed to its outer layers, were more effective in capturing styles. AlexNet, with fewer layers, showed that its outside layers captured colors well, while inside layers were better at structure representation.
- **Gram Matrix Rotation Experiment**: Rotating the Gram matrix yielded images with different style distortions, suggesting that orientation plays a role in style transfer. However, significant variations in style weren't observed, aligning with findings from related literature.
- **Multi-style Image Fusion**: Successfully combining two distinct artistic styles into one image was a significant accomplishment. This not only demonstrated NST's flexibility but also its potential in creating composite artistic expressions.

Conclusion and Impact:

The study concluded that Adam optimizer, combined with finely tuned parameters and the VGG19 model, forms an effective NST method. This method strikes a balance between preserving the content's integrity and imparting artistic style. Notably, the ability to blend two different styles into a single image was a breakthrough, enriching NST's application in digital art creation. The insights and methodologies developed in this research contribute significantly to the understanding and advancement of NST as a tool for artistic expression in the realm of AI.

Track: Language Analytics

I pursued the Language Analytics track, which encompassed two pivotal courses: Natural Language Processing (NLP) and Text Mining. This field merges language studies with data science to derive meaningful insights from extensive text datasets, uncovering patterns in how language is utilized across various contexts. The potential of language analytics extends to decoding human communication dynamics and advancing our understanding through deep learning implementations.

Machine Learning (ML) and Deep Learning are integral to the study of both NLP and Text Mining, underpinning the sophisticated analysis techniques used to process and understand large amounts of data. These technologies enable the automation of data analysis, allowing for more complex, accurate, and predictive analytics.

In practical terms, language analytics applies to analyzing customer sentiment, monitoring social media, and extracting patterns from healthcare data. Its relevance spans multiple sectors, enabling the development of skills in data analysis, machine learning, and NLP—skills that are in high demand across diverse industries. As digital and data-driven environments evolve, the significance of language analytics continues to grow.

Furthermore, this track does not confine my scope solely to language-related paths; it also opens the door to broader explorations within Machine Learning and Deep Learning fields. The skills and knowledge acquired are highly transferable, positioning me to delve into various aspects of AI and data science, beyond just language applications.

Natural Language Processing combines computer science with linguistics to interpret human language, aiming to make interactions with technology more seamless and intuitive. NLP is instrumental in improving cross-language communication, offering context-aware translations that foster global understanding. It also transforms industries like healthcare and finance by enabling the analysis of complex datasets to enhance diagnostics and financial strategies.

Text Mining leverages machine learning algorithms to mine insights from unstructured text, illuminating consumer behaviors, market trends, and more. This study equips me with valuable data science skills, including advanced machine learning techniques essential for handling large-scale data analyses. Text mining's adaptability makes it integral to sectors such as marketing and healthcare, where it continuously provides fresh insights and opportunities for innovation.

Embarking on this track has positioned me at the forefront of exploring and pushing the boundaries of language analytics and its applications. I am eager to contribute to and shape the future of this dynamic and crucial field, utilizing deep learning to enhance the interpretative power of language analytics tools, while also preparing to engage deeply with broader machine learning and deep learning pathways.

Data Science Learning Outcomes

The projects I undertook on Neural Style Transfer (NST), SkillSpotter, and Yelp review analysis have collectively broadened my technical expertise and analytical thinking in data science and artificial intelligence. Each project contributed uniquely to my skillset and professional development.

In the NST project, I mastered the art of blending images using deep learning, which enhanced my understanding of complex neural network architectures like VGG19 and AlexNet. It honed my skills in image processing and optimization algorithms, teaching me the subtleties of artistic style transfer. This project also developed my ability to visualize and interpret artistic data through technology, a skill that blends creativity with analytical thinking.

SkillSpotter deepened my experience in text mining and natural language processing, particularly in Named Entity Recognition (NER) using BERT models. It was a venture into the practical application of AI in extracting and analyzing textual data. This project sharpened my abilities in handling large datasets, understanding language nuances in job descriptions, and developing a keen eye for detail in data transformation and feature extraction.

The Yelp review analysis project enhanced my proficiency in handling real-world data. Working with a vast dataset of reviews, users, and businesses, I gained insights into user behavior and preferences. It allowed me to apply machine learning algorithms and data analytics techniques, improving my skills in data cleaning, preprocessing, and model evaluation. This project also strengthened my understanding of the business implications of data analytics.

Collectively, these projects have significantly contributed to my growth as a data science professional. They have equipped me with a diverse set of skills, from image and text processing to advanced machine learning and data analysis techniques. They have also developed my ability

to approach complex problems critically and analytically, delivering insights that drive decision-making. Moreover, these projects have enhanced my communication skills, enabling me to effectively present technical concepts and data-driven insights to varied audiences, ensuring clarity and stakeholder engagement. These experiences have been instrumental in preparing me for real-world challenges in the fields of data science and artificial intelligence.

Github References:

1. Yelp Recommendation System using PySpark: https://github.com/pankajyadav01/ADS_Portfolio/tree/main/Yelp%20Recommendation%20System
2. SkillSpotter: Mining and Skills from Job Descriptions using NER: https://github.com/pankajyadav01/ADS_Portfolio/tree/main/SkillSpotter
3. Sign-Vision: An Image Classification System For Sign Language Detection: https://github.com/pankajyadav01/ADS_Portfolio/tree/main/SignVision
4. Neural Artistry: A Journey through Style Transfer: https://github.com/pankajyadav01/ADS_Portfolio/tree/main/Neural%20Artistory