NATURAL LANGUAGE PROCESSING

Minor 1- Syllabus

Module 1: Foundations of Text Processing and Linguistic Structure

Text Processing

1.1 Introduction and overview of NLP, Data Extraction and Collection: types of datasets, sources (web, APIs, social media), web scraping (BeautifulSoup, Scrapy), handling large text data, handling noisy and imbalanced data, Preprocessing: Tokenization, stemming and lemmatization, stopword removal and punctuation handling, n-gram, Regex, POS Tagging, and NER.

Statistical NLP

1.2 Topics in Information Retrieval: Page Rank algorithm, vector space model, vector similarity, TF-IDF, BOW, CBOW, skip-gram, Annotating Linguistic Structure: Lexical analysis- word lexicons, word net, collocations, syntactic structure, dependency parsing, probabilistic context free grammars, BPE, text summarization- extractive and abstractive and multi-document text summarization.

1.3 Semantic relations: Word Sense Disambiguation; supervised, unsupervised and semi-supervised approaches, maximum entropy modeling, Lesk algorithm, Zipf's Law, WordNet and WordNet based similarity measures, EM algorithm, distributional measures of similarity, statistical machine translation, concept mining using latent semantic analysis.
Packages: BeautifulSoup, Scrapy ,NLTK, spaCy and networkx.

Suggested Readings:

1. Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

2. Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Link: https://web.stanford.edu/~jurafsky/slp3/