Erwin Cheng

L02_Reflection Journal

ITAI 2373 Natural Language Processing

Anna Rachapudi

Key Insights Gained

Key Insights:

Context Matters Preprocessing Needs: There is no one-size-fits-all approach to preprocessing. The best way to do it depends on the NLP task, text type, and end goal. Social media text will be vastly different than processing academic text.

Trade-offs Are Inherent: Any decision you make to preprocess will have trade-offs - for instance, removing stop words could reduce dimensions of the text but could also remove important context for some tasks. There are other strong methods for reducing dimensions, such as stemming, which can make your work portion of the process much more efficient, but could also change meaning.

Library Selection Matters: NLTK and spaCy have different strengths - NLTK offers more customization and control, while spaCy provides a better pre-trained model representation and out-of-the-box greater accuracy.

Real-world Text is Messy: The lab demonstrated how real-world text contains numerous challenges like emojis, hashtags, mixed cases, and irregular formatting that require sophisticated cleaning approaches.

Pipeline Design is Crucial: Building a modular, configurable preprocessing pipeline allows for experimentation and optimization based on known use cases rather than relying on fixed approaches.

Challenges I encountered:

Decision Complexity: Determining the right combination of techniques for different scenarios was challenging, as the impact of each preprocessing step isn't always obvious.

Information Loss Awareness: Recognizing what information might be lost during preprocessing and whether that loss would be causing some problem during downstream tasks requires careful consideration.

Library Limitations: Both NLTK and spaCy have limitations in handling certain edge cases, particularly with very informal text like social media content.

Connections to Real-World Applications

Search Engines: Understood how stemming helps match query variations while maintaining performance at scale.

Sentiment Analysis: Recognized the importance of preserving emotional cues (emojis, punctuation, intensifiers) that would be lost with aggressive preprocessing.

Chatbots: Appreciated the need for balanced approaches that maintain meaning while reducing complexity for real-time applications.

Content Recommendation: Saw how different preprocessing approaches could affect topic modeling and content similarity.

The most valuable insight was realizing that preprocessing isn't just a technical process but a strategic decision that should be driven by the specific use in each case. The same text might have a completely different preprocessing for sentiment analysis than for an information extraction purpose versus search optimization.

I also appreciated the importance of testing and validation - rather than assuming certain techniques will help, it's crucial to measure their actual impact on each task performance.

This lab fundamentally changed my perspective on text preprocessing from being a routine cleaning step to being a critical strategic component of NLP system design.