Text cleaning and preprocessing : Assignment

Score: 53 / 57

What are the common text cleaning techniques used in natural language processing?

Common text cleaning techniques are stop words removal, punctuation removal, emoji removal, tokenization, slang replacement, abbreviation expansion, stemming and lemmetization

Feedback

Great answer! You have provided a comprehensive list of common text cleaning techniques used in natural language processing.

How does tokenization help in text preprocessing for generative AI models?

Tokenization is the process of breaking down a sentence or text into smaller units called tokens. These small units can be at the word level, sub-word level, or character level. Each gen AI has a specific method of tokenization. These tokens are then embedded as dense vectors for processing through the Gen AI model. We must use the same tokenization that was used during training when we use the LLM model for either fine-tuning or inference.

Feedback

Great answer! The explanation is clear and comprehensive, covering the importance of tokenization in text preprocessing for generative AI models.

Explain the concept of stop words removal and its importance in text cleaning.

Stop words helps to remove common words which do not lend to semantic meaning thereby reducing the token count. Example of stop words are a ,the, is, are

Feedback

Great answer! You have provided a clear explanation of stop words and their importance in text cleaning.

What role does lemmatization play in text preprocessing and why is it important?

Lemmatization helps to reduce a word to its base form. Example going gone and go will all get reduced to go

Feedback

Great answer! Lemmatization is correctly defined and its importance is well explained.

Describe the process of handling special characters and symbols during text cleaning.

Removing special characters with Regular Expressions (Regex) or String manipulation functions - Replacing special characters - Unicode Normalization - Using Text Processing Libraries like NLTK (Natural Language Toolkit) or SpaCy.

Feedback

Great answer! Well explained and covers various methods of handling special characters during text cleaning.

Why is it necessary to perform text normalization in the context of generative AI models?

Text normalization involves transforming text into a consistent and uniform format by applying various rules and techniques. These techniques include:

Case normalization
Punctuation removal
Stop word removal
Stemming
Lemmatization
Tokenization and stop word removal
Parts-of-Speech (POS) taggin

Feedback

Great answer! The response demonstrates a clear understanding of text normalization techniques and their importance in generative AI models.

What are some challenges faced in text cleaning and preprocessing for generative AI?

Many modern text contain emojis and abbreviation. Converting them to meaningful text and context is a challenge. Handling of punctuation, stop words, slang varies from domain to domain. How to handle typos ? At what level segmentation or tokenization is to be done

Feedback

Great answer! You've provided a comprehensive understanding of the challenges in text cleaning and preprocessing for generative AI.

Can you provide an example of how regular expressions can be used for text cleaning?

We can use regular expression to remove HTML tags or links.

Feedback

Great answer! Well done!

Discuss the concept of word embedding and its role in text preprocessing for generative AI models.

Word embedding helps in convert text to dense vectors thereby mapping our text to numerical representations. Example of famour word embeddings are Glove, Word2Vec.

These embeddings are done in way that consider semantic meanings in words

Feedback

The answer provides a good understanding of word embedding and its role in text preprocessing. However, it could be improved by including more specific examples and discussing the impact of word embedding on generative AI models.

What is the difference between stemming and lemmatization in the context of text preprocessing?

Stemming chops off word endings without considering linguistic context, making it computationally faster. Lemmatization analyzes word forms to determine the base or dictionary form, which takes more processing time

Feedback

Great answer! Well explained the difference between stemming and lemmatization in terms of text preprocessing.

How can the handling of misspelled words and typos be addressed during text preprocessing for generative AI models?

Using built in libraries like (NLTK, TextBlob) can help in misspelled words and typeos

Feedback

Using NLTK and TextBlob to address misspelled words and typos is a good approach. However, it would be more comprehensive to mention other techniques such as spell-checking libraries, custom dictionaries, or phonetic matching algorithms. Additionally, providing specific examples of how these libraries can be used would enhance the answer.

Explain the concept of sentence segmentation and its significance in text preprocessing.

Sentence segmentation is the process of determining the longer processing units consisting of one or more words.

Feedback

Great explanation of sentence segmentation and its significance. Well done!