**Applied Text Mining in Python**

In this course, you will build your skill sets in text mining through lectures on core topics, code walk-throughs on tasks related to weekly assignments, reading material, quizzes, and programming assignments.

In **Module One**, you will be introduced to basic text mining tasks, and will be able to interpret text in terms of its building blocks – i.e. words and sentences, and reading in text files, processing text, and addressing common issues with unstructured text. You will also learn how to write regular expressions to find and extract words and concepts that follow specific textual patterns. You will be introduced to UTF-8 encoding and how multi-byte characters are handled in Python. This week’s assignment will focus on identifying dates using regular expressions and normalize them.

**Key Concepts**

* Interpret text in terms of its basic building blocks: sentences and words.
* Identify common problems with raw text and perform text cleaning tasks in Python.
* Write regular expressions to find textual patterns.

In **Module Two**, you will delve into NLTK, a very popular toolkit for processing text in Python. Through NLTK, you will be introduced to common natural language processing tasks and how to extract semantic meaning from text. For this week’s assignment, you’ll get a hands-on experience with NLTK to process and derive meaningful features and statistics from text.

**Key Concepts**

* Describe different natural language tasks.
* Process free text through the NLTK toolkit to tag language constructs onto text.
* Derive meaningful features from text.

In **Module Three**, you will engage with two of the most standard text classification approaches, viz. naïve Bayes and support vector machine classification. Building on some of the topics you might have encountered in Course 3 of this specialization, you will learn about deriving features from text and using NLTK and scikit-learn toolkits for supervised text classification. You will also be introduced to another natural language challenge of analyzing sentiment from text reviews. For this week’s assignment, you will train a classifier to detect spam messages from non-spam (“ham”) messages. Through this assignment, you will also get a hands-on experience with cross-validation and training and testing phases of supervised classification tasks.

**Key Concepts**

* Compare text classification to other classification approaches (covered in Applied Machine Learning in Python as well)
* Describe the Naive Bayes and Support Vector Machine algorithms
* Classify text in two classes using one of these approaches in Python
* Identify and extract features from text and transform them into feature vectors for the machine learning models

In **Module Four**, you will be introduced to more advanced text mining approaches of topic modeling and semantic text similarity. You will also explore advanced information extraction topics, such as named entity recognition, building on concepts you have seen through Module One and Module Three of this course. The final assignment lets you explore semantic similarity of text snippets and building topic models using the gensim package. You will also experience the practical challenge of making sense of topic models in real life.

**Key Concepts**

* Apply WordNet-based similarity measures on text
* Derive semantic topics from a large text collection using LDA
* List and describe techniques for named entity recognition and other information extraction tasks.