NLP Pipeline

NLP Presentation for SOS

Abhi Jain
23b0903

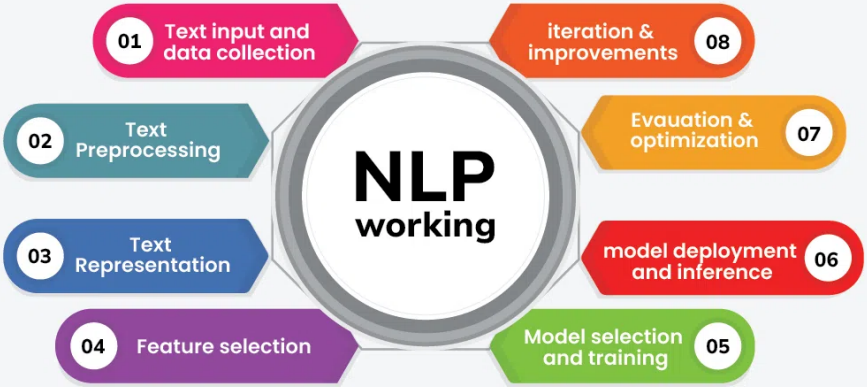
July 25, 2024

Outline

Introduction

- NLP pipeline: a series of steps to process raw text before ML algorithms.
- Ensures text data is clean, relevant, and usable.
- Example: Sentiment analysis of movie reviews.

Natural Language Processing



Data Acquisition

- Sources of data:
  - Internal databases (e.g., MongoDB, MySQL)
  - Public datasets (e.g., Google Dataset Search, U.S. Census Bureau)
  - Web scraping: using libraries like BeautifulSoup, Scrapy
  - Product intervention
  - Data augmentation: generating synthetic data
- Example: Scraping product reviews from e-commerce sites.

Text Extraction and Cleanup

- Discard irrelevant information: removing headers, footers.
- Extract required fields: using regular expressions or parsing libraries.
- Fix spelling errors: using spell-check libraries.
- Eliminate unnecessary new line characters.
- Example: Cleaning Twitter data by removing hashtags and mentions.

Text Preprocessing

- Tokenization: splitting text into words or sentences.
- Stop-word removal: removing common words like "and", "the".
- Stemming and Lemmatization: reducing words to their base form.
- Punctuation removal: removing punctuation marks.
- Example: Preprocessing a news article.

Text Representation

- Transform text into numerical format for ML algorithms.
- Techniques:
  - One-Hot Encoding: representing words as binary vectors.
  - Bag of Words (BOW): representing text as word occurrence vectors.
  - N-gram Models: capturing context with adjacent word sequences.
  - TF-IDF: weighing terms by frequency and importance.
  - Word Embeddings: using pre-trained models like Word2Vec, GloVe.
- Example: Representing text for a spam detection model.

Model Building

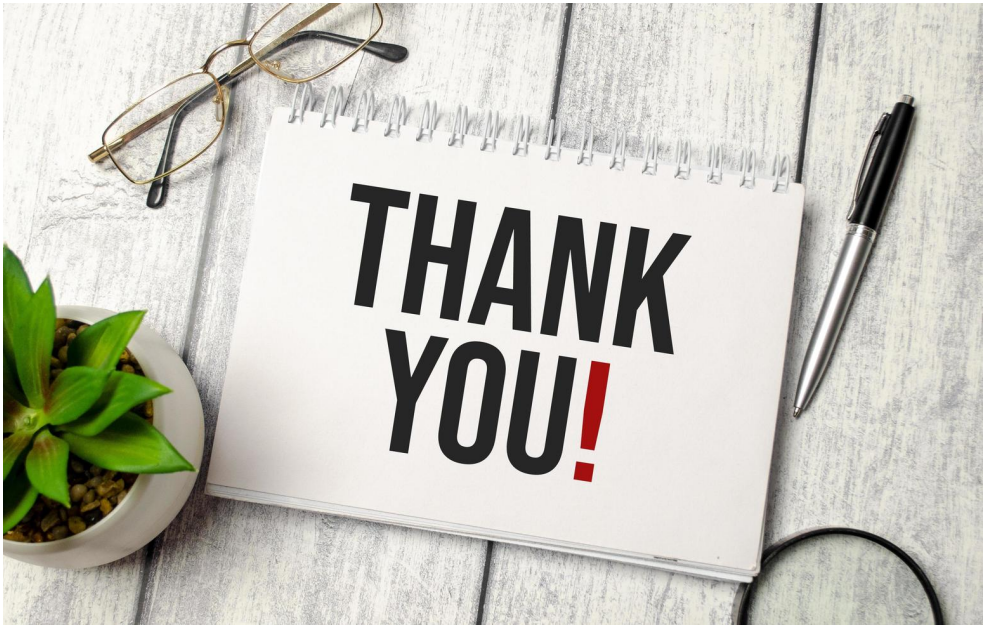
- Feature Extraction: selecting relevant features for the model.
- Model Selection: choosing algorithms like neural networks, RNNs, transformers.
- Hyperparameter Tuning: optimizing model parameters.
- Training: fitting the model to the training data.
- Evaluation: assessing model performance with metrics.
- Example: Building a chatbot using a transformer model.

Model Deployment

- Integration: embedding the model into an application or system.
- Monitoring: ensuring ongoing performance and accuracy.
- Maintenance: re-calibrating and updating with new data.
- Example: Deploying a sentiment analysis model in a customer service system.

Conclusion

- Creating an NLP application is a multi-stage process.
- Each stage is crucial for the model's success and adaptability.
- Ensures the model works as intended and adapts to new data.


THANK
YOU!