Вступ до Natural Language Processing

NLP - гаряча тема

Обробка природної мови (NLP) є однією з найгарячіших тем у галузі штучного інтелекту зараз, завдяки застосуванням, як-от генератори тексту, чат-боти та програми для перетворення тексту в зображення.

Останніми роками відбулася революція в можливостях комп'ютерів розуміти людські мови, програмні мови та навіть біологічні й хімічні послідовності, як-от структури ДНК і білків, що нагадують мову.

Останні моделі AI відкривають нові можливості для аналізу сенсів введених текстів і генерування осмислених, виразних результатів природньою мовою.

Natural language processing
Topic

+ Compare

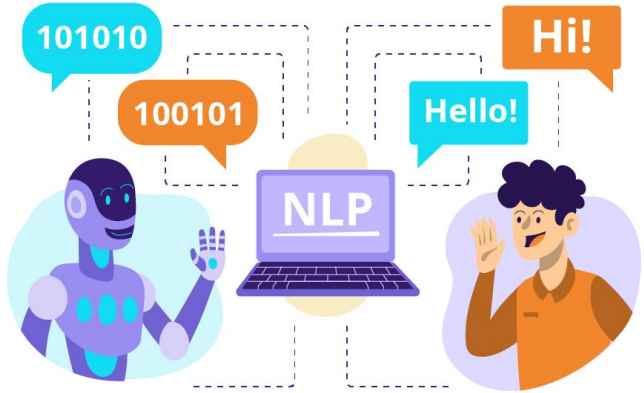
Worldwide | 2004 - present | All categories | Web Search

Interest over time

Що таке NLP

NLP – це дисципліна, яка спрямована на створення машин, здатних маніпулювати людською мовою або даними, що нагадують людську мову, так, як це написано, вимовлено та організовано.

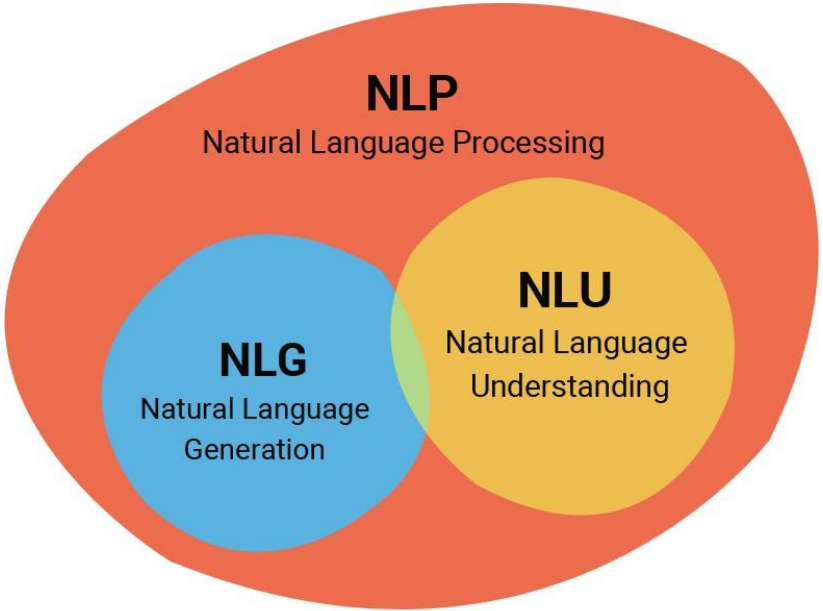
Еволюціонувавши з комп'ютерної лінгвістики, NLP зосереджується на розробці технологій для виконання корисних завдань.



Напрями в NLP

NLP можна розділити на два піднапрями:

- **Розуміння природної мови (NLU):** Фокус на семантичному аналізі або визначенні значення тексту.
- **Генерація природної мови (NLG):** Фокус на генерації тексту машиною.



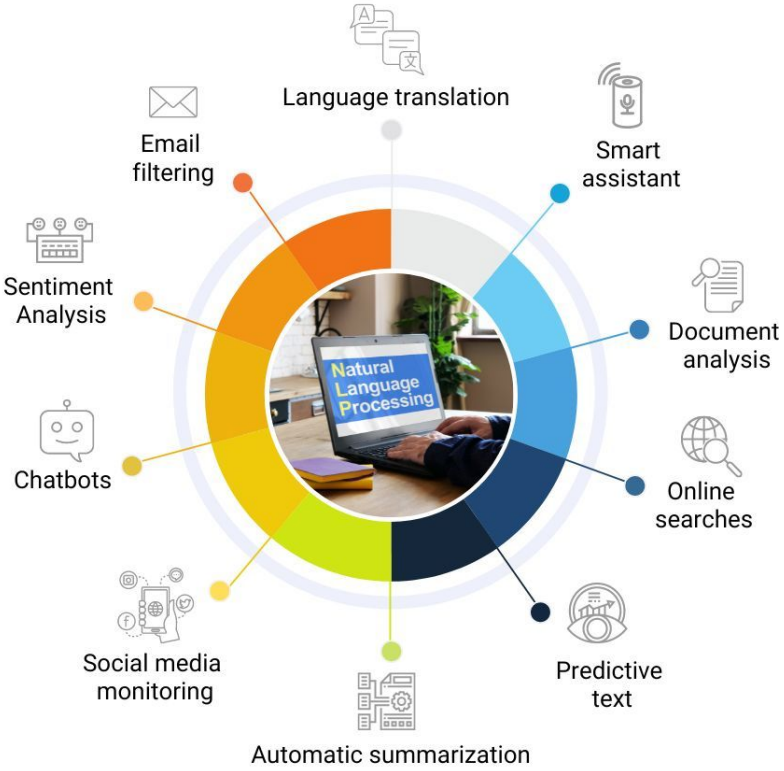
Важливість NLP в сучасному світі

NLP є невід'ємною частиною повсякденного життя: від роздрібної торгівлі (наприклад, чат-боти для обслуговування клієнтів) до медицини (інтерпретація або резюмування електронних медичних записів).

Приклади використання:

- Конверсійні агенти, як-от Amazon Alexa та Apple Siri, використовують NLP для обробки запитів користувачів.
- Моделі, як GPT-3, можуть генерувати складні тексти та підтримувати зв'язну розмову.
- NLP допомагає покращити результати пошуку Google та виявляти ненависну мову в соціальних мережах.

Застосувань NLP зараз дууже багато



Всі ці застосування досягаються розв’язанням кількох NLP задач

https://medium.com/nlplanet/two-minutes-nlp-33-important-nlp-tasks-explained-71e2eaed2b1b

Як працює NLP?

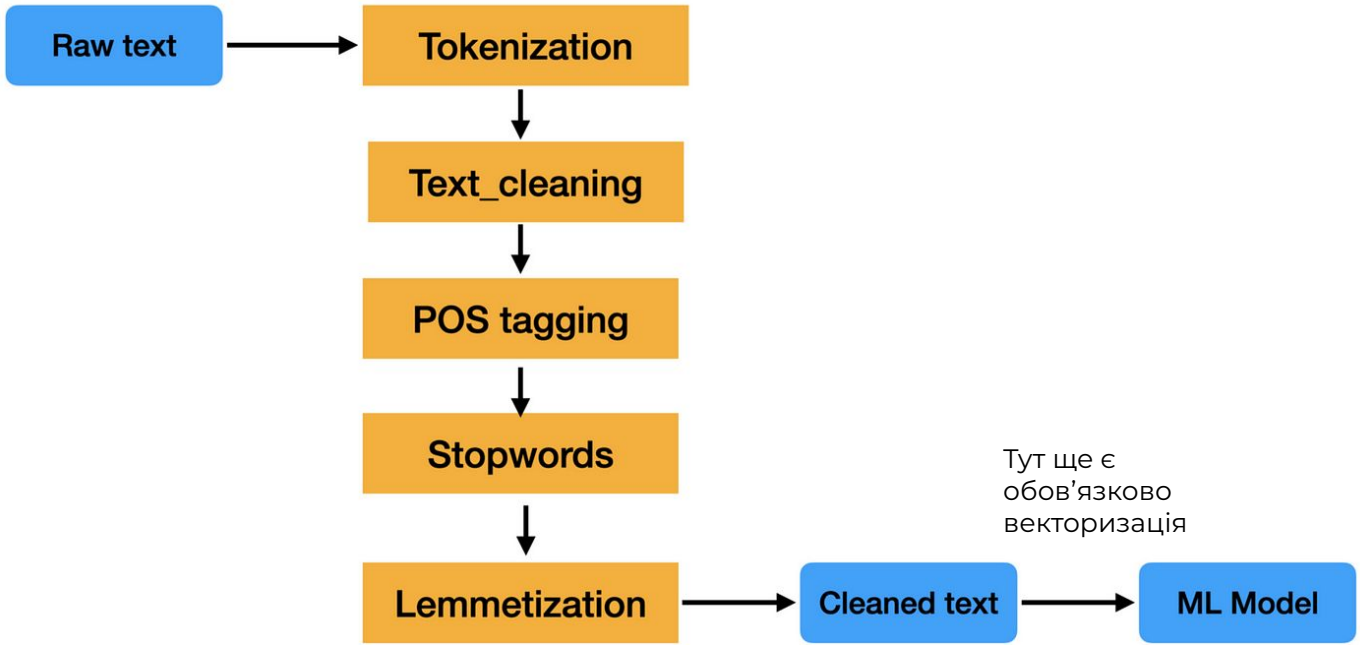
NLP моделі працюють, знаходячи зв'язки між складовими частинами мови — буквами, словами, реченнями у текстових наборах даних.

Основні етапи процесу NLP:

- **Попередня обробка даних:** Підготовка тексту для покращення продуктивності моделей.
- **Виділення ознак:** Перетворення тексту в числові дані для моделювання.
- **Моделювання:** Використання архітектур NLP для виконання різноманітних завдань.

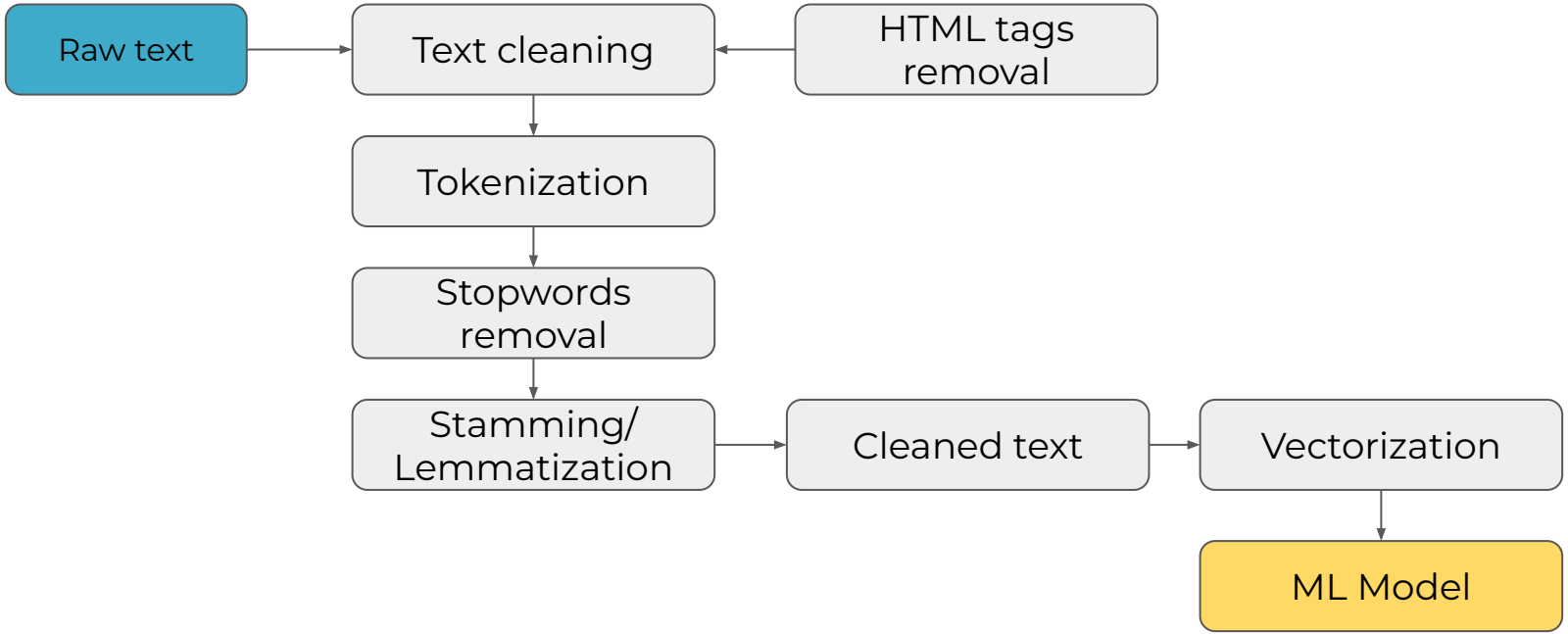
Стандартний процес розв’язку NLP задачі

Ця діаграма не вичерпна, але це — набір мінімум.



Процес розв’язку NLP задачі з класифікації тексту

Ця діаграма не вичерпна, це — набір мінімум.



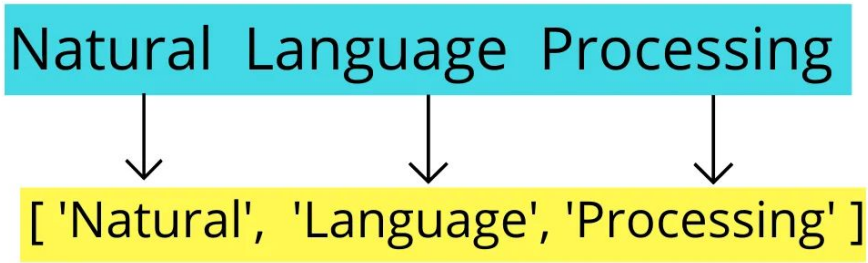
Етапи попередньої обробки тексту

Tokenization

Токенізація – це процес розбиття тексту на менші одиниці, звані токенами. Токени можуть бути окремими словами, фразами або навіть символами.

Мета токенізації – перетворити текст у структуру, яка зручна для подальшого аналізу і обробки моделями NLP. Наприклад, речення "NLP is fun!" може бути токенізовано на ["NLP", "is", "fun", "!"].

Токенізація дозволяє враховувати кожне слово або символ як окрему одиницю для аналізу.



Tokenization

Приклад найпростіших токенізацій.

Токенізації можуть робитись по-різному і бути специфічними до кожної NLP-моделі. Більше прикладів токенізації:

https://medium.com/@ajay_khanna/tokenization-techniques-in-natural-language-processing-67bb22088c75

Tokenization Using Python's split() function

```
text = """Natural language processing (NLP) is a subfield of
linguistics, computer science, and artificial intelligence concerned
with the interactions between computers and human language, in
particular how to program computers to process and analyze large
amounts of natural language data."""

# Splits at space

text.split()

['Natural','language','processing','(NLP)','is','a','subfield','of',
'linguistics,','compute','science,','and','artificial','intelligence',
'concerned','with','the','interactions','between','computers','and',
'human','language,','in','particular','how','to','program','computers'
,'to','process','and','analyze','large','amounts','of','natural','lan
guage','data.]
```

Tokenization Using Regular Expressions (RegEx)

```
import re

text = """Natural language processing (NLP) is a subfield of
linguistics, computer science, and artificial intelligence concerned
with the interactions between computers and human language,in
particular how to program computers to process and analyze large
amounts of natural language data. """

tokens = re.findall("[\w']+", text)

Output – ['Natural', 'language', 'processing', 'NLP', 'is', 'a',
'subfield', 'of', 'linguistics', 'computer', 'science', 'and',
'artificial', 'intelligence', 'concerned', 'with', 'the',
'interactions', 'between', 'computers', 'and', 'human', 'language',
'in', 'particular', 'how', 'to', 'program', 'computers', 'to',
'process', 'and', 'analyze', 'large', 'amounts', 'of', 'natural',
'language', 'data']
```

Чистка тексту

На цьому етапі видаляються слова та елементи з масиву текстових даних, щоб підвищити ефективність моделі машинного навчання. З текстових даних будуть видалені цифри, великі літери, розділові знаки, стоп-слова, одинарні лапки. Процес очищення тексту відбувається за допомогою регулярного виразу.

Стоп-слова

Стоп-слова — це загальновживані слова, які не мають суттєвого смислового навантаження і часто не несуть важливої інформації для розуміння змісту тексту. Приклади стоп-слів: "і", "або", "але", "що", "якщо", "на", "у", "з", "до", "від", "я", "він", "вона" тощо.

У задачах обробки природної мови (NLP) стоп-слова зазвичай видаляються з тексту, щоб зменшити розмір даних і підвищити ефективність моделей. Видалення стоп-слів дозволяє зосередитися на більш важливих, змістовних термінах, що допомагають краще розпізнавати патерни в тексті для різних NLP завдань, таких як класифікація тексту, аналіз настроїв або пошук інформації.

Lemmatization and stemming

Стемінг – це процес скорочення слова до його кореня або основної форми шляхом відкидання суфіксів чи префіксів. Наприклад, слова "running", "runner", "ran" скорочуються до "run". Стемінг є менш точним методом, оскільки він використовує евристичні правила для скорочення слів, що може призвести до отримання неіснуючих слів (наприклад, "univers" від "universe" та "university").

Лематизація – це процес перетворення слова у його базову або лемматичну форму, зважаючи на його морфологію та контекст у реченні. Наприклад, для слова "better" лематизація визначить основну форму як "good". Лематизація є більш формальною та точною, ніж стемінг, оскільки враховує граматичні правила та використання словника.

Всі ці застосування досягаються вирішеням кількох NLP задач

Кожне із застосувань NLP (аналіз настроїв, машинний переклад, чат-боти, резюмування тощо) досягається шляхом вирішення різних NLP задач:

- **Класифікація:** Sentiment Analysis, класифікація токсичності, виявлення спаму, визначення іменованих сутностей (Named Entity Recognition). Моделі класифікації використовуються для класифікації текстів або слів на різні категорії на основі їх змісту.
- **Переклад та генерація:** Машинний переклад, генерація тексту, автозаповнення. Ці задачі зосереджені на перетворенні тексту з однієї мови на іншу або на створенні нового тексту на основі заданого контексту. Може бути також продовження тексту за заданим початком.
- **Information Retrieval:** Пошук документів, що найкраще відповідають запиту, пошук подібних текстів, пошук фрагментів тексту які найкраще висвітлюють сенс тексту, відповіді на запитання (Question Answering). Задачі фокусуються на отриманні релевантної інформації з великих обсягів текстових даних.
- **Аналіз тексту:** Моделювання тем (Topic Modeling), корекція граматичних помилок (Grammatical Error Correction). Ці задачі пов'язані з аналізом текстових даних для виділення важливої інформації або вдосконалення структури тексту.

Named Entity Recognition