



CHỦ ĐỀ: NLP

# ĐỒ ÁN: CHUYỂN ĐỔI TỪ TIẾNG VIỆT ĐỊA PHƯƠNG SANG TIẾNG VIỆT CHUẨN 20CNTTHUC1

20127674 - LÊ ĐỨC ĐẠT



Ngôn ngữ → Qôn qữ  
Giáo dục → Záo zuk  
Tiếng nói → Tiếq nói  
Chữ viết → Cũ viết

# NỘI DUNG TRÌNH BÀY

VẤN ĐỀ

GIỚI THIỆU

Ý TƯỞNG

# VẤN ĐỀ

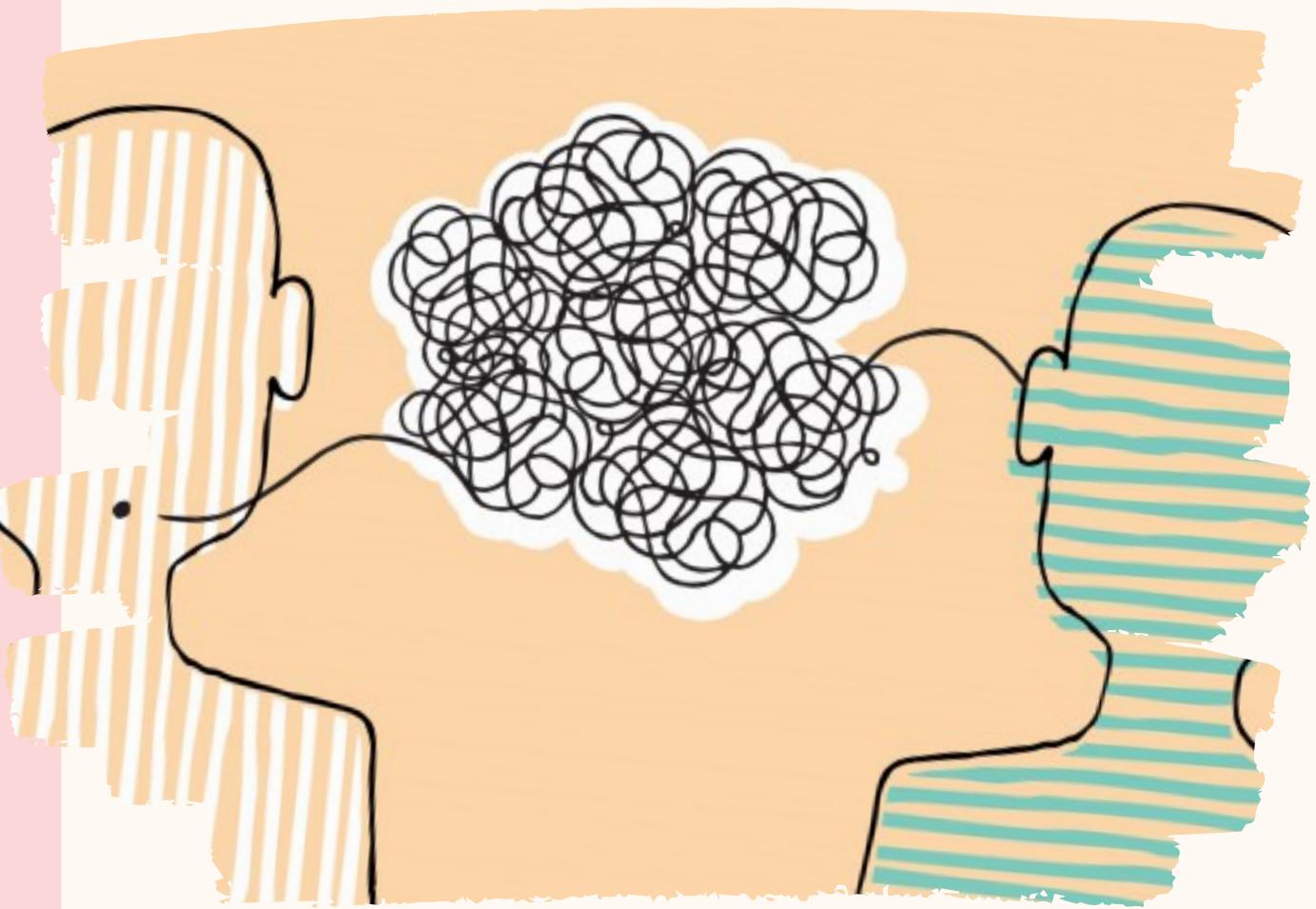
- Khi bạn đến các vùng khác nhau của Việt Nam, bạn phải học hoặc thay đổi giọng nói của mình cho phù hợp với các vùng này.
- -> Dành quá nhiều thời gian để học.



- Sự hiểu lầm

- Khó làm việc trong các lĩnh vực cụ thể : Giáo dục quan hệ công chúng

- RẤT NHIỀU



# GIỚI THIỆU

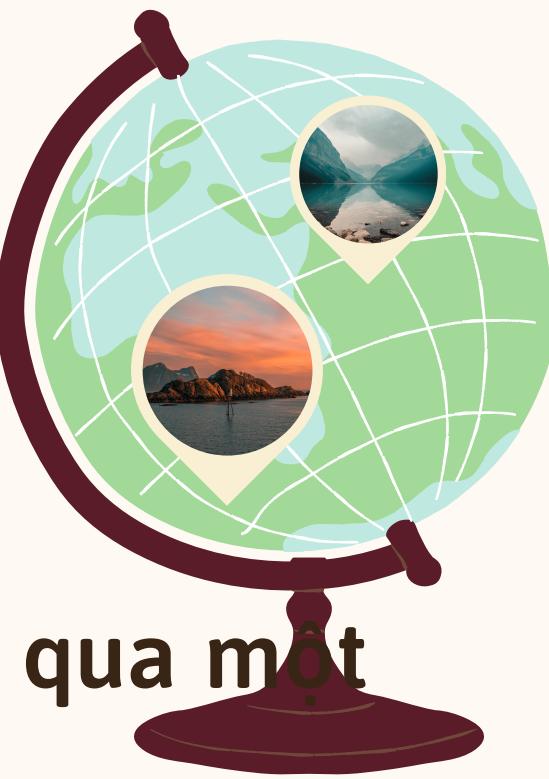
RIO DE JANEIRO

GUATEMALA

PERU



# GIỚI THIỆU



## DỊCH LÀ GÌ?

Đây là sự truyền đạt ý nghĩa của một văn bản ngôn ngữ nguồn thông qua một văn bản ngôn ngữ đích tương đương.

Tiếng Tây Ban Nha

vamos vamos  
argentina

×

Tiếng Việt

có lên argentina

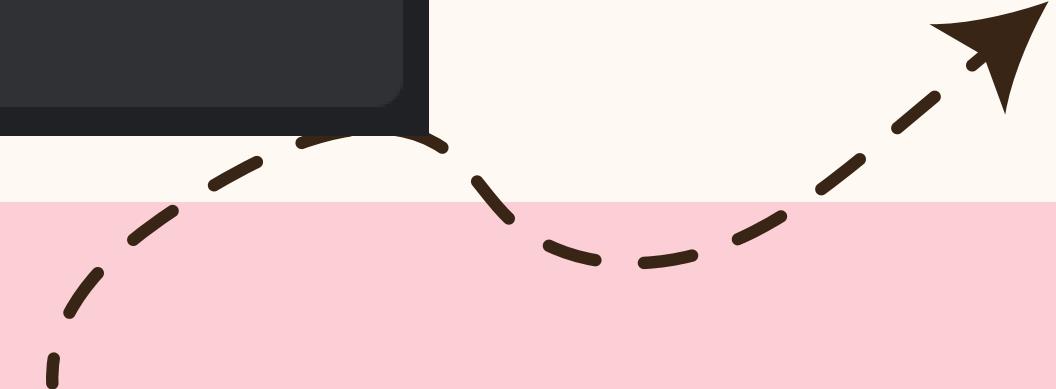
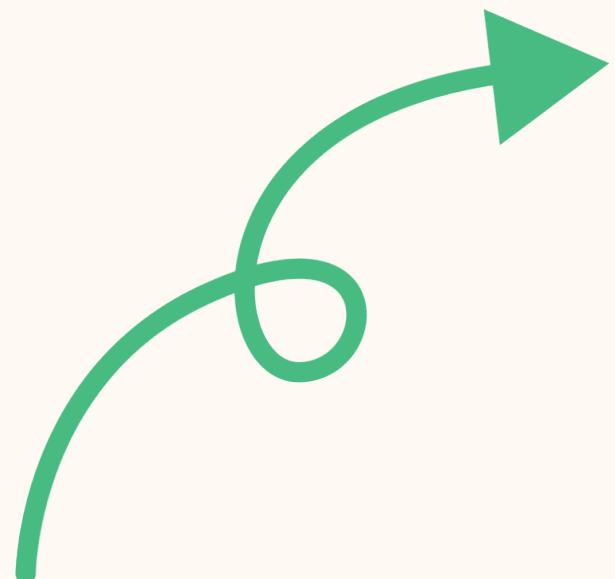
↔

Microphone icon

Speaker icon

Microphone icon

Speaker icon





# GIỚI THIỆU



## DỊCH MÁY LÀ GÌ?

Dịch máy (MT) là một trong những lĩnh vực nghiên cứu AI lâu đời nhất và những tiến bộ gần đây trong NLP đã dẫn đến những cải tiến lớn về chất lượng dịch thuật. Dịch máy là quá trình máy tính sử dụng để dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác, chẳng hạn như tiếng Anh sang tiếng Tây Ban Nha, mà không cần sự can thiệp của con người.

Các phiên bản đầu tiên của dịch máy có nhiều điểm không chính xác và lỗi dịch thuật. Trong những năm gần đây, sự phát triển như dịch máy thần kinh (NMT) đã giúp các công cụ AI xây dựng kiến thức của chúng để tạo ra các câu có sắc thái và chính xác hơn. Google Dịch, Microsoft Dịch, DeepL và IBM's Watson sử dụng công nghệ NLP mới nhất để cung cấp năng lượng cho hệ thống dịch máy của họ.



# GIỚI THIỆU

## LỢI ÍCH CỦA DỊCH MÁY

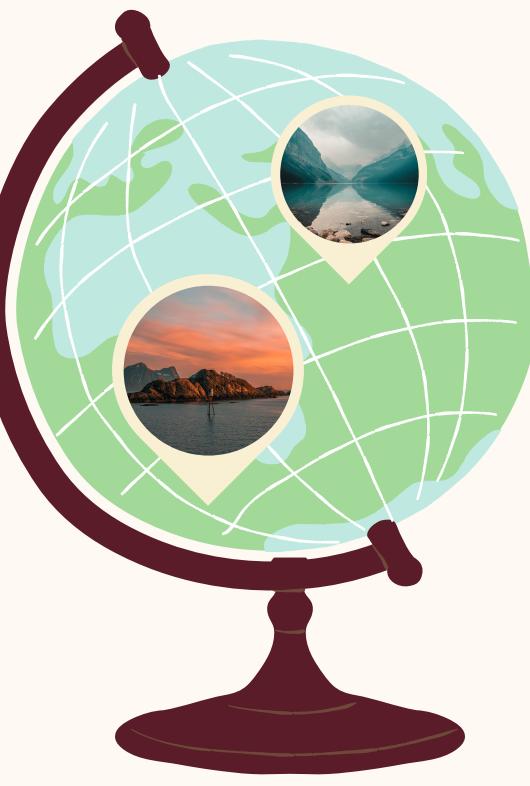
Các công cụ AI giúp quá trình dịch thuật trở nên dễ dàng hơn bao giờ hết. Một lợi ích quan trọng của dịch máy là khả năng xử lý nhanh như chớp của nó. Giờ đây, máy tính có thể dịch toàn bộ cơ sở dữ liệu sách, trang web hoặc sản phẩm chỉ trong vài giây. Một lợi ích chính khác là chi phí. Nhiều công cụ dịch AI hàng đầu có phiên bản doanh nghiệp chi phí thấp dành cho các công ty muốn bản địa hóa trang web của họ.

Mặc dù NLP đã dẫn đến những tiến bộ to lớn trong dịch thuật ngôn ngữ, nhưng các bản dịch của AI vẫn chưa hoàn hảo. Bản dịch máy không phải lúc nào cũng hiểu được sự khác biệt về văn hóa hoặc bối cảnh dịch thuật như người đọc. Vì lý do này, vẫn cần sự giám sát của con người để dịch chính xác nội dung từ ngôn ngữ này sang ngôn ngữ khác.

Trong ngành dịch thuật, dịch máy mang lại nhiều giá trị nhất như một công cụ để tăng tốc quá trình bản địa hóa. Khi được sử dụng cùng với chỉnh sửa hậu kỳ của con người, bạn sẽ có được điều tốt nhất của cả hai thế giới.



# GIỚI THIỆU



**TIẾNG ĐỊA PHƯƠNG -> TIẾNG VIỆT CHUẨN**

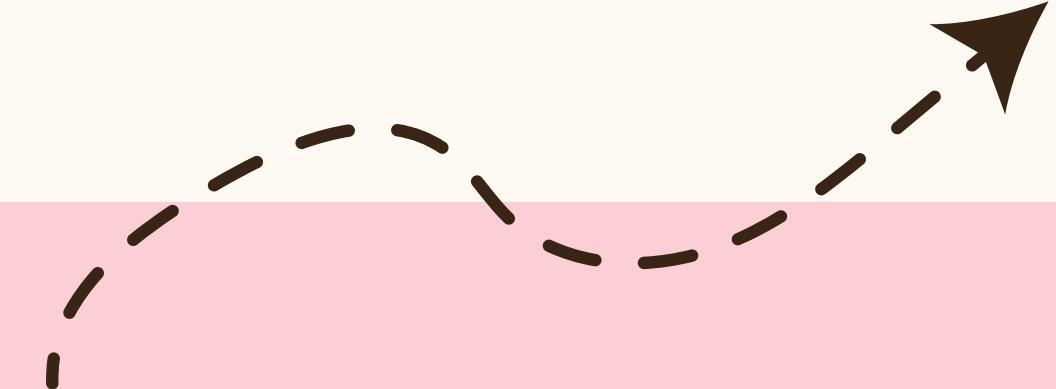
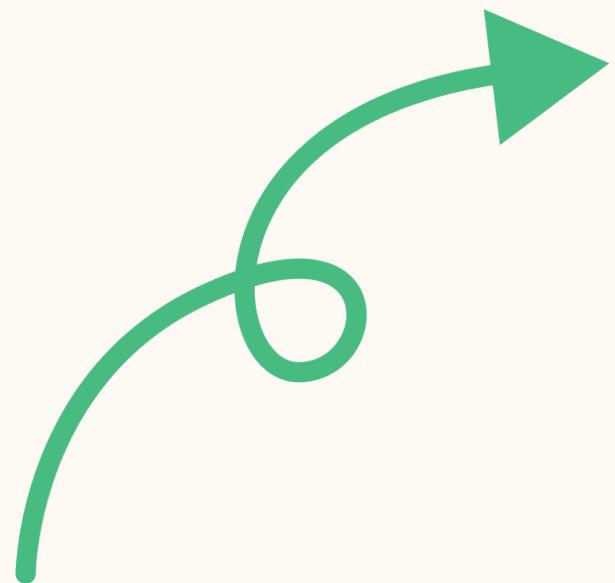
Tránh nhập nhằng về từ, cụm từ, ngữ nghĩa trong câu cần dịch.  
Mang bản sắc văn hóa Việt đến với mọi người dễ hơn.

PHIA RÂU, ĐI NGỦ THÂU

X



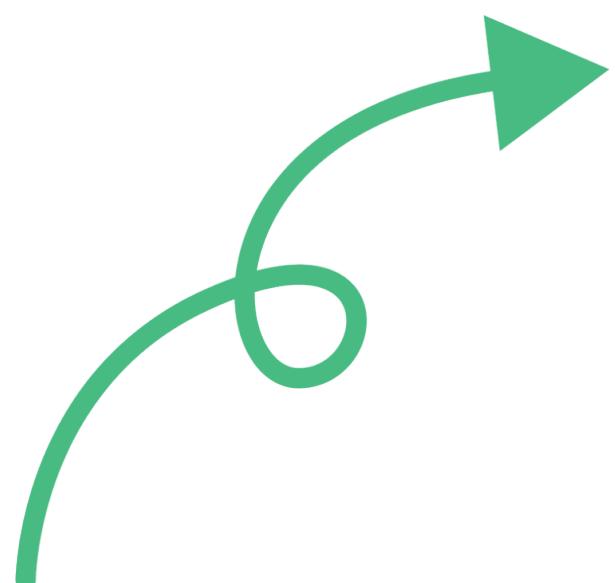
KHUYA RỒI, ĐI NGỦ THÔI



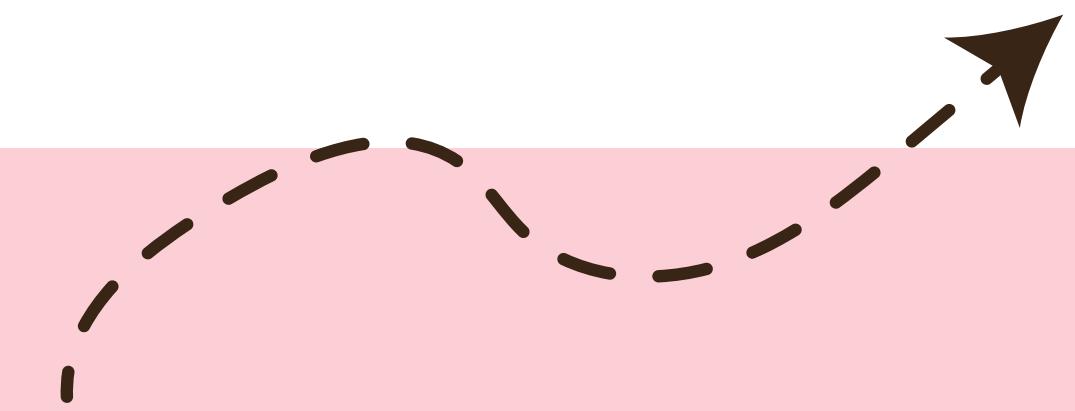


# GIỚI THIỆU

## VÍ DỤ VỀ CÁC TỪ ĐỊA PHƯƠNG - NGHĨA CỦA NÓ:



Ảnh	Anh ấy
Áo bà ba	Áo ngắn, tay áo dài, tra nút giữa, cổ kiềng, nam nữ đều mặc được.
Áo dà (chim)	Chim bìm bìm
Áo thun	Áo may ô
Áo mùng	Áo tang may bằng vải sô.
Áo ực	Rạo rực, thèm khát: <i>Mỗi bây lớn mà đã ực ực đòi vợ.</i>
Âm	1) Bồng, bẽ: <i>Âm em đi chơi chỗ khác.</i> 2) Lây, nhận về phần mình: <i>Thiên hạ có bao nhiêu tiên, nó ảm hết.</i>



Y TƯỞNG

RIO DE JANEIRO

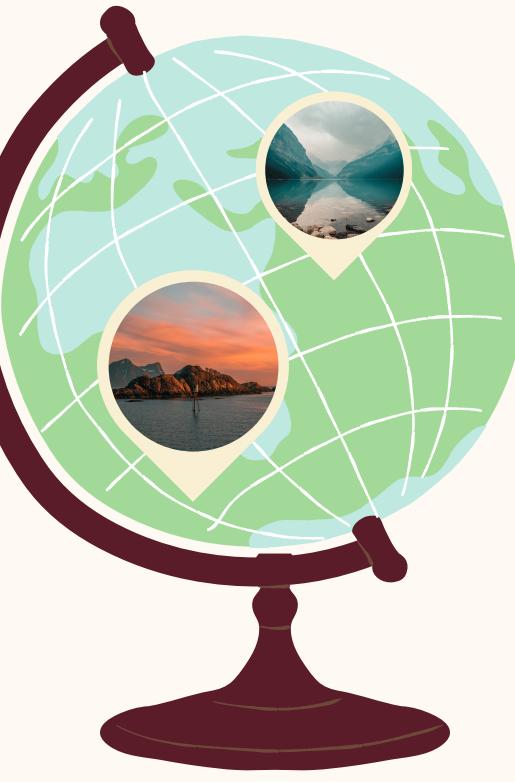
GUATEMALA





# Ý TƯỞNG

- Thiết lập mô hình:
  - + Tiếng Việt chuẩn
  - + Xử lí dữ liệu.
  - + Xử lí tiếng Việt.
- Xây dựng dữ liệu bằng SQL.
- Thiết lập web-app.



# CÁC THƯ VIỆN CẦN DÙNG





# CÁC THƯ VIỆN

- Xây dựng website: FLASK, GUNICORN.
- Phân tích audio: LIBROSA.
- TỐI QUAN TRỌNG: TENSORFLOW, SKIKIT-LEARN  
NUMPY, PYTORCH, SCIPY.



# MÔ HÌNH NGÓN NGỮ





# MÔ HÌNH TIẾNG VIỆT CHUẨN

- Data sources: VNEexpress



- Categories: Thời sự, Góc nhìn, Thế giới, Kinh doanh, Giải trí, Thể thao, Pháp luật, Giáo dục, Sức khỏe, Gia đình, Du lịch, Khoa học, Số hóa, Xe, Công đồng, Tâm sự.
- Number of articles: 210,109
- Number of different words: 157,362
- Total number of words: 92,956,179
- Top 10 từ phổ biến:

#	Word	Counts	#	Word	Counts
1	'và'	924063	6	'tôi'	655914
2	'có'	866613	7	'một'	636718
3	'của'	784810	8	'người'	620049
4	'không'	733558	9	'cho'	595462
5	'là'	685747	10	'trong'	587368





# SPEECH - TO - TEXT

- **Data sources:**

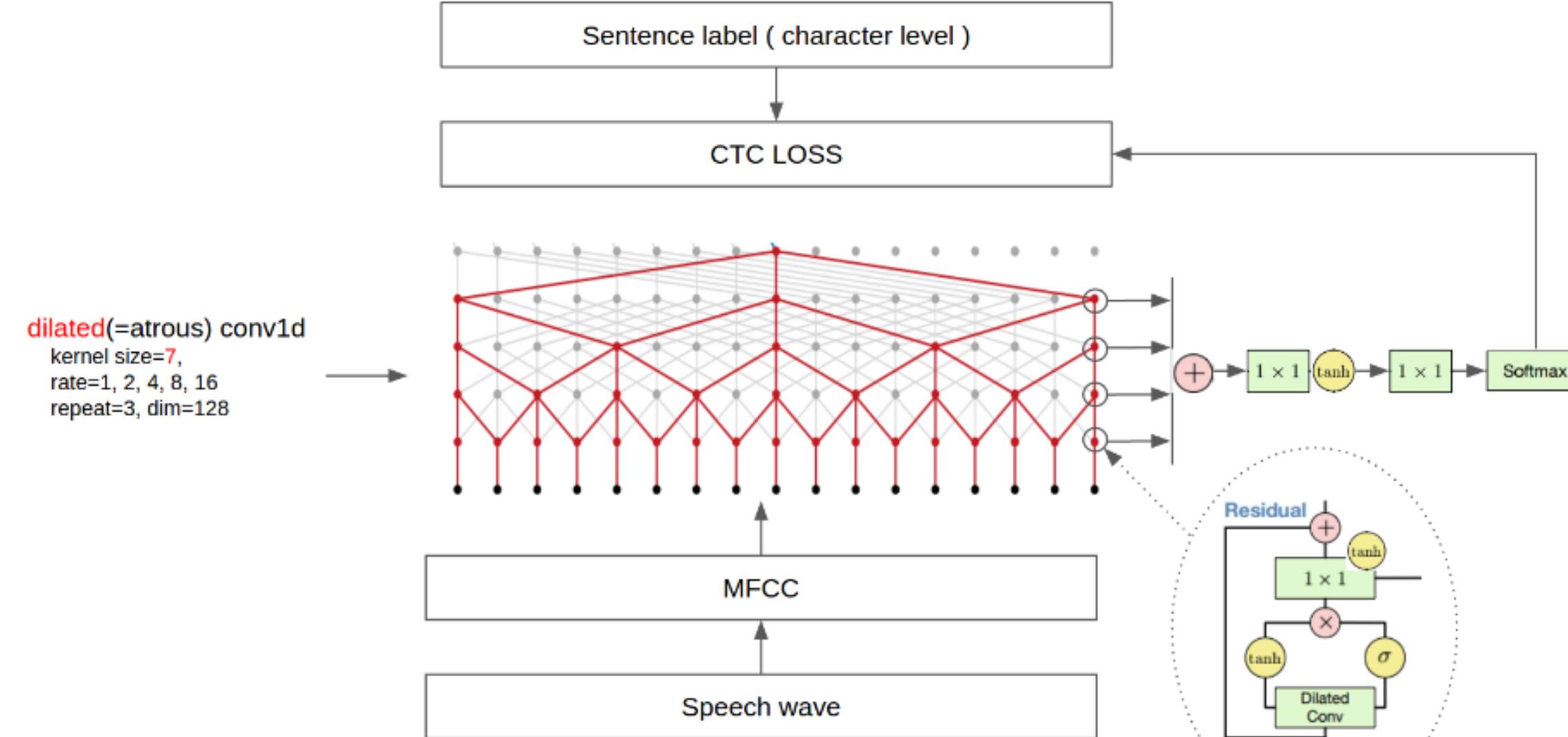
- + sachnoi.cc (Những trò ngụy biện sai thành trái).
- + sachnoiviet.net (Tam quốc diễn nghĩa).
- + soundcloud.com (Sông Đông êm đềm).

- **Dữ liệu xử lí:**

- Train set: 10000 samples ( > 5 hours)
- Test set: 7309 samples ( > 4.5 hours)



# TRAINING ARCHITECTURE (VIETNAMESE SPEECH-TO-TEXT)





# XỬ LÍ TIẾNG VIỆT

- Phân tách câu thành từng từ - cụm từ mang nghĩa địa phương.
- Kiểm tra từ - cụm từ trong database, nếu có thì ghép từ - cụm từ lại với nhau thành câu mới.
- Sử dụng bộ công cụ xử lí Tiếng Việt Underthesea với các chức năng chính:
  - + Word Segmentation.
  - + POS Tagging.
  - + Text Normalization.
  - + Sentence Segmentation.



## Underthesea





# XỬ LÍ TIẾNG VIỆT - XỬ LÍ ĐẦU VÀO

- Câu nói/text input: "Cỏi, ăn cơm đi con coi chứ nguội".
- Phân tách từ: "Cỏi", "ăn", "cơm", "đi", "con", "coi", "chứ", "nguội".
- POS Tagging:



- *Key = [('Cỏi', 'Np')] -> Kiểm tra key trong database, thấy 'Cỏi': Coi kìa.*
- Đưa ra kết quả: Coi kìa, ăn cơm đi con coi chứ nguội.





# DỮ LIỆU TIẾNG ĐỊA PHƯƠNG

- Data sources: Các nguồn/website trên google như: Ngosaokim, tiengvietonline,....
  - Lưu trữ trong SQL, từ Scratch, với thư viện Tensorflow.
  - Dữ liệu xử lí:
  - Train set: 1043 samples ( > 1 hours)
  - Test set: 541 samples ( > 0.5 hours)
- 





**WEB-APP**



RIO DE JANEIRO



GUATEMALA



# WEB-APP

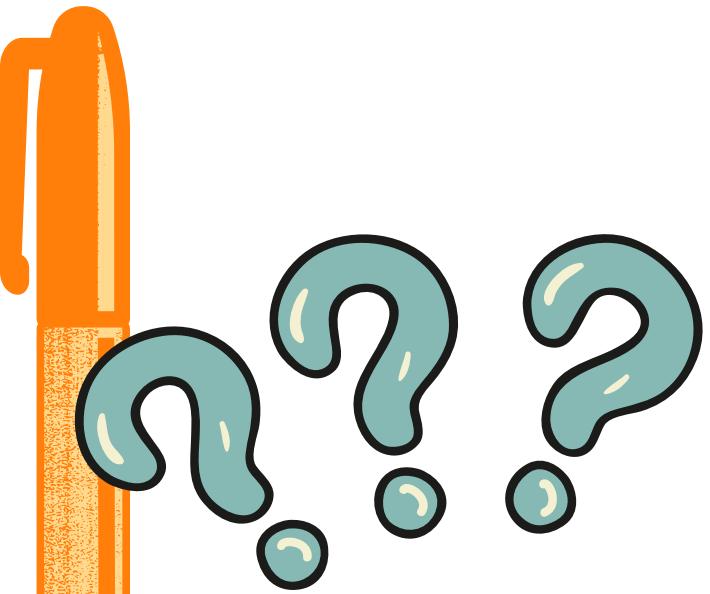
## MÔ HÌNH HUẤN LUYỆN

Epoch: 20

Step: 59283

Training loss: 11.84

Testing loss: 19.09



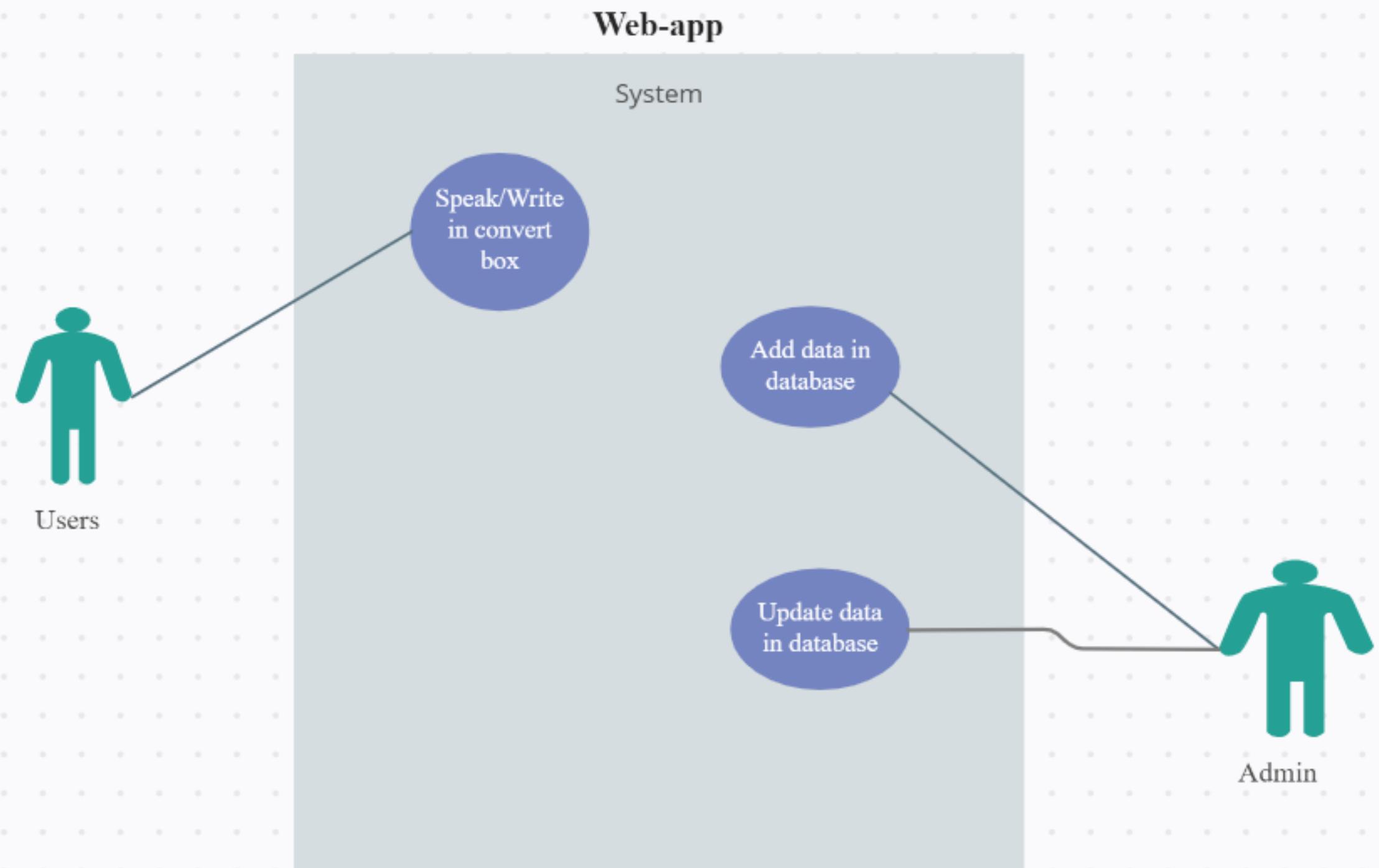
# WEB-APP

## USE-CASE SPECIFICATION

Use case ID	U001
Use case name	Speak/Write in convert box.
Summary	Users
Factors	Everyone whose devices connect to Internet
Condition	Access the Internet
Result	Is appeared by what you wrote (Include text and speech).
Script	<ol style="list-style-type: none"><li>1. Access the web-app.</li><li>2. Speak/Write something (in Vietnamese) to convert box.</li><li>3. Results will appear, users choose suitable for them.</li></ol>
Non-functional requirements	<ul style="list-style-type: none"><li>- Loading under a minute.</li><li>- Convert under two minutes</li></ul>

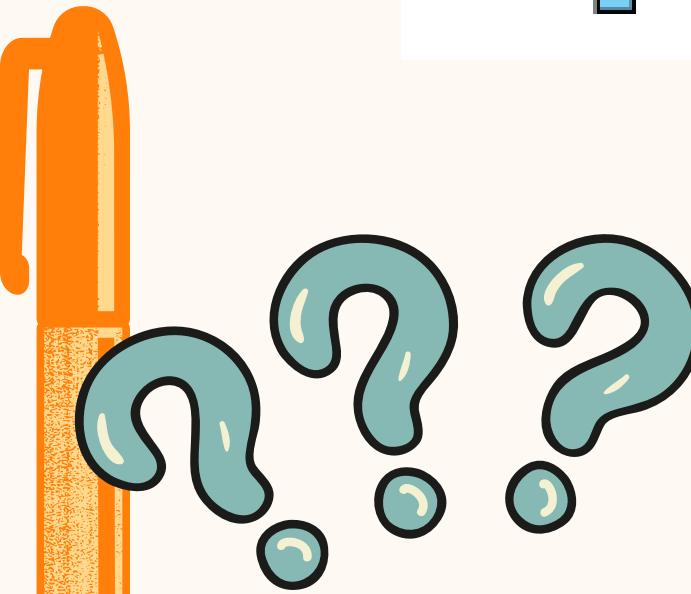
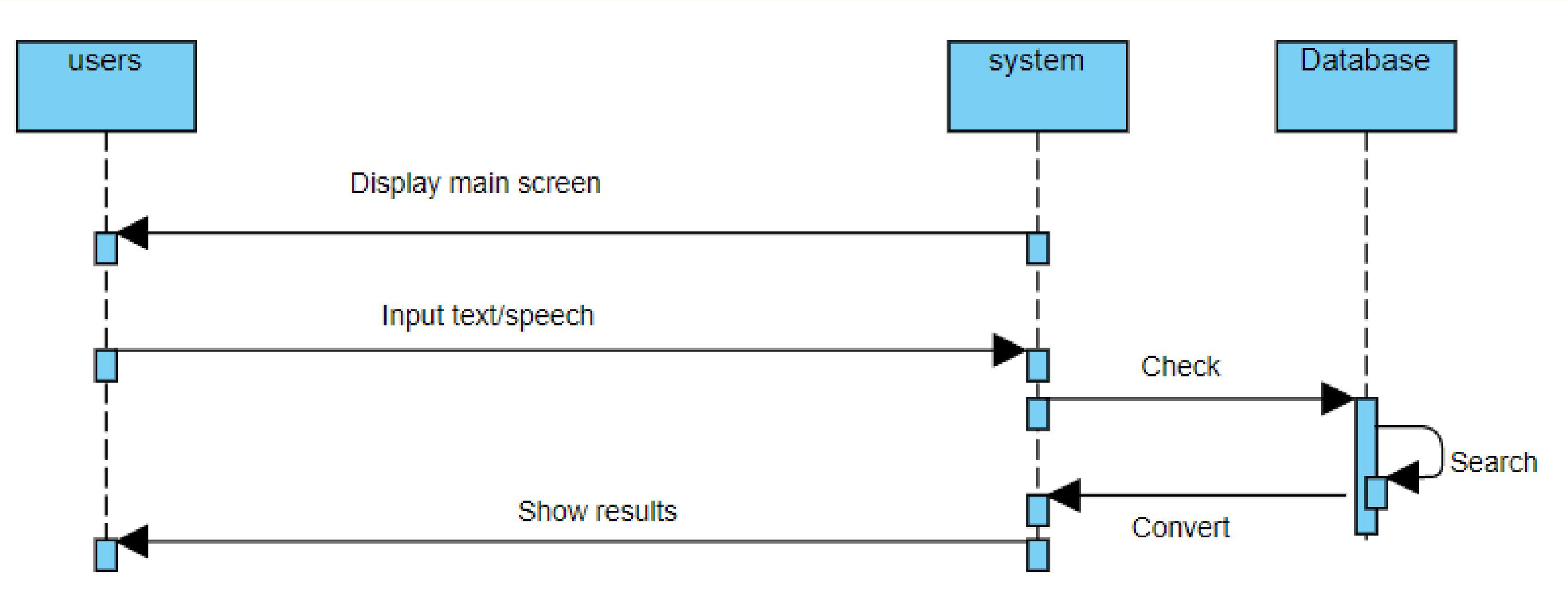
# WEB-APP

## SIMPLE USE-CASE DIAGRAM



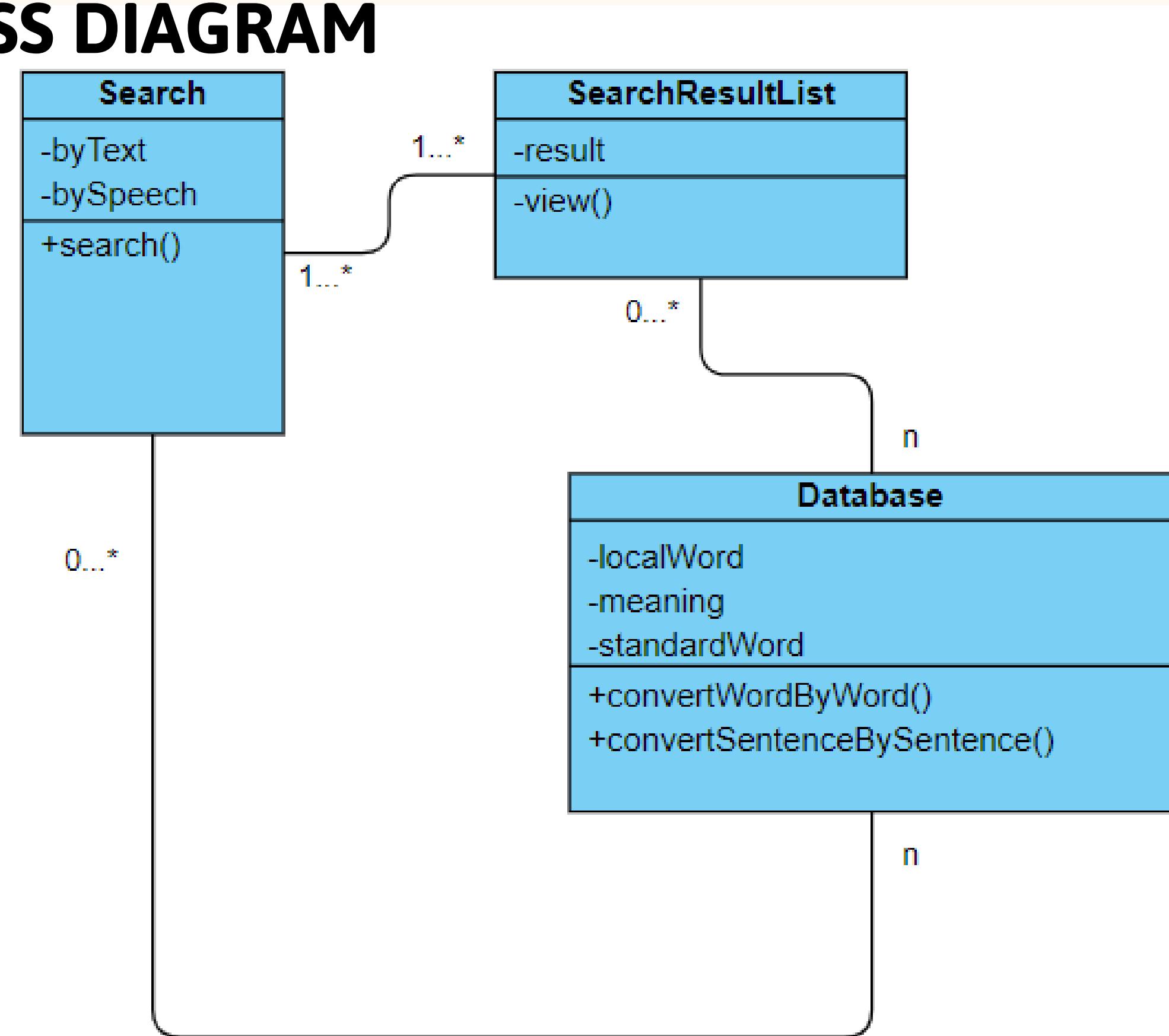
# WEB-APP

## SIMPLE SEQUENCE DIAGRAM



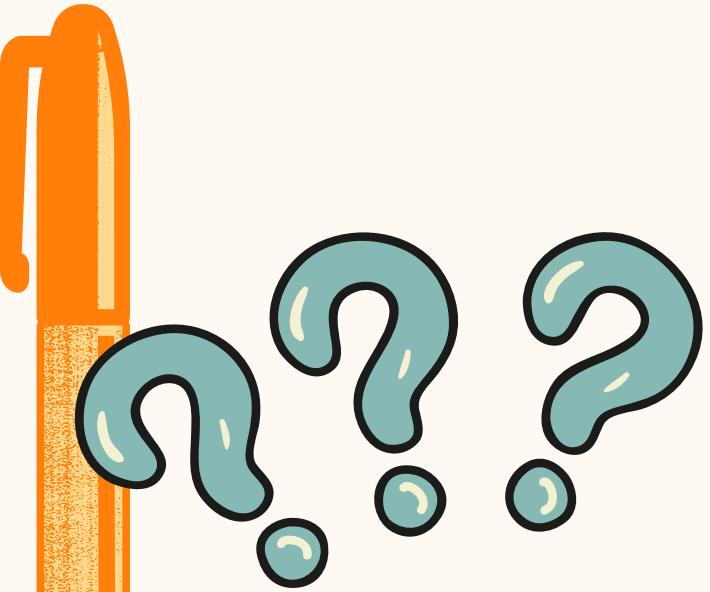
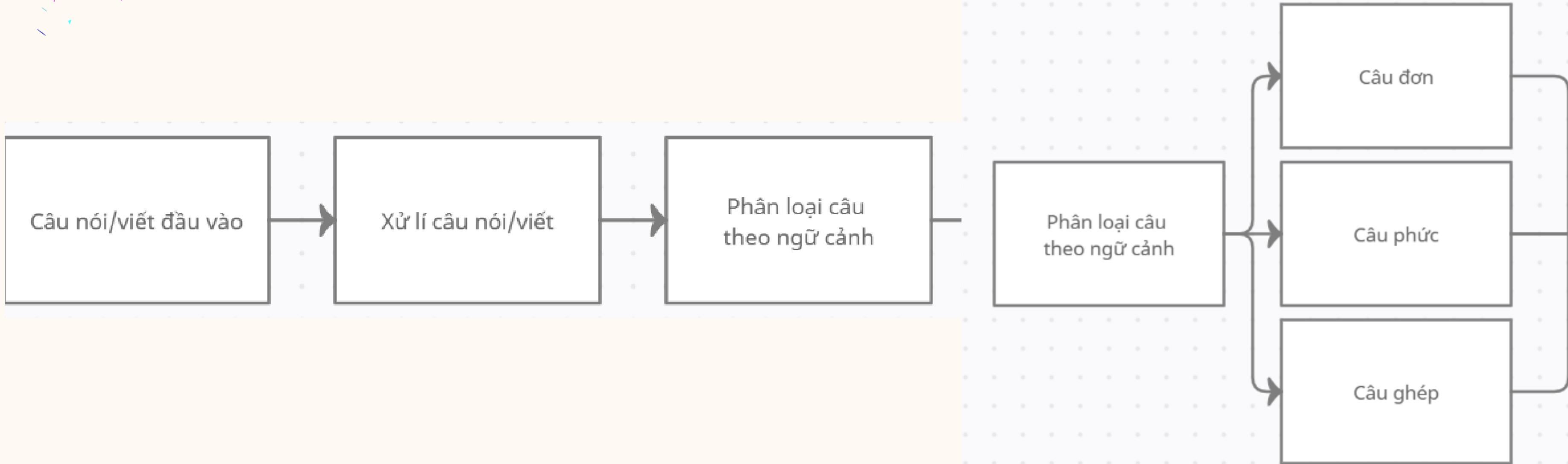
# WEB-APP

## SIMPLE CLASS DIAGRAM



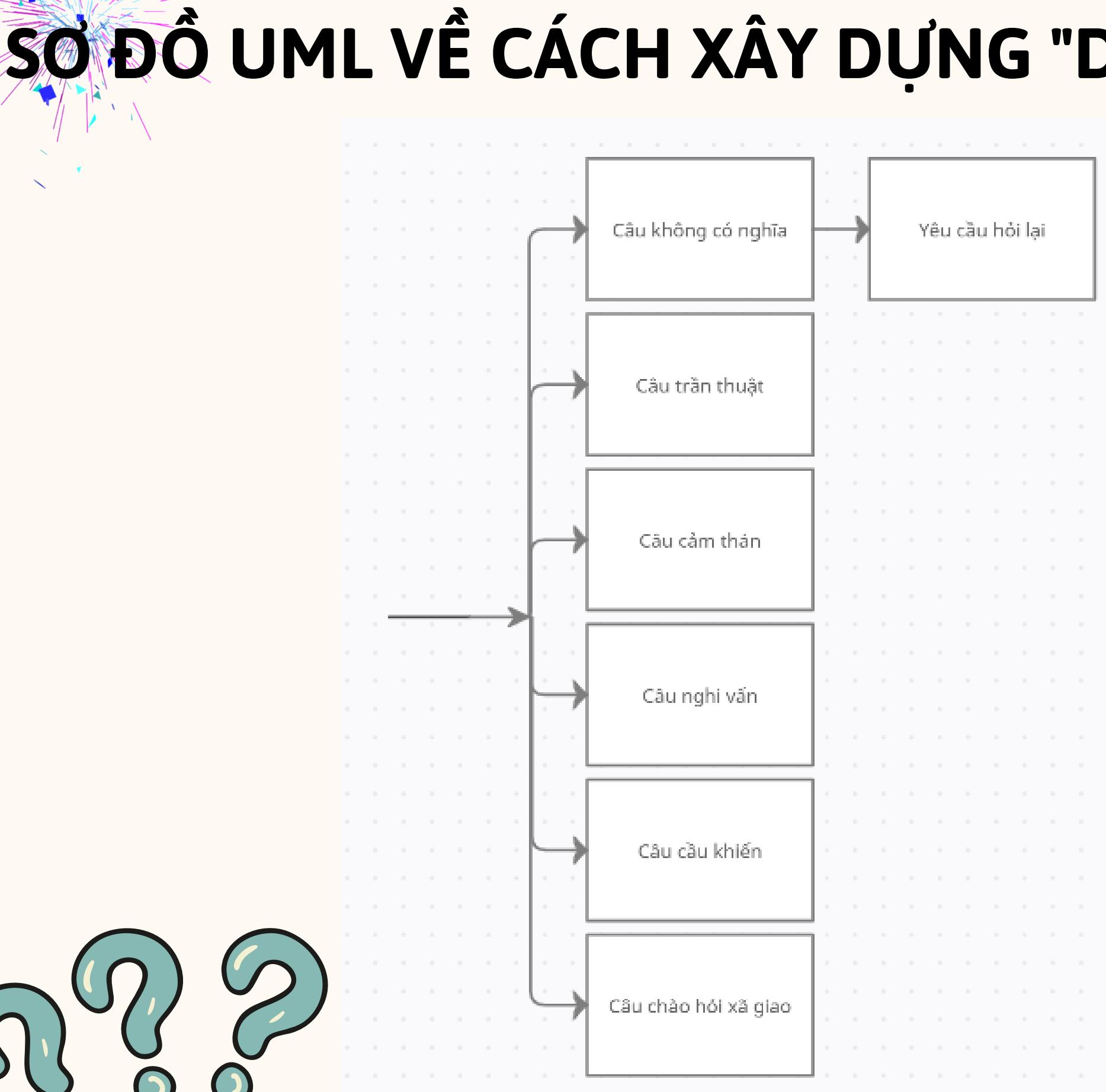


# SƠ ĐỒ UML VỀ CÁCH XÂY DỰNG "DỊCH"





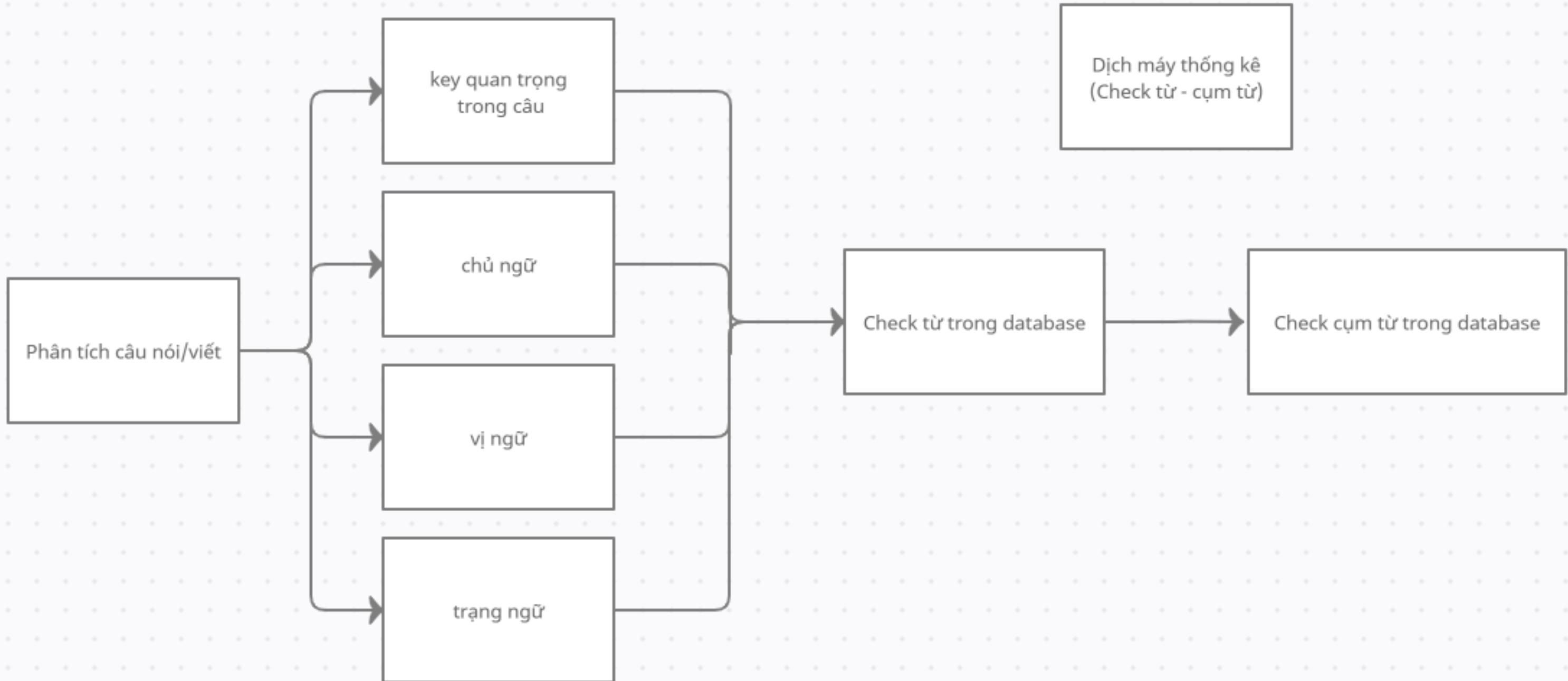
# SƠ ĐỒ UML VỀ CÁCH XÂY DỰNG "DỊCH"





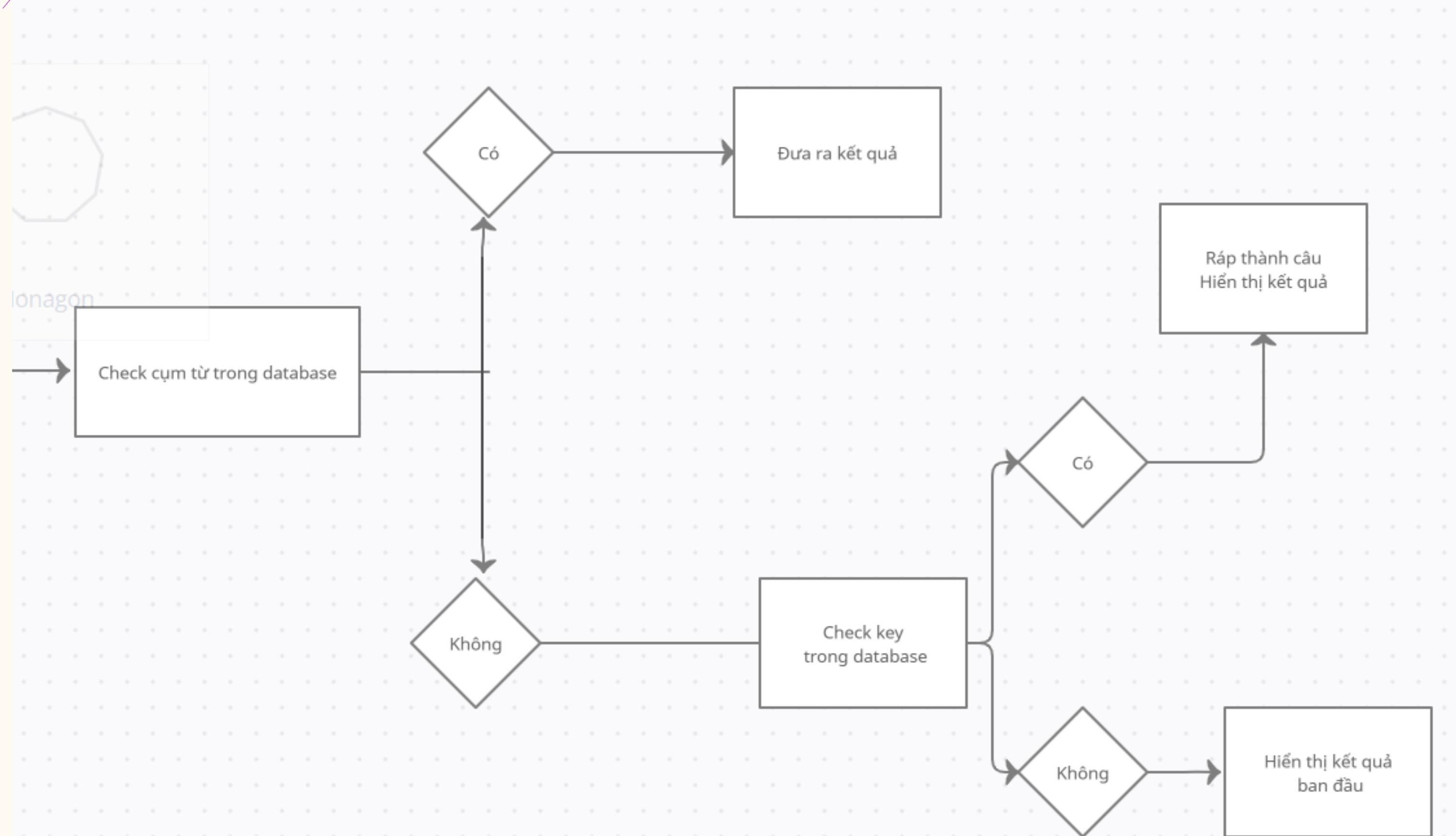
# WEB-APP

## SƠ ĐỒ UML VỀ CÁCH XÂY DỰNG "DỊCH"



# WEB-APP

## SƠ ĐỒ UML VỀ CÁCH XÂY DỰNG "DỊCH"





# THANK YOU FOR LISTENING!

Đừng quên đặt câu hỏi nghen.

Contact: 20127674@student.hcmus.edu.vn