**🔹 Giới thiệu học phần**

* **Thời lượng:** 30 tiết lý thuyết + 30 tiết thực hành.
* **Mục tiêu:** Trang bị kiến thức nền tảng và kỹ thuật xử lý ngôn ngữ tự nhiên (NLP).
* **Đánh giá:** Giữa kỳ và cuối kỳ.

**📘** Chương 1: Tổng quan về Xử lý Ngôn ngữ Tự nhiên

**✅ Tóm tắt nội dung chính Chương 1–2: Xử lý ngôn ngữ tự nhiên (XLNNTN)**

**1. Ngôn ngữ tự nhiên là gì?**

* Ngôn ngữ được con người phát triển, sử dụng để giao tiếp.
* Dữ liệu văn bản là dạng phi cấu trúc, chứa cú pháp và ngữ nghĩa.

**2. Xử lý ngôn ngữ tự nhiên (NLP)** **Natural Language Processing**

* Một nhánh của **trí tuệ nhân tạo (AI)**.
* Cho phép máy tính hiểu, phân tích và tạo ra ngôn ngữ giống con người.
* Tập trung vào việc phân tích văn bản và lời nói.

**3. Thách thức trong NLP**

* Ngôn ngữ **đa nghĩa**, **có cấu trúc phức tạp**, **đệ quy**, và **ẩn chứa ngữ cảnh**.

**4. Ứng dụng của NLP**

* **Chatbot**, **dịch máy**, **phân tích cảm xúc**, **kiểm tra chính tả**, **tóm tắt**, **truy vấn thông tin**, **nhận dạng giọng nói**, **OCR**, v.v.
* Nền tảng cho **AI tạo sinh** như ChatGPT, GitHub Copilot, DALL-E...

**📘 Chương 2: Kiến thức cơ bản về ngôn ngữ học & Kỹ thuật NLP**

**1. Các cấp độ phân tích ngôn ngữ**

| **Cấp độ** | **Mô tả** |
| --- | --- |
| **Phonology** (Âm vị học) | Nghiên cứu âm thanh trong ngôn ngữ |
| **Morphology** (Hình thái học) | Cấu trúc từ, hình vị (gốc từ, tiền tố, hậu tố) |
| **Syntax** (Cú pháp) | Quy tắc sắp xếp từ thành câu |
| **Semantics** (Ngữ nghĩa) | Ý nghĩa của từ/câu |
| **Pragmatics** (Ngữ dụng) | Ngữ cảnh và hành vi giao tiếp |

**2. Một số khái niệm chính**

* **Từ loại (Part-of-Speech):** danh từ, động từ, tính từ,...
  + **Lớp mở**: danh từ, động từ, tính từ...
  + **Lớp đóng**: giới từ, liên từ, mạo từ...
* **Cấu trúc cú pháp:**
  + **Cấu trúc cụm từ (Phrase Structure)**.
  + **Cây phụ thuộc (Dependency Tree)**.
* **Ngữ nghĩa từ vựng:** Nghĩa nội tại và khả năng kết hợp của từ.
* **Ngữ dụng:** Phân tích mục đích và ngữ cảnh của hành vi ngôn ngữ.

**🔧 Kỹ thuật và Tác vụ NLP phổ biến**

| **Tác vụ** | **Mô tả** |
| --- | --- |
| **Sentence Splitting** | Chia văn bản thành các câu |
| **POS Tagging** | Gán nhãn từ loại |
| **Parsing** | Phân tích cấu trúc câu |
| **Named Entity Recognition** | Nhận diện thực thể như tên người, tổ chức |
| **Word Sense Disambiguation** | Hiểu đúng nghĩa của từ theo ngữ cảnh |
| **Semantic Role Labeling** | Gán vai trò ngữ nghĩa (chủ ngữ, vị ngữ, đối tượng...) |

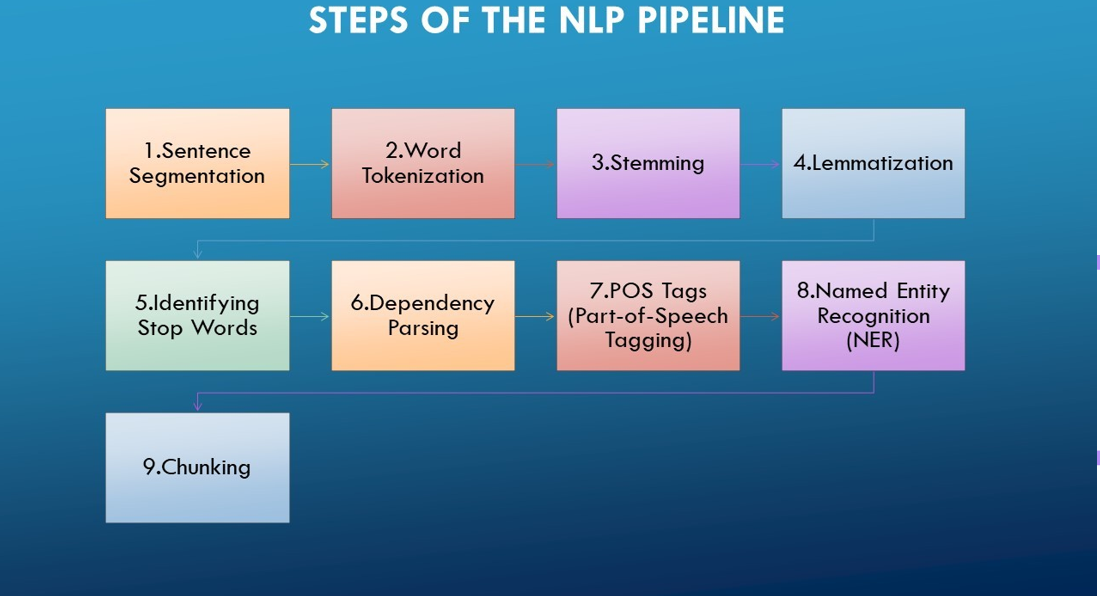
**✅ Tóm gọn**

* **Chương 1**: Giới thiệu NLP và ứng dụng.
* **Chương 2**: Cơ sở ngôn ngữ học và các kỹ thuật nền tảng trong NLP.

# Chương 2: Tiền xử lý dữ liệu trong Xử lý Ngôn ngữ Tự nhiên (NLP)

**🔹 1. Tổng quan quy trình NLP**

* Trước khi thực hiện các tác vụ NLP như phân loại, tóm tắt hay dịch máy, **văn bản phải được chuẩn hóa và xử lý ban đầu**.
* Một quy trình điển hình gồm:  
  **Phân đoạn → Chuẩn hóa → Trích xuất đặc trưng → Ra quyết định**



**🔹 2. Các bước tiền xử lý văn bản**

**📌 a. Phân đoạn câu (Sentence Segmentation)**

* Chia văn bản thành **các câu riêng lẻ**.
* Sử dụng các dấu câu như dấu chấm (.), chấm hỏi (?), chấm than (!).
* Mục tiêu: **hiểu ngữ cảnh tốt hơn và dễ xử lý từng câu**.

**📌 b. Tokenization (Tách từ)**

* Chia câu thành **các đơn vị nhỏ hơn (token)**.
* Có 3 dạng:

| **Loại Tokenization** | **Ví dụ token** | **Ưu điểm** | **Nhược điểm** |
| --- | --- | --- | --- |
| Word | ['love'] | Dễ hiểu, phổ biến | Không xử lý tốt từ mới/lạ |
| Character | ['l','o'] | Tốt cho lỗi chính tả/OCR | Mất ngữ nghĩa của từ |
| Subword | ['happi'] | Tối ưu, giảm từ ngoài từ điển | Cần huấn luyện/phức tạp hơn |

| **Loại Tokenization** | **Đặc điểm chính** |
| --- | --- |
| **Word** | Tách theo khoảng trắng, dấu câu |
| **Character** | Tách từng ký tự (phù hợp với tiếng Trung, lỗi chính tả, v.v.) |
| **Subword** | Cắt từ hiếm thành phần nhỏ, giữ nguyên từ phổ biến (GPT/BERT dùng) |

**🔹 3. Chuẩn hóa token (Token Normalization)**

* Đưa từ về dạng chuẩn/tối giản để **tăng độ khớp và giảm kích thước từ vựng**.
* Ví dụ:
  + **Xóa dấu chấm, dấu gạch ngang**
  + Chuẩn hóa chính tả tiếng Anh–Anh vs. Anh–Mỹ (colour ↔ color)

**🔹 4. Stemming và Lemmatization**

**✂️ Stemming**

* Cắt đuôi từ để tìm "gốc từ" (dù không phải từ thật)
* Ví dụ: running → run, connected → connect
* Hai vấn đề thường gặp:
  + **Over-stemming**: cắt quá nhiều → mất nghĩa
  + **Under-stemming**: cắt quá ít → chưa tối ưu
* Các bộ stemming phổ biến:
  + **Porter Stemmer (1980)** – Dựa vào quy tắc
  + **Krovetz Stemmer (1993)** – Kết hợp từ điển + quy tắc

**🧠 Lemmatization**

* Đưa từ về **dạng gốc có nghĩa (lemma)**, có xét **ngữ cảnh và từ loại**
* Ví dụ:
  + am, are, is → be
  + cars, car’s → car

✅ **So sánh**:

|  | **Stemming** | **Lemmatization** |
| --- | --- | --- |
| Dễ thực hiện | ✅ | ❌ (phức tạp hơn) |
| Xét ngữ cảnh | ❌ | ✅ |
| Cho từ có nghĩa thật | ❌ (không đảm bảo) | ✅ |

**🔹 5. Stop Words (Từ dừng)**

* Là các từ phổ biến **không mang nhiều thông tin ngữ nghĩa** (vd: "the", "of", "and", "to"...)
* Thường được loại bỏ để:
  + Giảm kích thước dữ liệu
  + Tăng hiệu quả xử lý
* Có thể tùy chỉnh danh sách stop words theo **lĩnh vực cụ thể**

**✅ Tóm tắt nhanh**

| **Kỹ thuật** | **Chức năng chính** |
| --- | --- |
| Sentence Segmentation | Chia văn bản thành câu |
| Tokenization | Chia câu thành từ/ký tự/subword |
| Normalization | Chuẩn hóa từ (viết hoa, dấu, chính tả...) |
| Stemming | Cắt đuôi từ để tìm gốc |
| Lemmatization | Đưa từ về dạng gốc có nghĩa |
| Stop Words Removal | Loại bỏ từ không mang nhiều thông tin ngữ nghĩa |