Natural Language Processing

NLP Fundamentals

Definition

NLP, or Natural Language Processing, is a field of artificial intelligence focused on enabling computers to *process* human language. It offsets into multiple applications that leverage computers for language related tasks

Text Processing Stages

There varying numbers of stages when processing text for NLPs, and the existence of some steps are dependent on the solution we are trying to work on.

The more common stages go as follows…

1. Tokenization

Tokenization splits raw text into smaller units (tokens), tokens could be a word, parts of a word, or even a letter.

For example, the text "Cats are running fast." can be converted to the following tokens "Cats", "are", "running", "fast", "."

This step makes data more readable for the various implementations that can follow. Those can be translations, searching, categorization… e.t.c

2. Pre-Processing

Pre-Processing is just a term for normalizing text for easier processing and in some cases to remove unwanted variations, this includes

- Lowercasing
- Removing punctuation, special characters

And more specific processing like

- Removing special patterns like urls
- Expanding contractions (e.g., "don't" → "do not")
- Removing stop words

3. Stemming or Lemmatization

Stemming is basically stripping parts of word possibly front start or end, into its most base format. Implementations use things like suffix removal and linguistic rules on letter composition to perform stripping.

For example a word like “driving” would convert to “drive”

This could output *nonsensical* words.

Continued…

Lemmatization is word normalization like stemming, but it accounts for the *meaning* of a word and it's possible *relevance* to the piece of text.

These implementations often use dictionaries of large texts to look up the normalizations

This method of processing is useful for when the context or meaning of a word matters.

4. Representation or Feature Extraction

Feature extraction is the process of transforming raw, unstructured text data into it's numerical representations or features.

One of the more traditional ways to represent a body of text is the "**Bag-of-Words**" method, a piece of training text is converted into a single set containing just unique words and they are represented in a vector with length of the set *[dcnTxt1, dcnTxt2, ….dcnTxtn]*, then the values are populated according to the method we are working with.

This method neglects word sequences and is prone to dimensional problems as we process larger texts.

Continued…

Non traditional methods include **Word Embedding**, a method of text representation using words that occur together or **Sentence Embedding** these methods try to represent a segment of text instead of a word. These implementations usually use learned embeddings from large bodies of text

These solutions transfer learning, Pre-trained embeddings can be used for more custom solutions.

Challenges

- Bias Amplification and Propagation
- Lack of True Understanding & Common Sense
- Possible Privacy Breaches (Depends on the data-set, and implementation)
- And Computation challenges

Applications

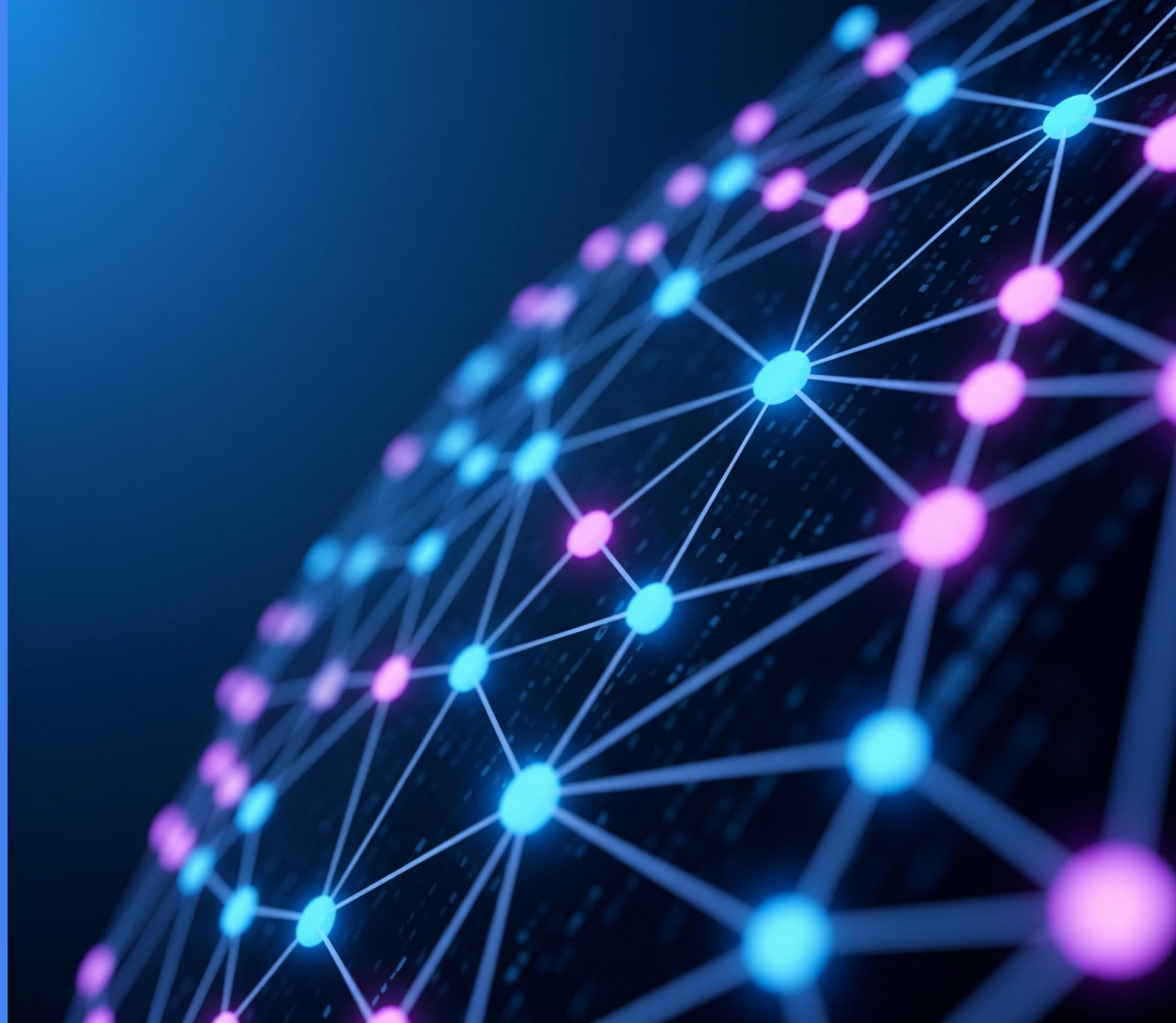
- Content Generation
- Assistance
- Automation (spam identification, highlighting large texts….)
- Big Data Analysis (converting raw internet data into useful information)
- Enhanced Human-Computer Interaction

Final point

NLP = Good For Humans

Thanks You!!

Questions?



For Additional Info

Ask our AI friends

Gemini

ChatGPT

Grok