‘Text Generation with Hugging Face’

## Course Description:

This course explores the use of Hugging Face Transformers, a leading tool in natural language processing (NLP), for text generation. Students will learn to implement, fine-tune, and deploy various transformer models to produce human-like text, exploring both the technical implementations and creative applications.

To complete this course, you would benefit from having already completed ‘Deep Learning for Text with PyTorch’ and ‘Working with Hugging Face’. Upon finishing this course, you will have much more confidence using the Hugging Face state-of-the-art ‘Transformers’ library to complete challenging ‘Text Generation’ tasks. It will also serve as a helpful stepping stone to the larger ‘Introduction to LLMs in Python’ course.

## Course Outline

**Chapter 1: Using the Hugging Face Transformers Library**

**Lesson 1.1** | Predicting Next Word with RNNs

* Learner will be able to
  + identify the shift from traditional machine learning techniques to deep learning in NLP.
  + discuss key innovations such as word embeddings and RNNs.
  + describe the role of word embeddings and recurrent neural networks (RNNs) in advanced NLP.

**Lesson 1.2** | Using Hugging Face

* Learner will be able to:
  + understand the components of the Hugging Face ecosystem.
  + explain how Hugging Face components integrate to facilitate NLP tasks.
  + identify and utilize the core components of the Hugging Face library.
  + integrate these tools to successfully perform complex NLP tasks.

**Lesson 1.3** | Utilizing Pre-trained Models

* Learner will be able to:
  + set up a Python development environment tailored for NLP applications
  + utilize pre-trained models for initial text generation and classification tasks.

**Chapter 2: Generating Text using Transformers**

**Lesson 2.1** | Basics of Text Generation with GPT-2 (out of the box)

* Learner will be able to:
  + deploy an out-of-the-box GPT-2 model to generate text.
  + understand basic GPT-2 functionality and limitations.
  + create interactive, adaptive, and engaging stories with text-generation.

**Lesson 2.2** | Interactive Storytelling with GPT-2

**Lesson 2.3** | Fine-Tuning GPT-2

* Learner will be able to:
  + prepare and preprocess datasets specifically tailored for fine-tuning GPT-2
  + adjust and optimize hyperparameters for fine-tuning GPT-2 to improve the model’s performance
  + apply transfer learning techniques to enhance GPT-2’s genre-specific text generation.

**Lesson 2.4** | Using Llama or Mistral

* Learner will be able to:
  + describe the unique architecture and features of Llama and Mistral models.
  + compare Llama and Mistral models to traditional GPT architectures.
  + set up and deploy Llama or Mistral models for various AI applications.
  + consider the specific strengths and use cases of these different transformer models.

**Chapter 3: Fine-Tuning and Improving Inference**

**Lesson 3.1** | Quantization (taking an existing model and running it on lesser hardware)

* Learner will be able to:
  + understand and explain the necessity of quantization in machine learning
  + apply both post-training quantization and quantization-aware training techniques.
  + quantize AI models effectively, while maintaining performance considerations.

**Lesson 3.2** | Fine-Tuning Llama or Mistral

* Learner will be able to:
  + understand the fundamentals of fine-tuning to distinguish it from training a model from scratch.
  + prepare datasets for fine-tuning by selecting, cleaning, and labeling them appropriately.
  + set hyperparameters and use evaluation metrics to effectively measure and troubleshoot the fine-tuning process.

**Lesson 3.3** | Creating a Custom Writing Assistant

* Learner will be able to:
  + design and implement a custom writing assistant to generate specific text
  + refine generated text to meet qualitative standards and user-specific criteria.
  + integrate user feedback to continuously improve the writing assistant’s performance.

**Chapter 4: Deployment of Transformer Model**

**Lesson 4.1** | Creating a Chatbot Interface

* Learner will be able to:
  + develop advanced conversational agents using transformer-based models.
  + incorporate techniques that maintain context across interactions.
  + deliver a seamless and coherent user experience

**Lesson 4.2** | Performance Evaluation and Model Optimization

* Learner will be able to:
* use quantitative metrics like BLEU and ROUGE to evaluate text generation models.
* apply qualitative techniques to further refine and optimize AI text generation models.
* integrate quantitative and qualitative metrics for model evaluation and optimization.

**Lesson 4.3** | Ethical Considerations and Bias Mitigation

* Learner will be able to:
  + recognize biases in AI-generated text and understand their ethical implications.
  + apply strategies to mitigate biases in AI-generated content to enhance fairness and inclusivity.
  + evaluate and adjust AI models to promote ethical practices in content generation.

**Lesson 4.4** | Packaging a Model as an API

* Learner will be able to:
  + compare models of different sizes and inference speeds
  + understand relevant metrics for selection, such as tokens per minute.
  + evaluate AI models using TensorFlow or PyTorch
  + deploy AI models using Flask or FastAPI and implement security features

## (‘Exhaustive’) Course Objectives

* Evaluate and explain the transition from traditional machine learning models to deep learning in the context of NLP advancements.
* Analyse and describe the components of the Hugging Face ecosystem and their contributions to simplifying NLP tasks.
* Demonstrate the application of pre-trained Hugging Face models to execute basic NLP tasks like text classification and generation.
* Implement a text generation script using GPT-2 and assess the influence of various model parameters on the output quality.
* Distinguish between the architectures of GPT-2 and GPT-3 and evaluate their respective capabilities in generating text.
* Apply fine-tuning techniques to GPT models for generating text in specific genres and critically assess the performance improvements.
* Execute advanced sampling strategies in text generation projects and evaluate their effectiveness in enhancing text quality and coherence.
* Develop interactive fiction using text generation models and analyse the user engagement and narrative adaptability.
* Fine-tune models for poetic text generation and evaluate the stylistic accuracy and creativity of the output.
* Implement text generation for automated content creation and critically examine the relevance and quality of the generated content.
* Construct conversational agents using text generation and assess their effectiveness in maintaining context and dialogue flow.
* Deploy multilingual text generation models and evaluate their accuracy and adaptability across different languages.
* Develop strategies for scaling and deploying text generation models and evaluate their performance in production environments.
* Utilise various performance metrics to optimise text generation models and interpret the outcomes to guide further improvements.
* Identify and mitigate biases in text generation models and develop strategies to ensure ethical use of text generation technologies.

Looking at the existing DataCamp lessons on text generation using the Hugging Face ‘Transformers’ library, this is a ‘frontier’ area of generative AI that I feel can be deepened. There is definitely room for expanding on the topic of ‘text generation’, especially in different use-case scenarios.