Bootcamp Information Sheet

Instructor

Name: *Elliot Stern*

Bio: *Elliot Stern is a freelance Data Scientist with 7+ years of experience, including working with an Olympic team and creating an NBA salary model for Hazan Sports Management, an NBA sports agency. He co-founded a soccer computer vision company that secured $250,000 of funding. Elliot was a leader in creating computer vision products for the vertical jump and juggling a soccer ball. His sports algorithms have generated over $100,000 in profit and are more accurate than those of industry leaders such as ESPN and Yahoo. Elliot currently works as a data science freelancer taking on a myriad of data science challenges, including regression, computer vision, dashboard creation, and clustering problems.*

Bootcamp Details

ID: *< Assigned by Skillsoft >*

Bootcamp Title: *Introduction to NLP and Text mining in Python*

Number of Days: *4 days*

Hours per Day: *3 hours (2.5 hours of instruction + 0.5 hours for Q and A)*

Type of Instruction: *Lecture, polling questions (knowledge checks), and exercises*

Description: *This program will give students the foundational skills they need in order to process, clean, and format text data for analysis. By the end of this course, students will be able to integrate text mining into their work. Students will also learn to clean and process large amounts of text data, segregating text*

into different groups and topics, as well as finding similarities between different documents. As natural language can be vague and subjective, the course also presents ways to evaluate and interpret these language models.

Target Audience: *Learners who have experience using Python to manipulate and visualize data and want to expand their skills in programming in Python and text mining analysis.*

Technologies: *Python and Anaconda*

Prerequisites: *Attendees must be comfortable using Python to manipulate data and must know how to create advanced visualizations in Python.*

Student References: *Class slides, class exercises and code, datasets, list of packages needed to be installed for the class.*

Bootcamp Syllabus

Day 1

- **Introduction to NLP - 1**
    - Concept of NLP and its usage in industry
    - Corpus in Text data

- **Text Processing - 1,2**
    - Tools in Python to work with text data and create and inspect a corpus object
    - Steps to pre-process text for the bag-of-words approach
    - Text cleaning steps for bag-of-words approach
    - Implement the text cleaning steps to pre-process documents

Day 2

- **Text Processing - 3**
    - Create Term-Document Matrix
    - Distribution of words in corpus

- **TfIdf - 1,2**
    - Use cases for bag-of-words approach
    - Supervised vs. unsupervised learning
    - Need for weighting terms in a corpus
    - Weigh text data with term frequency inverse document frequency (TF-IDF)

Day 3

- **Topic Modeling - 1,2**
    - Topic modeling
    - Process of LDA (latent dirichlet allocation)
    - Load data and perform text cleaning on one document
    - Perform text cleaning on entire corpus

Day 4

- **Topic Modeling - 3,4**
    - Perform LDA on frequency counts
    - Evaluate results and choose optimal number of topics

- Visualize results of LDA using interactive LDAvis plot
- Extract and interpret document-topic information

Package Versions for Code Scripts Execution of Introduction to NLP & Text mining in Python

```
# Python version: 3.11.0
pandas==2.2.1
matplotlib==3.8.3
wordcloud==1.9.3
nltk==3.8.1
scikit-learn==1.4.1.post1
gensim==4.3.2
pyLDAvis==3.4.1
```