Text Preprocessing and Language Basics

Summer

Agenda

- Regular Expressions
- Corpora
- Linguistic Basics
- Tokenization
- Lemmatization
- Stemming
- Part of Speech (POS) Tagging
- Stop Word Removal
- Data Cleaning
- More Preprocessing Techniques

Regular Expressions

- Regular expressions (RegEx) are tools for efficiently finding, matching, and manipulating patterns in text.
- Key features:
  - Brackets `[]` for character disjunctions or ranges (e.g., `[a-z]`).
  - Wildcard `.` matches any character.
  - Quantifiers:
    - `*`: Zero or more occurrences.
    - `+`: One or more occurrences.
    - `?`: Optional matching.
- Applications: Text search, validation, and replacement.

Corpora and Text Mining

- **Corpora**: Collections of written or spoken material stored digitally.
- **Text Mining**: Transforming unstructured text into structured formats to extract insights.
- Best practices include providing dataset cards with metadata (source, content, ethical considerations).

Linguistic Basics

- Tokenization: Splitting text into tokens (words, subwords, or characters).
- Lemmatization: Extracting the base dictionary form of words (e.g., *running* $\rightarrow$ *run*).
- Stemming: Crude heuristic to remove affixes (e.g., *processed* $\rightarrow$ *process*).
- POS Tagging: Identifying grammatical categories (e.g., noun, verb).
- Stop Word Removal: Removing frequent but semantically unimportant words (e.g., *and, the*).

Tokenization Techniques

- Top-Down (Rule-Based): Define standards and rules.
- Bottom-Up: Use statistics to derive subword tokens (e.g., Byte Pair Encoding).
- Subword Tokenization: Efficient for handling unknown words and reducing vocabulary size.

Normalization Techniques

- Case folding: Converting text to lowercase.
- Lemmatization and stemming: As described above.
- Sentence Segmentation: Dividing text into sentences.

Advanced Processing

- Dependency Parsing: Identifying syntactic relationships between words.
- Shallow Parsing: Grouping phrases without full grammatical analysis.
- Data Cleaning: Preparing datasets (e.g., filtering out irrelevant content).

NLP Libraries

- **spaCy**: Fast, extensible, and well-documented.
- **NLTK**: Comprehensive library for NLP tasks.
- **Stanza**: Provides linguistic annotations.
- **TextBlob**: Simplifies common NLP tasks.

Resources

- Stanford NLP: `https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf`
- Hugging Face NLP Course: `https://huggingface.co/learn/nlp-course`