Modern NLP

Based on Deep Learning and Language models.
Day 1 Afternoon

Alexandre Gazagnes

1st Day

1. Morning (~ 2h)
    - ML – IA – NLP – A gentle introduction
    - Python, Anaconda, Juypter, Git, Github : All the tools you need
    - Python 101 – A very small remainder

2. Afternoon (~ 4h)
    - ML 101 – A very first implementation
    - Text Processing – From words to vectors
    - Basic and Advanced NLP techniques

ML - 101

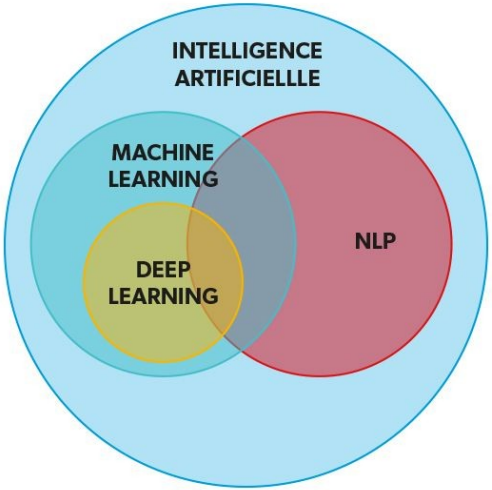
Centrale
CASABLANCA

Practice !

Introduction to NLP

Definition and overview of NLP

NLP, or natural language processing, is a field of computer science and artificial intelligence that deals with the **interaction between computers and human languages**. It involves using techniques from linguistics, computer science to **process, analyze, and understand human language**.



Some common applications of NLP include language translation, text classification, sentiment analysis, and chatbot development.

Deep learning is a technique used in NLP to process and understand human language.

Applications of NLP

NLP has a wide range of applications in various industries, including:

- **Healthcare**: NLP can be used to extract information from medical records, identify trends and patterns in patient data, and assist with diagnoses and treatment planning.
- **Finance**: NLP can be used to analyze financial news articles and social media posts to predict stock market trends or detect fraudulent activity.
- **Marketing**: NLP can be used to analyze customer reviews and social media posts to understand customer sentiment and preferences, and to target advertising and marketing efforts.
- **Customer service**: NLP can be used to develop chatbots and virtual assistants to assist customers with inquiries and support.

Usage in the food industry

Here are some examples:

- **Recipe generation:** Generate new recipes based on a given set of ingredients or dietary restrictions.
- **Nutritional analysis:** Extract nutritional information from food labels or recipes, making it easier for consumers to track their intake or for companies to comply with regulatory requirements.
- **Food safety monitoring**: Analyze social media posts or news articles for mentions of food safety incidents or outbreaks, allowing companies to identify and address potential issues more quickly.

Introduction to Python and its role in NLP

Python is a popular programming language that is widely used in the field of NLP due to its simplicity and ease of use.

There are many libraries and frameworks available in Python that make it easy to implement NLP projects, such as NLTK, SpaCy, Gensim or HuggingFace.

Python also has a large community of developers and researchers who contribute to the development of NLP tools and techniques.

Practice !

Alexandre Gazagnes

Text Preprocessing

Alexandre Gazagnes

Tokenization

Tokenization is the process of breaking down a piece of text into smaller units, or tokens, that can be more easily analyzed and processed.

Tokens can be for example: words, sentences, paragraphs. The choice of token type depends on the specific task.

The tokenization process is a pre-processing step to convert the raw text into smaller unit for the main task of NLP. Each unit will be encoded as a numeric vector (**Embedding**).

There are many techniques for tokenization: split function, regular expressions, NLTK library …

Tokenization - examples

If the task is at the **word level,** such as in **part-of-speech** tagging or **named entity recognition**, the token type is typically set to be individual words. In these cases, the model is able to analyze the grammatical function of individual words in a sentence, and make predictions about the roles they play in the sentence.

If the task is at the **sentence level**, such as in sentiment analysis or machine **translation**, the token type is typically set to be complete sentences. In these cases, the model can analyze the meaning and context of a full sentence and make predictions based on that.

In some task that require more context, such as text **summarization** or automatic text generation, token type can be set as a paragraphs or even as a whole document. These model are able to analyze and generate context based on the relationship between different sentences.

Stemming and lemmatization:

Definition: the process of reducing a word to its base form, or stem, in order to better analyze the meaning and context of the word

Examples: *running* -> *run*, dogs -> dog, thought -> think

It helps to reduce the complexity of text analysis by reducing the number of unique words that need to be processed

Stop word removal:

Definition: the process of removing common, non-meaningful words from a piece of text in order to focus on the more important content.

Examples: the, a, an, is, are, was

Importance: helps to reduce the complexity of text analysis by removing noise from the text

Techniques: NLTK library, custom stop word list

Part-of-speech tagging:

Definition: the process of identifying the part of speech of each word in a piece of text, such as noun, verb, adjective, etc.

Importance: helps to better understand the structure and meaning of a piece of text.

The full name of the part-of-speech tags used in the Tagged Tokens output are:

DT: Determiner, JJ: Adjective, NN: Noun, singular or mas, NNS: Noun, plural, IN: Preposition or subordinating conjunction

Techniques: NLTK library, rule-based systems, machine learning algorithms

Embedding

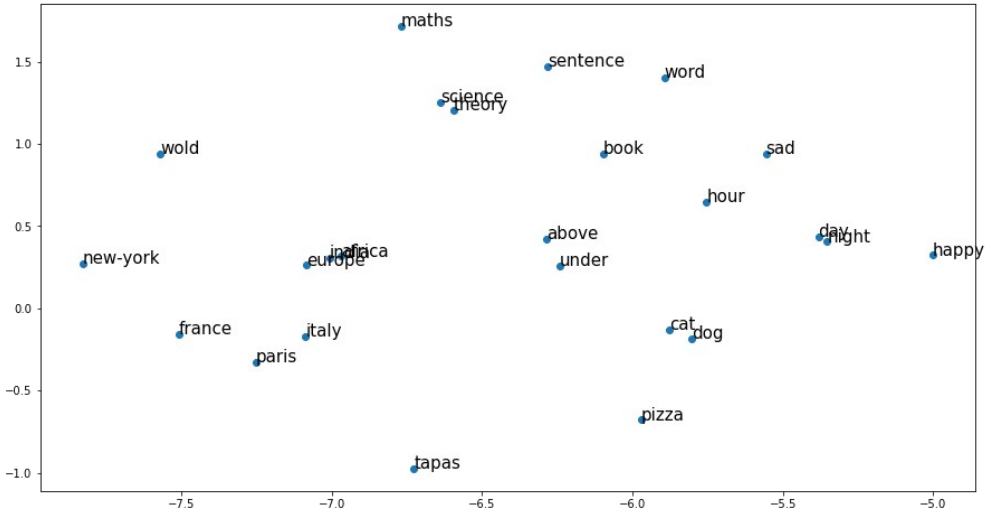
Embedding is a technique used in natural language processing (NLP) to represent words, phrases, or documents as dense, low-dimensional vectors in a continuous vector space. These vectors capture the semantic meaning of the words and are useful for a wide range of NLP tasks, such as text classification, language translation, and information retrieval.

Embedding methods can be broadly divided into two categories: frequency-based methods, such as word count or term frequency-inverse document frequency (TF-IDF), and prediction-based methods, such as word2vec and GloVe. The latter methods learn the embeddings by training a model to predict the surrounding context of a word, while the former methods rely on the statistics of word co-occurrence in a corpus of text.

Therefore, the term embedding is a technique used in the field of text processing and more specifically in the NLP area to represent natural language in a numerical format that machine learning models can work with.

Word2vec

Word2Vec is a technique for creating dense, numerical representations of words, also called **"word embeddings."** These embeddings capture the meaning and context of a word in a continuous, multi-dimensional space. Word2Vec is trained using **neural networks** on large corpora of text.



Practice !

Alexandre Gazagnes