Jade Sanchez

2376 ITAI

2/26/2025

Reflective Journal

Lab 01

Learning Insights

My experience included practical work with text processing techniques that are essential for natural language processing (NLP). My understanding of text transformation into machine learning modelready structured data improved through performing tasks which included word cloud generation along with stemming and lemmatization and part-of-speech tagging.

Every exercise in this category addressed crucial data preprocessing methods that form essential parts of improved model accuracy in machine learning. The text processing skills acquired from this stage enable raw text data to be suitable for more sophisticated NLP application usage.

The process of discovering how to create the word cloud proved to be the standout impactful discovery during the exercise. The experience demonstrated to me that visualizations assist people to quickly identify primary elements which exist within extensive text collections. Word clouds created a clear representation which showed the key terms prominently so that text cleaning in advance of visualization proved essential.

Challenges and Struggles

I faced problems employing the Stemmer class function from PyStemmer during my work. The initial ambiguity about library division selection produced several incorrect attempts before the correct utilization.

Detailed readings of instructions as well as documentation served as my approach to handle these difficulties. My success in mastering spaCy syntax demanded full understanding of tokenization methods along with stemming while joining them with spaCy system operations.

My solution strategy consisted of checking preprocessing operations first followed by reviewing function output after which it proceeded to the next steps.

Personal Growth

My knowledge about machine learning was initial at the start because I failed to recognize the specific details of text processing before this point. These preprocessing steps have become vital to my understanding of approach building for effective and accurate modeling because they focus on unstructured text data.

Stemming and lemmatization demonstrated enhanced power during this part of the project.

Observing the extent to which these techniques decrease word variations and enhance model performance made me understand their essential nature.

The skills I acquire now will enable me to reach my future AI-related objectives by intensifying my work in Natural Language Processing as well as machine learning. My professional responsibilities that include data analysis and AI-based work such as sentiment analysis tools and chatbots need the fundamental mastery of text preprocessing methods.

Critical Reflection

My strategy for better repetition of the labs would include increased test time for experimenting with multiple types of text data to study the performance of stemming and lemmatization approaches in varied contexts.

The application of text processing methods on NLP models particularly BERT and GPT remains an unsolved research area. This research aims to discover what advanced processing procedures complex models need for their operation.

These labs provide a basis for grasping text-based machine learning model mechanics and their operation. The text preprocessing practices fit within the overall machine learning research by showing how textual data needs specialized precleaning methods that allow its effective usage in AI systems.

L02

Learning Insights

My work during this lab provided important knowledge about the Bag-of-Words (BoW) text vectorization process and its multiple machine learning applications. The major lesson revolved around teaching machines to process unstructured text data through numerical conversion to become analyzable with computers.

The BoW technique enables conversion of text into numerical vectors according to the methods I learned during this evaluation. Binary classification practice allowed me to identify how words appearing or not appearing in sentences gets recorded for multiple machine learning purposes such as detecting sentiment and spam.

Understanding the role which word frequency plays in document representation became possible through implementing word counts and Term Frequency (TF). Document importance analysis becomes better when word occurrence data is more critical than basic text presence data.

TF-IDF: This text analysis method extended my knowledge of text data by implementing the Inverse Document Frequency (IDF) measurement system. I understand the importance of TF-IDF by recognizing how it selects important words through document frequency analysis combined with overall dataset term distribution.

The hands-on exercises created the most valuable learning experience as they allowed me to see the vectorization results presented through Pandas DataFrames. Each data transformation method became more straightforward to comprehend through visual representation of structural changes.

Challenges and Struggles

The main challenge I faced was grasping the distinctions between the BoW variants that included binary classification and word counting and TF-IDF. At first it seemed like the transformation process was dull since I failed to recognize why specific methods held better value than the others.

I eliminated this difficulty by returning to the method descriptions and examining the generated output results simultaneously. I devoted time to test CountVectorizer and TfidfVectorizer parameters so I could comprehend how parameter adjustments affect final vector output.

I discovered effectiveness in problem-solving by performing small separate tests which allowed me to identify how various techniques affected the data results. Process segmentation enabled me to recognize potential problem points so I addressed them step by step.

Personal Growth

Through this lab my knowledge of machine learning progressed at an important level. Prior to this experience I regarded machine learning as an algorithmically based field yet I understand now that effective feature extraction with vectorization methods constitute fundamental parts towards building effective machine learning models.

Critical Reflection

My repetition of these labs would concentrate on performing experiments using extensive datasets to understand method scalability. I would focus on investigating Word2Vec and GloVe word embedding methods in my second try since I want to understand how they differentiate from BoW at retaining word semantic meaning.

I want to investigate how the text representation methodologies can be integrated with different features obtained through word embedding and character sequence modeling.

The wider machine learning framework includes this lab because it showcases how preprocessing techniques matter to NLP operations. Basic knowledge about text vectorization serves as a necessary base for multiple advanced aspects of machine learning which include document clustering, machine translation and automated summarization.

L03

LEARNING INSIGHTS

This lab provided instruction about word embeddings with a focus on GloVe (Global Vectors for Word Representation) as the technique for vector representation of words in high-dimensional space. The lab demonstrated to me how pre-trained word embeddings including GloVe serve to identify word semantics. The part I found most meaningful in this experience was understanding how cosine similarity functions as a method for word comparison based on word vector data. A cosine similarity model demonstrated to us how artificial intelligence evaluates word meaning alignments by analyzing the relationship between "cat" and "dog" within computational analysis. NLP tasks heavily depend on vector space word relationship theory which remains essential in the field.

CHALLENGES AND STRUGGLES

The installation process of required libraries proved to be a major struggle which I had to overcome. I started by encountering a dependency installation error from the requirements.txt while fixing it manually by installing torchtext specifically. The technical difficulty of this task required me to focus on setup procedures in the environment which remains essential knowledge for data scientists. Understanding cosine similarity and its practical implementation for word vector evaluations required me to spend time learning this aspect. The examination of vector comparison required me to repeat tests to validate understanding and validate accurate vector comparison. I improved the situation by reviewing the lab directions and using the code to detect various word pair results.

My knowledge about machine learning and NLP has made substantial progress from the beginning to the end of this laboratory work. Word embeddings were previously my only knowledge about word embeddings. Presently I feel capable to describe the functionality of GloVe pre-trained models in addition to demonstrating cosine similarity applications for word meaning comparison. My upcoming NLP and machine learning projects will require the knowledge I learned in this subject to operate text analysis or develop chatbots. Basic techniques proved to be effective methods for machines to understand and process human language at their current capability level.

For a repetition of this experiment I would dedicate additional time to testing GloVe vectors of multiple dimensions (100d as well as 200d) to determine their effect on word similarity assessment performance. I would develop a sophisticated model to accept these embeddings for performing NLP tasks including sentiment analysis or text classification. Studying word embedding integration within complicated machine learning models would provide me with improved knowledge about embedding integration at this level. The relationship between word embeddings stands out as an aspect which I wish to investigate further with more advanced models specifically transformers. GloVe embeddings present themselves against modern BERT models through evaluation of both their operational performance and resource utilization.

My experience in this lab proved valuable because it enabled improved knowledge of Natural Language Processing methods and word embeddings. The GloVe training activities offered fundamental knowledge which serves as the basis for advancing with machine learning and NLP applications. The content about word embeddings has provided me with increased self-assurance along with the motivation to implement this expertise in practical applications where large text databases are present.

L04

Learning Insights

The course has broadened my knowledge about machine learning by focusing on Support Vector Machines (SVM) techniques in classification methods. By modifying kernel type parameters I mastered the process of fine-tuning models so they generated better results. The practical application of convolutional neural networks (CNNs) and natural language processing (NLP) techniques reinforced the need for stemmings and lemmas in data cleanup steps. Real-world applications involving the "Chihuahua vs. Muffin" dataset during the workshop allowed me to fully understand how CNNs extract features because this experience became the highlight of my learning journey.

Challenges and Struggles

The main struggle necessitated finding the right configuration for neural network parameters. I faced initial challenges in understanding adjustments of learning rates combined with batch sizes until I mastered the concept by applying practice and research from outside resources. Cross-validation became my chosen method for preventing both overfitting and underfitting since it enhances model generalization.

Personal Growth

Machine learning knowledge has progressed for me from abstract theoretical knowledge into a concrete practical application. The key revelation during learning consisted of how rapidly models evolve through slight alteration adjustments. I possess solid confidence when implementing machine learning methods for my present academic work alongside my professional activities in AI and data science.

Critical Reflection

I would center my lab work on exploratory data analysis (EDA) since it leads to better model results if I had another chance to conduct these labs. Research will focus on ethical aspects particularly related to AI bias because this issue plays a crucial role in ensuring sensitive domains remain unbiased. I am interested in the process of transfer learning which enables the reprogramming of pretrained models for new applications.

The laboratory assignments have delivered me a strong base of machine learning understanding that leads me toward advanced field topics.