Student: Adejare Fasiku
Course: ITAI 2373 - NLP - L02 Basic Preprocessing Techniques
Date: July 23, 2025

Key Learning Insights

Completing Lab 02 has fundamentally transformed my understanding of text preprocessing as the critical foundation for all medical NLP applications, the most profound insight I gained is that preprocessing decisions in healthcare contexts carry significantly higher stakes than in general NLP tasks when working with clinical notes, patient records, or medical research papers, seemingly minor preprocessing choices can impact patient safety and clinical decision-making.

The comparison between NLTK and spaCy revealed crucial differences that directly apply to my career goals in medical AI, while NLTK's granular tokenization approach preserves every textual element including punctuation that might indicate urgency in clinical notes—spaCy's sophisticated linguistic analysis provides the semantic understanding necessary for complex medical entity extraction, this distinction became particularly clear when processing the social media text containing health related emojis, in medical contexts, patients increasingly use social platforms to discuss symptoms and treatment experiences, making spaCy's robust handling of informal text invaluable for monitoring patient sentiment and adverse drug reactions.

Challenges Encountered and Overcome

The most significant challenge I encountered was understanding when to preserve versus remove certain textual elements in medical contexts, the stop words removal exercise highlighted a critical issue: words typically considered "stop words" like "no," "not," and "without" are absolutely essential in medical texts for indicating negation and absence of symptoms, this realization forced me to recon sider the standard preprocessing pipeline and think more critically about domain specific adaptations.

Another challenging concept was the trade-off between stemming and lemmatization in medical applications, Initially I was drawn to stemming's efficiency, but the demonstration of its limitations—particularly how it failed to connect "better" with "good" or produced non-words like "wa" from "was"— revealed its inadequacy for medical NLP, In healthcare, maintaining semantic relationships between related terms is crucial for accurate clinical outcome analysis and patient progress tracking.

The tokenization comparison also presented unexpected complexity, while I initially assumed more sophisticated tokenization was always better; I learned that the choice depends heavily on the specific medical NLP task, for clinical documentation analysis, NLTK's preservation of all punctuation elements might be preferable, while spaCy's linguistic analysis excels in medical entity extraction and relationship identification.

Connections to Medical AI Applications

Throughout the lab, I consistently connected each preprocessing technique to real world medical scenarios, which deepened my appreciation for the field's complexity. The tokenization exercises made me consider how clinical abbreviations and medical terminology require specialized handling. For instance, "Dr." versus "Dr" or "2.5 mg" versus "25 mg" represent critical distinctions where preprocessing errors could have serious consequences.

The stop words analysis particularly resonated with my interest in clinical decision support systems. Understanding that negation words are crucial for accurate diagnosis extraction has implications for how I would approach building systems that analyze electronic health records. Similarly, the lemmatization discussion connected directly to medical literature mining, where connecting related concepts like "effective," "efficacious," and "beneficial" is essential for comprehensive research analysis.

The social media text processing opened my eyes to the growing importance of patient-generated health content. As patients increasingly share health experiences online, the ability to accurately pre- process informal medical discussions become crucial for public health monitoring and personalized healthcare applications.

Future Applications and Career Development

This lab has provided me with a solid foundation for my career development in medical AI. I now understand that successful medical NLP requires not just technical proficiency with preprocessing libraries, but also deep domain knowledge about healthcare contexts and clinical workflows. The preprocessing techniques I've learned will be directly applicable to projects I envision working on, including clinical decision support systems, patient sentiment analysis tools, and medical literature mining systems.

Looking Forward

The most valuable takeaway from this lab is the recognition that preprocessing is not a one-size-fits-all process. Each medical NLP application requires careful consideration of domain-specific requirements, potential consequences of

preprocessing decisions, and the balance between efficiency and accuracy. As I progress through the remaining labs in this course, I will carry forward this nuanced understanding of preprocessing as both a technical skill and a critical thinking exercise.

This foundation in text preprocessing has prepared me to tackle more advanced NLP concepts with a medical perspective, always considering how each technique can be adapted and optimized for healthcare applications where accuracy and reliability are paramount.