Modern NLP

Based on Deep Learning and Language models.

Jawad ALAOUI

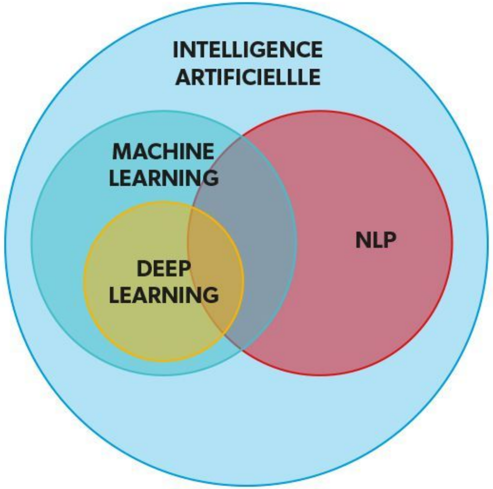
Agenda

1. Introduction to NLP
2. Text Preprocessing
3. NLP Tasks with pretrained models
4. Training or fine-tuning a model
5. NLP in Practice: Project Work

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.

Hugging Face

GENSIM
topic modelling for humans

NLTK

spaCy

Practice - Colab

Text Preprocessing

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

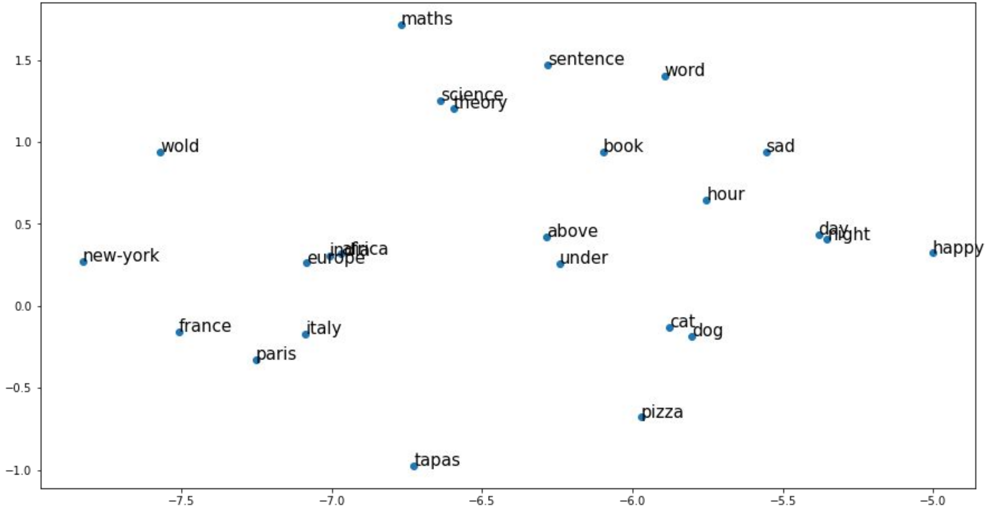
Embedding is a technique used in natural language processing (NLP) to represent words, phrases, or documents as dense, low-dimensional vectors in a continuous vector space. These vectors capture the semantic meaning of the words and are useful for a wide range of NLP tasks, such as text classification,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

NLP Tasks with pretrained models

What is a pretrained model ?

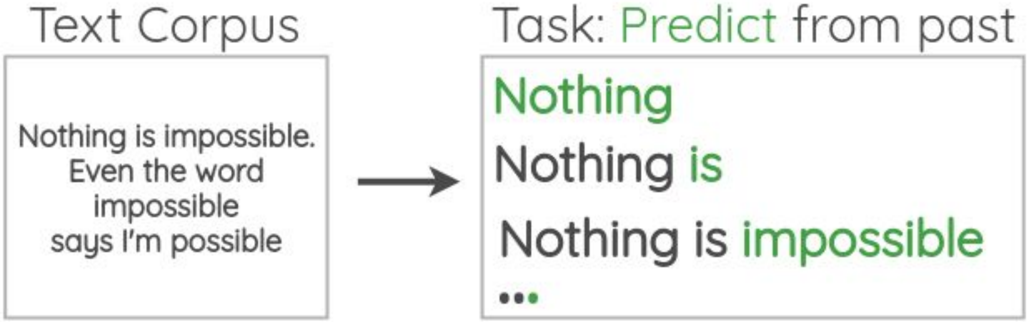
A **pretrained** model is a neural network that has been trained on a large dataset and can be used directly on your data or as a starting point for other tasks.

These models can s**ave a significant amount of time and computational resources,** as it is not necessary to train a model from scratch on a new task.

Finetuning is the process of further training a pretrained model on a specific task using task-specific data (usually small dataset). The degree of finetuning varies depending on the size and similarity of the new task to the pretraining task.

Large language models case

LLM model is presented with a large dataset of text and is then trained **to predict the next word** in a sequence of words (**Self-supervised learning**).



Large language models have achieved **state-of-the-art performance** on a wide range of natural language processing tasks, including language translation, text summarization, and question answering.

Most used pretrained models in NLP

BERT: Developed by **Google**, it is a transformer-based model with over **110 million parameters,** trained on a dataset of over 3 billion words (**16G** of text). Used in Gmail for email auto-completion

RoBERTa: Developed by Facebook AI, it is a transformer-based model with over **125 million parameters**, trained on a dataset of over **160GB** of text data. Used in facebook for Content Moderation.

GPT3: Developed by **OpenAI**, it is a transformer-based model with over **175 billion parameters**, trained on a dataset of petabytes of data collected over 8 years of web crawling. GPT4 has almost one trillion parameters.

Classification & Sentiment Analysis

The process of d**etermining the category or a given text**. Sentiment Analysis is one of the most used case.

Examples:

- Determining whether a customer review of a product is positive or negative.
- Classifying a news article as belonging to a specific category (politics, sports, entertainment, etc.).

Zero-Shot Classification

Zero-shot classification is an NLP algorithm that can classify text data into a set of predefined categories or labels without the need for any labeled training data. Instead, the algorithm relies on natural language understanding and general knowledge to make predictions based on the input text and a set of given parameters.

Examples:

- Classifying news articles into predefined topics such as politics, sports, and entertainment, without being trained on any specific labeled data for those categories. The algorithm relies on its ability to understand the language and the context of the articles to make accurate predictions.
- Classifying customer feedback into different categories such as positive, negative, or neutral sentiments, without any labeled training data for sentiment analysis. The algorithm can use its understanding of language and common patterns in text to identify the sentiment expressed in the feedback.
- Classifying documents or texts into different languages, without any labeled data for language classification. The algorithm can analyze the input text's linguistic features and identify the most likely language based on its knowledge of language structure and patterns.
- Classifying products or services into predefined categories based on their features and characteristics, without any labeled data for product classification. The algorithm can analyze the text describing the product's features and characteristics and classify it into the most appropriate category based on its understanding of the product domain.

Information Extraction & Questing Answering

The process of extracting **specific information from a text** and answering questions based on that information.

Examples:

- Extracting information about a person from a Wikipedia article and answering questions about their education and career.
- Extracting information about a company from a financial report and answering questions about their revenue and profits.

Translation

The process of converting text from one language to another.

Example:

- Translating a Spanish news article into English.
- These models can also be adopted to generate voice : Text-to-Speech

Summarization

The process of **condensing a text** to its main points or **key information**.

Examples:

- Summarizing a long research paper into a brief summary for a conference presentation.
- Summarizing a news article about a complex legal case into a summary for a non-legal audience.

Topic Modeling

The goal of topic modeling is to uncover the underlying structure and patterns in the data, and to group together documents that are related to each other based on their content.

Topic modeling algorithms typically use techniques such as latent Dirichlet allocation (LDA) to identify the underlying topics present in a set of documents. These topics may be represented as a set of keywords or phrases, and each document may be assigned a probability distribution over the topics, indicating the degree to which it relates to each topic

Examples:

- A news organization may use topic modeling to identify the most common themes and issues discussed in their articles, and to group similar articles together for easier organization and analysis.
- A social media monitoring tool may use topic modeling to identify the most frequently discussed topics and hashtags on a particular platform, and to track changes in these topics over time.

Text Generation & Prompting

The process of creating new text based on a given input or set of parameters.

Prompting refers to the process of providing input, context and/or starting point for a text generation.

Examples:

- Generating personalized email responses based on the recipient's name and previous interactions with the company.
- Generating product descriptions for an e-commerce website based on the product's features and target audience.
- Generating creative headlines for news articles based on the content of the article.
- Generating social media posts for a brand based on current events and trending topics.

Evaluation

There are several common evaluation metrics used in natural language processing (NLP) tasks, depending on the specific task and dataset. Some of the most common metrics include:

- **Accuracy**: This is the most basic and commonly used metric, which simply calculates the percentage of correctly predicted labels or outputs.
- **Precision, Recall and F1-score**: These are used for classification tasks and take into account both true positives and false positives. Precision is the number of true positives divided by the number of true positives and false positives, recall is the number of true positives divided by the number of true positives and false negatives, and F1-score is the harmonic mean of precision and recall.
- **BLEU, ROUGE and METEOR:** These are used for text generation tasks such as machine translation, summarization, and text completion. **BLEU** compares the generated text to one or more reference texts, **ROUGE** compares the generated text to a reference text, and **METEOR** compares the generated text to a reference text, taking into account synonyms and stemming.

It is important to note that these metrics should be chosen based on the specific NLP task, and that a single metric may not provide a complete picture of the model's performance.

Practice - Colab

Training or fine-tuning a model

What is model Fine-tuning ?

Model fine-tuning refers to the process of **adapting a pre-trained model to a specific task or dataset** by training it further on a new set of data.

Examples:

- Fine-tuning on a dataset of customer reviews to improve ability to classify the sentiment of the reviews.
- Fine-tuning on a dataset of legal documents to improve ability to extract specific information and answer legal-related questions.

Fine-tuning is not only applied to NLP, all deep learning algorithms can be fine tuned. For example in computer vision, a pre-trained image classification model can be fine-tuned on a dataset of medical images to improve its ability to identify specific diseases in the images.

Practice - Colab

Annexes

NLP in Practice: Project Work

Conduct a study about trends in healthy food in 2022. This project could involve the following steps:

- Collecting data on healthy food trends by scraping relevant information from websites, blogs, and social media platforms.
- Preprocessing the collected data by cleaning, normalizing, and tokenizing the text
- Using NLP techniques such as sentiment analysis, topic modeling, and named entity recognition to analyze the data and uncover trends in healthy food.
- Visualizing the results using charts and graphs to make the findings more clear and actionable.
- Creating a report or presentation summarizing the findings and discussing their implications for the healthy food industry.

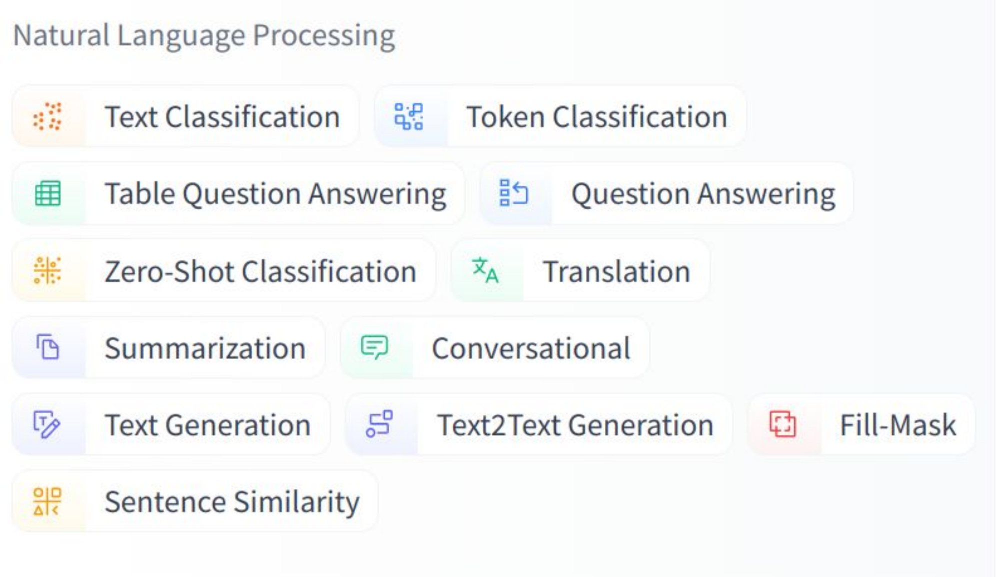
This project would allow students to apply their NLP skills to a real-world problem and gain a better understanding of trends in healthy food. Additionally, it could be possible to share the findings with relevant stakeholders in the food industry and potentially have an impact.

Collecting data on healthy food trends

One example of where you could collect data for this NLP project on healthy food trends in 2022 would be from social media platforms such as Twitter, Instagram, and Facebook. You could use the social media APIs to access data on posts and comments related to healthy food, such as those containing keywords like "organic," "plant-based," "gluten-free," etc. You could also use hashtags such as #healthyeating or #plantbaseddiet to find relevant content.

Another example would be scraping websites that specialize in healthy food recipes and nutrition information, such as Whole Foods, EatingWell, and Healthline. These websites contain a lot of useful information on healthy food trends, such as recipes, ingredients, and nutritional information.

HuggingFace NLP Tasks classification



Natural Language Processing

Text Classification

Token Classification

Table Question Answering

Question Answering

Zero-Shot Classification

Translation

Summarization

Conversational

Text Generation

Text2Text Generation

Fill-Mask

Sentence Similarity

https://github.com/wandb/examples

https://wandb.ai/ayush-thakur/huggingface/reports/How-To-Fine-Tune-HuggingFace-Transformers-on-a-Custom-Dataset--Vmlldzo0MzQ2MDc

Annexes

To find dataset you can start by exploiting:

https://datasetsearch.research.google.com/

https://huggingface.co/datasets/viewer/