mini_day_nlp

February 7, 2024

Corpus = "Natural language processing (NLP) is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and humans using natural language. It encompasses various techniques such as sentiment analysis, part-of-speech tagging, and named entity recognition. Sentiment analysis aims to determine the attitude or emotional tone of a piece of text, while named entity recognition is a technique that classifies named entities into predefined categories, such as names of persons, organizations, and locations.

Part-of-speech tagging assigns a grammatical category to each word in a sentence, which is crucial for syntactic parsing, a process involving the analysis of the grammatical structure of sentences to understand their meaning. Word embedding is another technique used in NLP, representing words in a continuous vector space, facilitating tasks such as information retrieval and document summarization.

Chatbots are computer programs designed to simulate human conversation using natural language. They leverage techniques like language modeling and text generation to produce human-like responses. Information retrieval, on the other hand, involves obtaining relevant information from a collection of data, often using techniques like topic modeling to discover abstract topics within documents.

Dependency parsing is a technique used to analyze the grammatical structure of sentences based on dependencies between words. It plays a crucial role in tasks such as machine translation and text classification. Word sense disambiguation, another important task, aims to determine the correct meaning of a word based on its context, facilitating accurate language understanding.

Machine translation is the process of automatically translating text or speech from one language to another. This task often relies on techniques such as named entity recognition and part-of-speech tagging to preserve the meaning and grammatical structure of the original text. Language modeling, which involves predicting the next word in a sequence of text given the previous words, is fundamental to many NLP tasks, including machine translation and text generation.

In summary, NLP encompasses a wide range of techniques and tasks, including sentiment analysis, part-of-speech tagging, named entity recognition, language modeling, and machine translation. These techniques play a vital role in enabling computers to understand and generate human language, paving the way for applications such as chatbots, information retrieval systems, and automated translation services."

Sentence 1: "Natural language processing is a subfield of artificial intelligence focused on the interaction between computers and humans using natural language." Sentence 2: "Text mining is the process of deriving high-quality information from unstructured text."

Sentence 1: "Sentiment analysis aims to determine the attitude or emotional tone of a piece of text." Sentence 2: "Named entity recognition is a technique in natural language processing that

aims to classify named entities into predefined categories."

Sentence 1: "Part-of-speech tagging assigns a grammatical category to each word in a sentence." Sentence 2: "Word embedding is a representation of words in a continuous vector space."

Sentence 1: "Chatbots are computer programs designed to simulate human conversation using natural language." Sentence 2: "Information retrieval is the process of obtaining relevant information from a collection of data."

Sentence 1: "Topic modeling is a technique used to discover abstract topics within a collection of documents." Sentence 2: "Word sense disambiguation is the task of determining the correct meaning of a word based on its context."

Sentence 1: "Syntactic parsing involves analyzing the grammatical structure of sentences to understand their meaning." Sentence 2: "Text classification is the process of assigning predefined categories or labels to text documents."

Sentence 1: "Named entity recognition identifies entities such as names of persons, organizations, and locations in text." Sentence 2: "Dependency parsing is a technique used to analyze the grammatical structure of sentences based on dependencies between words."

Sentence 1: "Word sense disambiguation is important for tasks such as machine translation and information retrieval." Sentence 2: "Document summarization involves generating a concise summary of a longer document while preserving its key information."

Sentence 1: "Language modeling involves predicting the next word in a sequence of text given the previous words." Sentence 2: "Text generation is the task of automatically producing human-like text based on a given input or prompt."

Sentence 1: "Named entity recognition is widely used in applications such as information extraction and question answering systems." Sentence 2: "Machine translation is the task of automatically translating text or speech from one language to another."

Task:

The aim of this exercise is to contextualize the NLP concepts that we have covered in the past few days. To achieve this, you will use the above corpus to perform:

- Tokenization
- Stopwords removal
- Lemmatization and/or stemming
- Wordcloud
- TF (Term Frequency)
- IDF (Inverse Document Frequency)
- TF-IDF (Term Frequency-Inverse Document Frequency)
- Similarity between each pair of sentences provided above