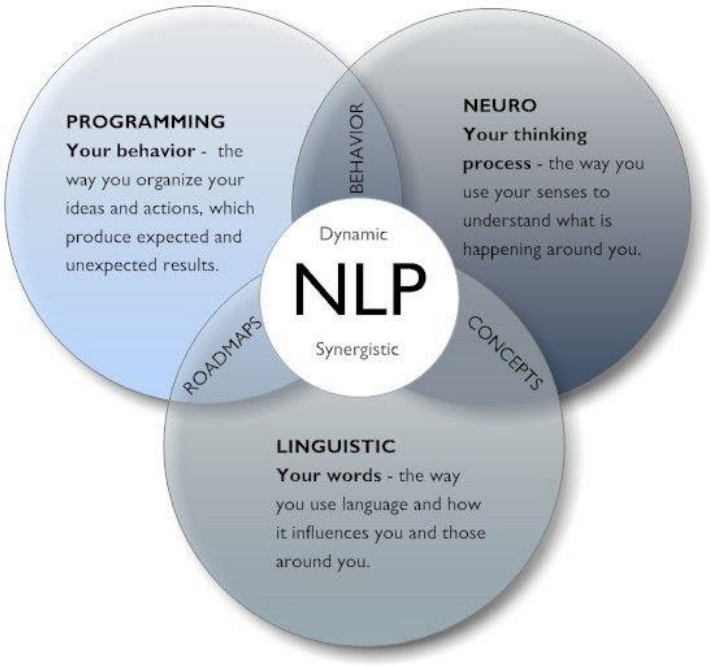
ĐỒ ÁN CUỐI KÌ
MÔN: NHẬP MÔN XỬ LÍ NGÔN NGỮ TỰ NHIÊN 20CNTTHUC01
ĐỒ ÁN: CHUYỂN ĐỔI TỪ TIẾNG ĐỊA PHƯƠNG SANG TIẾNG VIỆT CHUẨN

(Xem kết hợp với slide)



DECEMBER 31

NHÓM 16
GV HD: PGS.TS Đinh Điền
Thầy Lương An Vinh
Cô Lê Thị Thúy Hằng

Authored by: 20127674 – Lê Đức Đạt

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHTN
ĐHQG-HCM

Mục Lục

I. THÔNG TIN NHÓM 16: 3
II. PHÂN CÔNG CÔNG VIỆC – KẾ HOẠCH: 3
III. CÁC CHỨC NĂNG CỦA SẢN PHẨM ĐỒ ÁN: 3
IV. PHÂN TÍCH THIẾT KẾ ĐỒ ÁN: 3
1. Kiến trúc huấn luyện: 3
2. Kiến trúc web-app: 4
V. MÔ HÌNH DÙNG TRONG SẢN PHẨM – NGỮ LIỆU HUẤN LUYỆN: 5
VI. TƯ LIỆU THAM KHẢO: 6

I. THÔNG TIN NHÓM 16:

Họ và tên	MSSV
Lê Đức Đạt	20127674

II. PHÂN CÔNG CÔNG VIỆC – KẾ HOẠCH:

Chia làm 2 giai đoạn chính như sau:

Giai đoạn 1: Xây dựng các mô hình – database tiếng địa phương (100%).

- Xây dựng database về tiếng Việt địa phương từ mọi miền Tổ quốc bằng SQL.
- Tìm hiểu về các thư viện mới: LIBROSA (phân tích audio), UNDERTHESEA (toolkit phân tách câu thành từ/cụm từ tiếng Việt).
- Xây dựng các mô hình như: Tiếng Việt chuẩn (VNExpress), Speech-to-text (Nguồn: Wavenet), Xử lí tiếng Việt bằng Underthesea.
- Kết nối với nhau để tạo 1 sản phẩm tạm thời (chạy bằng Anaconda/Google Colab đều được).

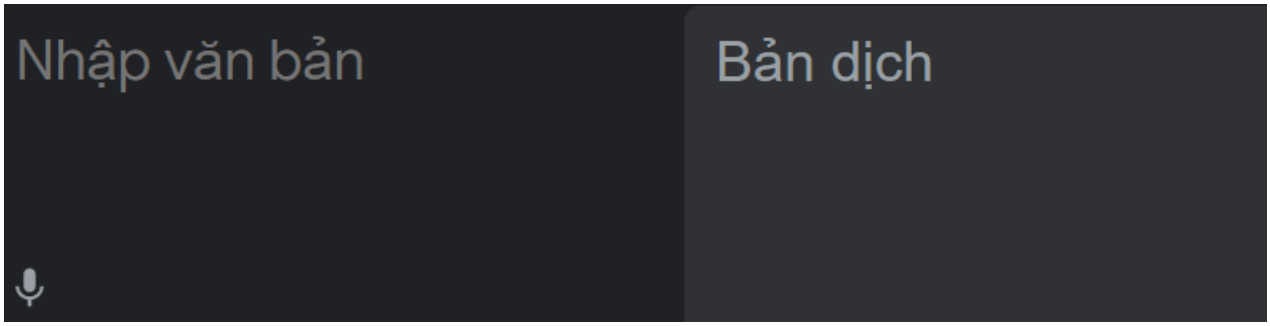
Giai đoạn 2: Xây dựng web-app (50%).

- Tìm hiểu về các thư viện mới: FLASK, GUNICORN (cho Web-app), SCRAPY(mã nguồn mở - tìm kiếm web, trích xuất dữ liệu).
- Xây dựng Web-app (khó khăn ở đây là làm Web, thế thôi).

"VIẾT BÁO CÁO CHUNG – LÀM TIẾP WEB".

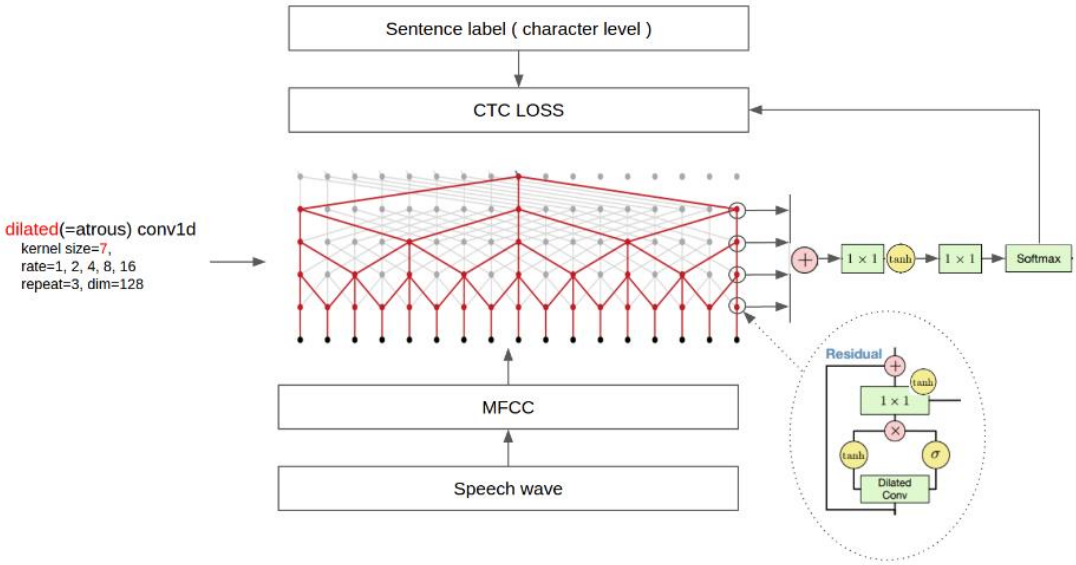
III. CÁC CHỨC NĂNG CỦA SẢN PHẨM ĐỒ ÁN:

Chức năng của nó rất đơn giản: Nói/gõ vào ô nội dung cần được chuyển đổi, ngay lập tức sẽ có kết quả ở ô đích, nó sẽ như thế này, tương tự như Google dịch, chỉ có điều là ở bản dịch sẽ không có dạng "Text-to-Speech":

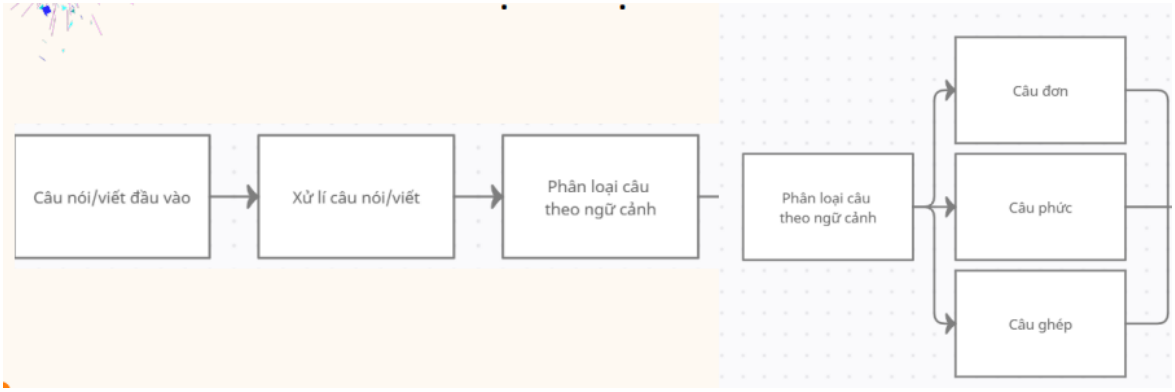


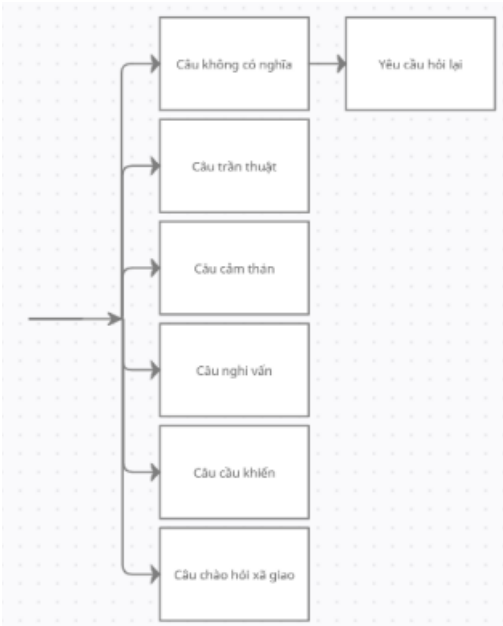
IV. PHÂN TÍCH THIẾT KẾ ĐỒ ÁN:

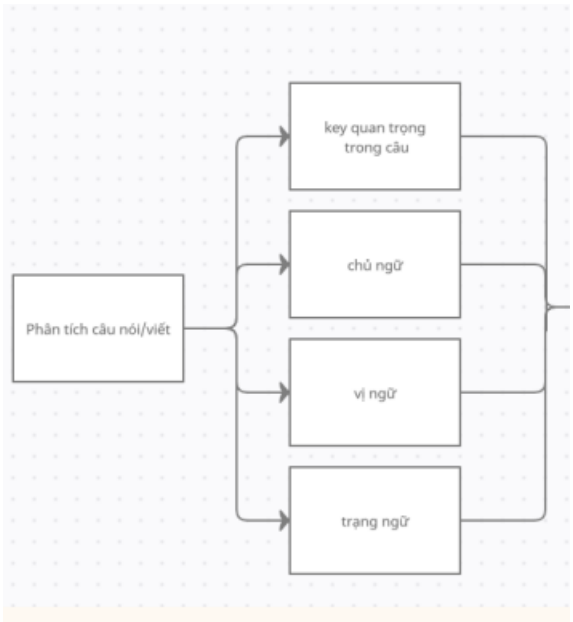
1. Kiến trúc huấn luyện:

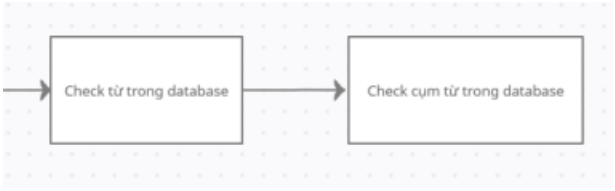


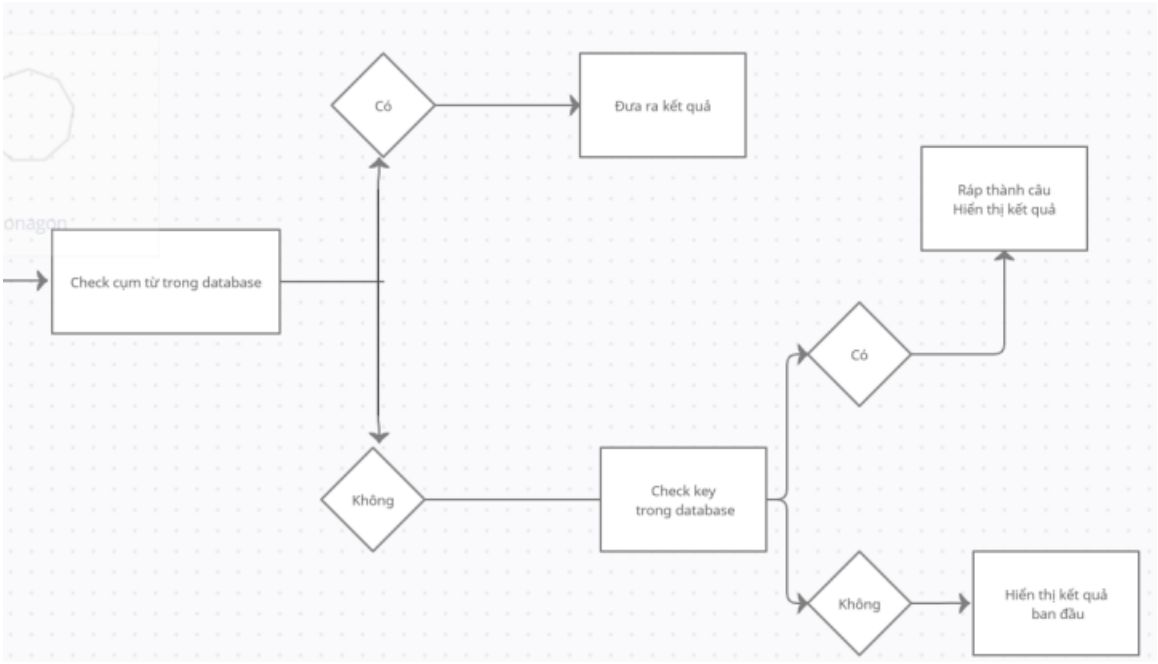
2. Kiến trúc web-app:











V. MÔ HÌNH DÙNG TRONG SẢN PHẨM – NGỮ LIỆU HUẤN LUYỆN:

SOMETHING	TIẾNG VIỆT CHUẨN	SPEECH-TO-TEXT	UNDERTHESEA	TIẾNG VIỆT ĐỊA PHƯƠNG
Chú thích	Có	Có	Có	Có
Đơn vị ngôn ngữ	Mẫu tự	Giọng nói	Câu, từ, cụm từ	Câu, từ, cụm từ
Bình diện	Hình thái, ngữ pháp, ngữ nghĩa	Ngữ âm, ngữ pháp, ngữ nghĩa	Hình thái, ngữ pháp, ngữ nghĩa	Ngữ âm, hình thái, ngữ pháp, ngữ nghĩa
Tag set	Tất cả từ, cụm từ, câu tiếng Việt chuẩn	Giọng nói, ngữ âm, ngữ điệu của con người	Tất cả từ, cụm từ, câu, dấu câu tiếng Việt (kể cả Chuẩn và địa phương).	Tất cả từ cụm từ, câu, ca dao, tục ngữ, ngữ âm, ngữ điệu, thành ngữ địa phương
Đầu vào	x	Văn bản nói, hoặc file audio.	Văn bản nói (Từ Speech-to-text)/văn bản viết (câu, từ, cụm từ).	Văn bản nói (Từ Speech-to-text)/văn bản viết (câu, từ, cụm từ).
Đầu ra	x	Text tiếng Việt chuẩn	Chuẩn hóa câu, thông tin các từ trong câu, vị trí chủ-vị-trạng của câu, key quan trọng	Như Underthesea, chỉ thêm kết quả là câu tiếng Việt chuẩn sau khi "chuyển đổi" trong

			của câu (là các từ địa phương).	database.

VI. TƯ LIỆU THAM KHẢO:

- GITHUB Vietnamese Speech-to-text: https://github.com/npanguyen412/vietnamese-speech-to-text-wavenet.
- GITHUB Underthesea: https://github.com/undertheseanlp/underthesea.
- Xây dựng Database với Tensorflow: https://towardsdatascience.com/natural-language-to-sql-from-scratch-with-tensorflow-adf0d41df0ca.
- Clip nhận dạng giọng nói tiếng Việt: https://www.youtube.com/watch?v=P3mhEngL1us.