TECH LADDER

APRIL 5, 2024
VOLUME 2

Newsletter on Natural Language Processing

BY TAVISH AGGARWAL

INTRODUCTION

In the world of Large Language Models (LLMs), it is important to understand the basics of the language and how machines can understand such language. This will not only help one to build better language models but at the same time will also help in developing a deep understanding of biases that may exist in the language model.

With this as an aim and helping one to develop a deep and intuitive understanding of the language I am sharing this newsletter.

This newsletter is divided into three sections: lexical processing, syntactic processing, and semantic processing, which will help how machines can sequentially understand complex language structures starting from basics to advanced understanding.

With all of the excitement, let's start by exploring the meaning of lexical processing, syntactic processing, and semantic processing.

DEMYSTIFYING NLP: EXPLORING LEXICAL, SYNTACTIC, AND SEMANTIC PROCESSING FOR POWERFUL NATURAL LANGUAGE UNDERSTANDING

This is the master article covering lexical, syntactic, and semantic processing.

Lexical processing involves text cleaning and preparation, syntactic processing examines sentence structure and grammar, and semantic processing seeks to understand the hidden meanings behind words. It discusses various methods and tools for each stage, such as tokenization, stop word removal, stemming, lemmatization, and TF-IDF representation, providing examples and Python code snippets to illustrate these concepts.

The article serves as a comprehensive guide for those working on NLP projects.

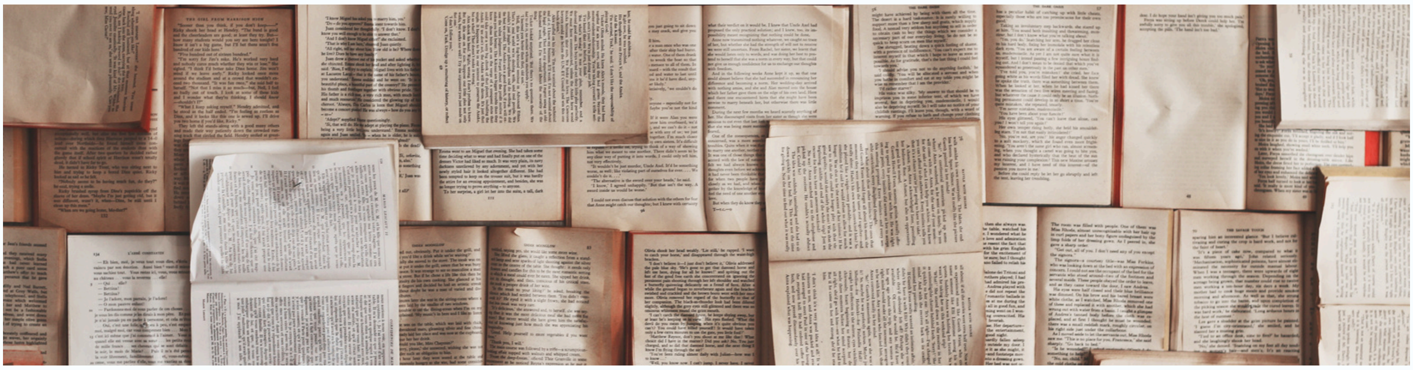
READ MORE

LEXICAL PROCESSING

In the field of natural language processing (NLP), lexical processing plays a crucial role. It is a starting stage where we do basic text processing and text cleaning. Some of the techniques that we use as part of lexical processing include calculating word frequencies, removal of stop words, tokenization, representing text sentences in the bag-of-words representation or TF-IDF representation, and treating words with stemming and lemmatization. These techniques are discussed in the article “Demystifying NLP: Exploring Lexical, Syntactic, and Semantic Processing for Powerful Natural Language Understanding” under the Lexical Processing section.

- **Tokenization**: This involves breaking down a text into smaller units, often words or subwords, known as tokens. It can be as simple as splitting text by whitespace or punctuation.
- **Stemming**: Stemming is the process of reducing words to their root or base form by removing affixes (prefixes or suffixes). For example, "running" would be stemmed to "run."
- **Lemmatization**: It is similar to stemming, but it aims to reduce words to their canonical or dictionary form, known as the lemma. For example, "ran" would be lemmatized to "run."
- **Normalization**: It involves transforming text to a standard or canonical form. This may include converting all characters to lowercase, removing punctuation or special characters, etc.

These lexical processing techniques are often applied as preprocessing steps before more advanced NLP tasks such as sentiment analysis, named entity recognition, or machine translation. They help to simplify and standardize the text data, making it easier for algorithms to process and analyze.



UNLEASHING THE POWER OF ADVANCED LEXICAL PROCESSING: EXPLORING PHONETIC HASHING, MINIMUM EDIT DISTANCE, AND PMI SCORE

We’ve explored various techniques, including Word Frequencies, Stop Words Removal, Tokenization, Bag-of-Words Representation, Stemming, Lemmatization, and TF-IDF Representation. However, some cases remain challenging, such as handling misspellings and multi-word terms. To address this, in this article, we delve into advanced techniques like Phonetic Hashing, Edit Distance algorithms, and Pointwise Mutual Information (PMI). Phonetic Hashing reduces words to base forms, while Edit Distance helps correct misspellings. PMI scores guide tokenization decisions for multi-word terms. These techniques enhance NLP’s robustness and accuracy.

READ MORE

SYNTACTIC PROCESSING

Following lexical analysis, the subsequent phase involves extracting additional information from the sentence, this time utilizing its syntax. Rather than solely focusing on individual words, we examine the syntactic structures, such as the grammar of the language, to grasp the intended meaning of the sentence by using the syntax and grammar of the sentence.

The syntactical analysis looks at word order and meaning, the role of stop words, the morphology of words, parts-of-speech, and dependencies aspects in the sentence that lexical processing doesn't. The article "Demystifying NLP: Exploring Lexical, Syntactic, and Semantic Processing for Powerful Natural Language Understanding" under the Syntactic Processing section covers the basics of this.

- **Part-of-Speech Tagging (POS tagging)**: POS tagging involves assigning a part-of-speech tag (e.g., noun, verb, adjective) to each word in a sentence. This information helps in determining the role of each word in the sentence and aids in syntactic analysis.
- **Parsing**: Parsing involves analyzing the syntactic structure of sentences to identify the grammatical relationships between words. This process often results in the creation of a parse tree or syntactic structure tree, which represents the hierarchical organization of words and phrases according to the rules of grammar.
    - **Dependency Parsing**: Dependency parsing is a more detailed form of parsing that focuses on identifying the syntactic dependencies between words in a sentence. It represents these dependencies as directed links between words, indicating which words are dependent on (or modify) others.
    - **Constituency Parsing**: Constituency parsing involves identifying the constituent phrases or units within a sentence, such as noun phrases, verb phrases, and clauses. It breaks down the sentence into its grammatical components based on the rules of syntax.

Syntactic processing is widely used in applications such as question-answering systems, information extraction, sentiment analysis, grammar checking, etc. By understanding the syntactic structure of sentences, NLP systems can better comprehend the meaning and context of textual data, leading to more accurate and effective language processing.

DEMYSTIFYING PART-OF-SPEECH (POS) TAGGING TECHNIQUES FOR ACCURATE LANGUAGE ANALYSIS

POS tagging is a natural language processing technique that assigns grammatical labels (such as nouns, verbs, adjectives, etc.) to words in a sentence. Various techniques exist for POS tagging, including lexicon-based approaches, rule-based methods, and probabilistic models like Hidden Markov Models (HMMs). Lexicon-based taggers assign the most frequent tag to each word, while rule-based approaches define rules for tag assignment. HMMs use probabilities to assign tags based on context. Learn more about such techniques...

READ MORE

APRIL 5, 2024

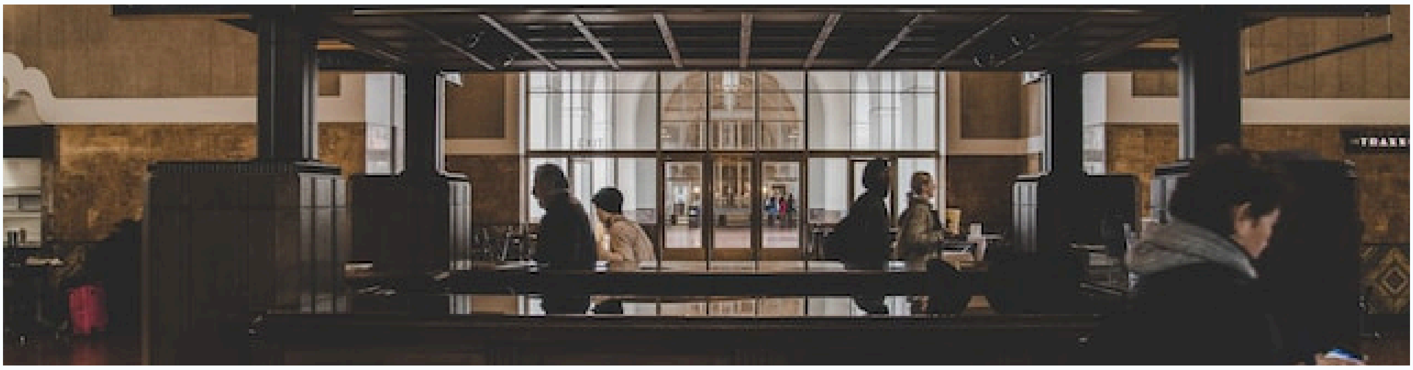
VOLUME 2

SYNTACTIC PROCESSING CONTINUE....

DECODING LANGUAGE STRUCTURE: EXPLORING CONSTITUENCY PARSING AND DEPENDENCY PARSING IN NLP

Parsing involves breaking down sentences into grammatical constituents. Constituency parsing uses context-free grammar (CFGs) to divide sentences into noun phrases, verb phrases, and other constituents. It assigns POS tags and builds parse trees. Dependency parsing, on the other hand, establishes relationships directly between words, focusing on subject-verb-object dependencies. While constituency parsing is based on CFGs, dependency parsing relies on supervised learning and projectivity. Both techniques have their strengths and applications in natural language processing. Refer to article to learn more......

READ MORE



APPLICATION: MASTERING NAMED ENTITY RECOGNITION: UNVEILING TECHNIQUES FOR ACCURATE ENTITY EXTRACTION IN NLP

It’s time to put all of the learnings into action by building an information extraction system. Information Extraction (IE) processes unstructured text to extract specific information, transforming it into a structured format. IE identifies entities, relationships, and events, enabling machines to comprehend human expression. Techniques include rule-based models, probabilistic sequence labeling, and Conditional Random Fields (CRFs). Building an IE system involves preprocessing, POS tagging, entity detection, and relationship extraction. Are you excited to build your own IE?

READ MORE

SEMANTIC PROCESSING

There has been a lot that we have covered so far. You deserve a break as going through all this is not easy. This is the last topic on a journey of understanding how machines learn language.

Lexical and syntactic processing alone is insufficient for developing sophisticated NLP applications like language translation and chatbots. Semantic processing is what we are missing in a puzzle. It goes beyond the individual words and their grammatical relationships to grasp the intended message or interpretation conveyed by the text. It aims to capture a deeper understanding of language, including nuances, context, and ambiguity.

Various techniques such as Word Sense Disambiguation, Distributional Semantics, and Topic modeling help machines to understand the meaning of sentences like we humans understand the meaning of the sentence. The article “Demystifying NLP: Exploring Lexical, Syntactic, and Semantic Processing for Powerful Natural Language Understanding” under the Semantic Processing section covers the basics of this.

- **Word Sense Disambiguation (WSD)**: It is the task of determining the correct meaning of a word based on its context within a sentence, aiding in more precise natural language understanding and interpretation. There are various supervised approaches and unsupervised approaches (the popular one is the Lesk algorithm) used to find the sense of the word.
- **Distributional Semantics**: It describes the words that appear in the same contexts and have similar meanings. We use word vectors to represent words in a format that encapsulates their similarity with other words. We use occurrence matrix such as LSA or non-occurrence matrix such as HAL, Word2Vec, or GloVe to represent words as vectors.
- **Topic Modeling**: It is a statistical technique used to identify abstract topics or themes present in a collection of documents, providing insights into the underlying structure and content of the text data. Techniques such as Matrix Factorisation-based Topic Modelling, Explicit Semantic Analysis (ESA), and Probabilistic Models such as PLSA and LDA are used to find a common theme among collection of documents.

Semantic processing enables NLP systems to understand and generate human-like interpretations of text, leading to more accurate and meaningful language processing.

BUILDING BLOCK OF SEMANTIC PROCESSING: INTERPRET THE MEANING OF THE TEXT

In the field of natural language processing (NLP), making machines understand language like humans is a challenging task. This article covers tools and concepts that can help machines understand the “meaning” of the word in a given sentence like humans.

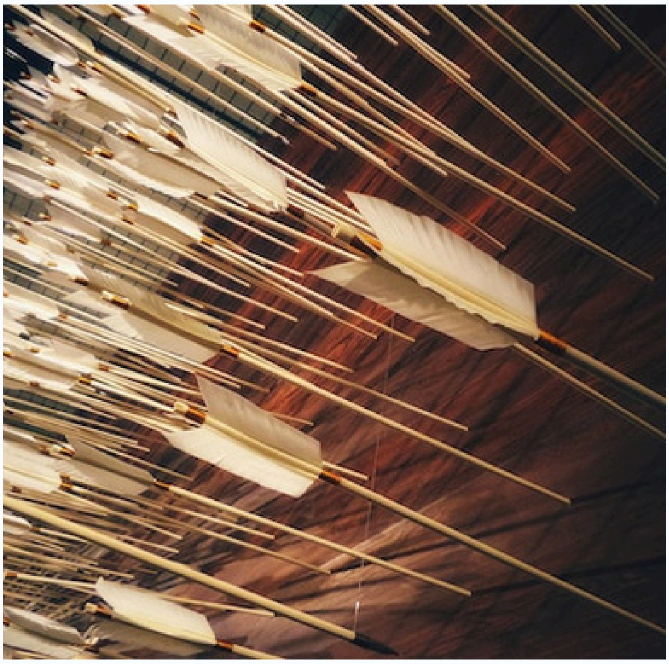
It covers concepts such as Ordinary Language Philosophy, Entity and Entity Types, Predicates, Arity and Reification, and the Principle of Compositionality which serves as a building block of semantic processing. Refer to the post to learn more about the concepts.....

READ MORE

APRIL 5, 2024

VOLUME 2

SEMANTIC PROCESSING CONTINUE....



DISTRIBUTIONAL SEMANTICS: TECHNIQUES TO REPRESENT WORDS AS VECTORS

The distributional hypothesis posits that the context words of an ambiguous word determine its meaning. Techniques like word embeddings capture this context-based meaning. Occurrence matrices and co-occurrence matrices are used to represent words as vectors. LSA (Latent Semantic Analysis) reduces dimensionality, while Word2Vec and GloVe use prediction-based approaches. These methods help understand word semantics and are crucial for tasks like information retrieval and document similarity analysis. Read more about such a technique of representing words as vectors ......

READ MORE

DIVING DEEP INTO TOPIC MODELING: UNDERSTANDING AND APPLYING NLP'S POWERFUL TOOL

Topic Modeling helps decipher the underlying themes in a collection of documents. This post explores how Topic Modeling can reveal latent patterns within textual data. From understanding aboutness to defining topics, we delve into techniques like Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). These models allow us to infer topics being discussed in a given set of documents. Refer to the post to learn about this interesting topic.....

READ MORE