Báo cáo tiến độ luận văn chatbot

**Trực Gia Phú**

**Huỳnh Trọng Nghĩa**

Table of Contents

[**1.** **Phân loại thêm câu hỏi:** 2](#_Toc35198421)

[**2.** **Mô hình câu hỏi đề xuất & các dạng câu hỏi:** 2](#_Toc35198422)

[**3.** **Kiến thức đã tìm hiểu thêm** 5](#_Toc35198423)

[3.1 Kiến thức cơ bản về machine learning: 5](#_Toc35198424)

[3.2 Kiến thức cơ bản về Python: 7](#_Toc35198425)

[3.3 Kiến thức chung về NLP (xử lí ngôn ngữ tự nhiên): 7](#_Toc35198426)

[**4.** **Cài đặt thử RASA:** 9](#_Toc35198427)

# **Phân loại thêm câu hỏi:**

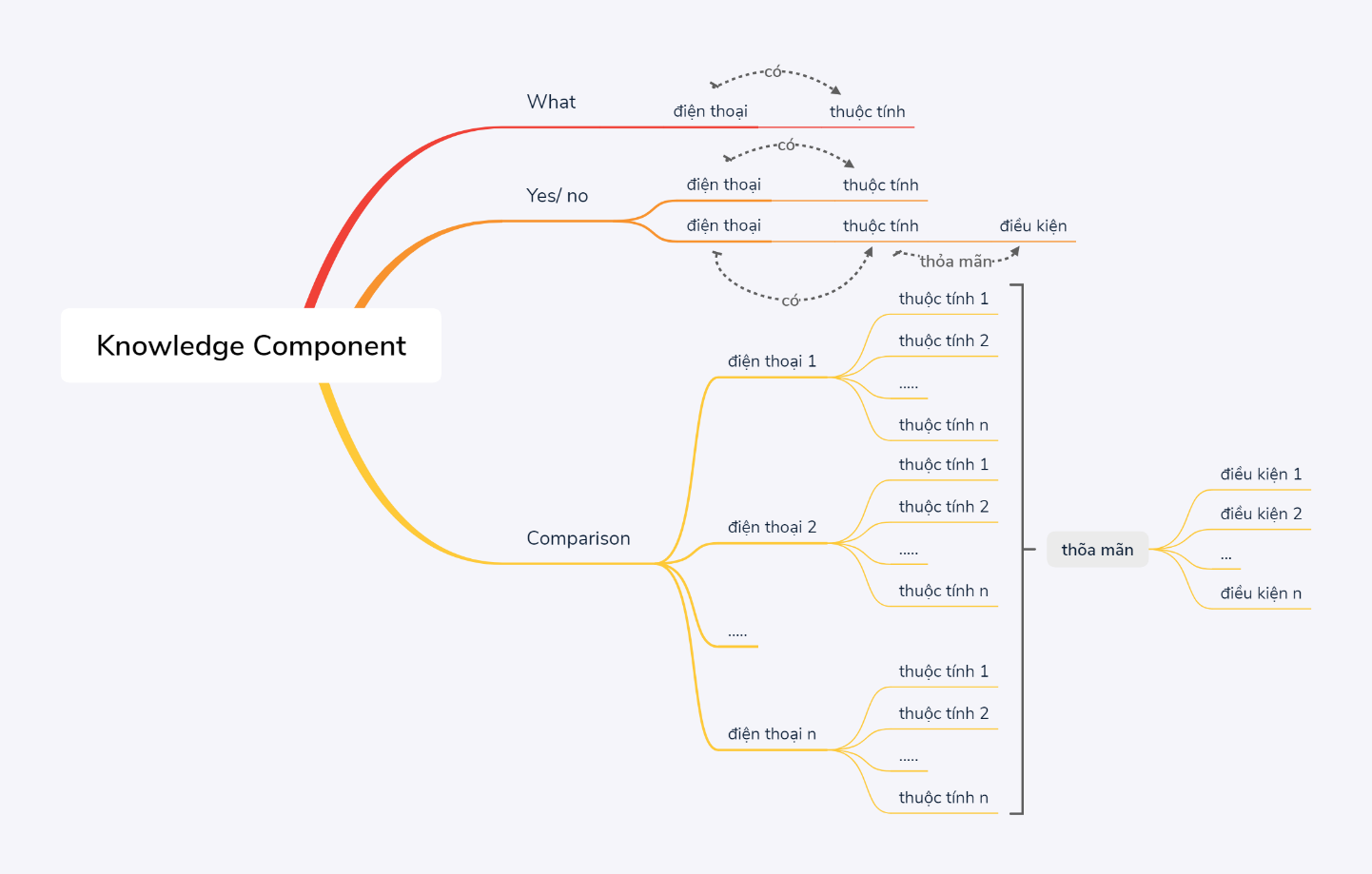
Nhóm đã thực hiện crawl dữ liệu thêm 761 câu hỏi từ các trang thegioididong.com, fptshop.com.vn, trong đó có 165 câu hỏi what, 22 câu hỏi why, 52 câu hỏi how, 9 câu hỏi when, 60 câu hỏi so sánh, 383 câu hỏi yes/no, 62 câu hỏi chưa xác định được dạng.

Tổng lượng câu hỏi nhóm đã phân loại: 1721 câu hỏi, trong đó có:

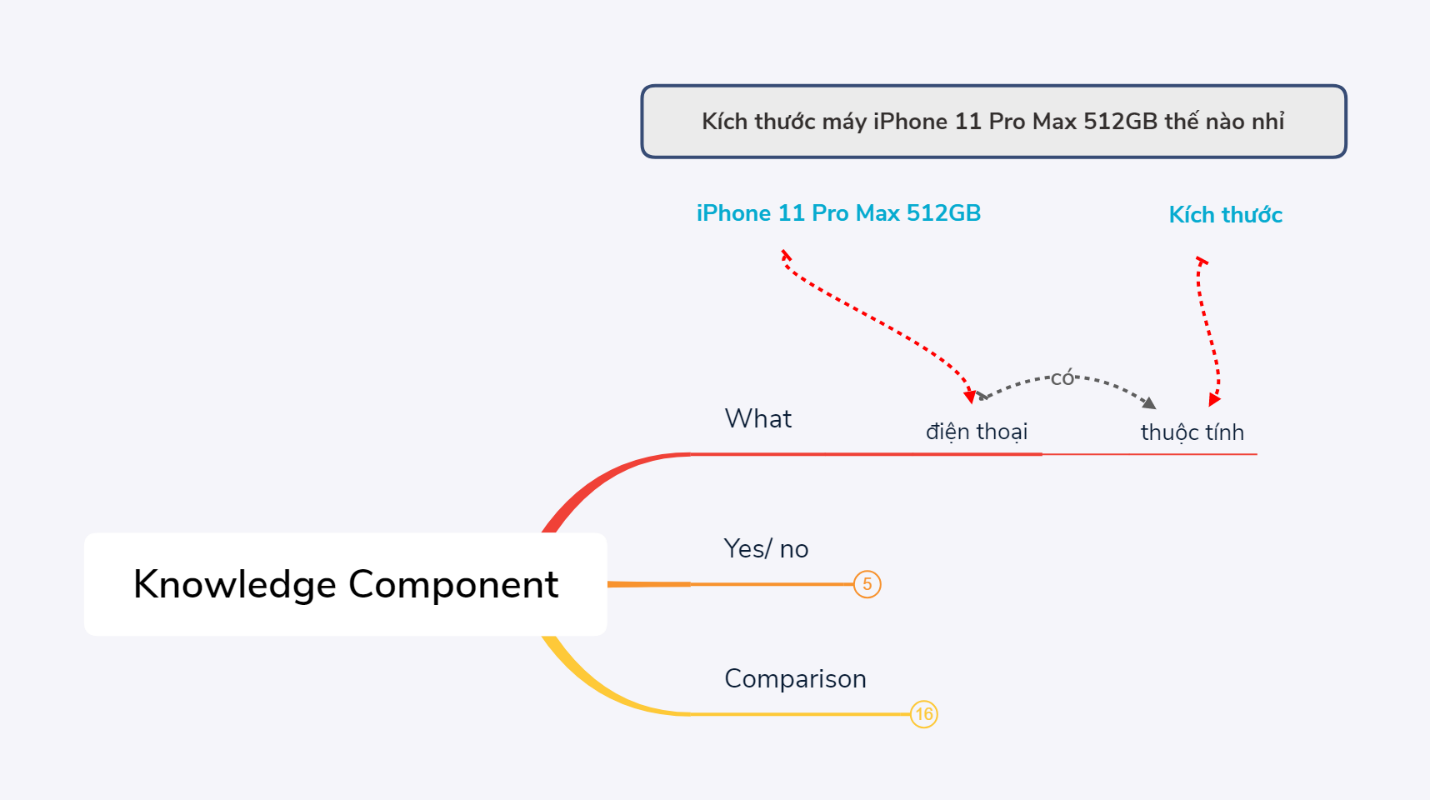
* 356 câu hỏi what
* 134 câu hỏi why
* 152 câu hỏi how
* 22 câu hỏi when
* 156 câu hỏi so sánh
* 780 câu hỏi yes/no
* 113 câu hỏi chưa xác định dạng (thường là để lại số điện thoại/ phàn nàn về chất lượng,…)

# **Mô hình câu hỏi đề xuất & các dạng câu hỏi:**

* Mô hình tổng quát:

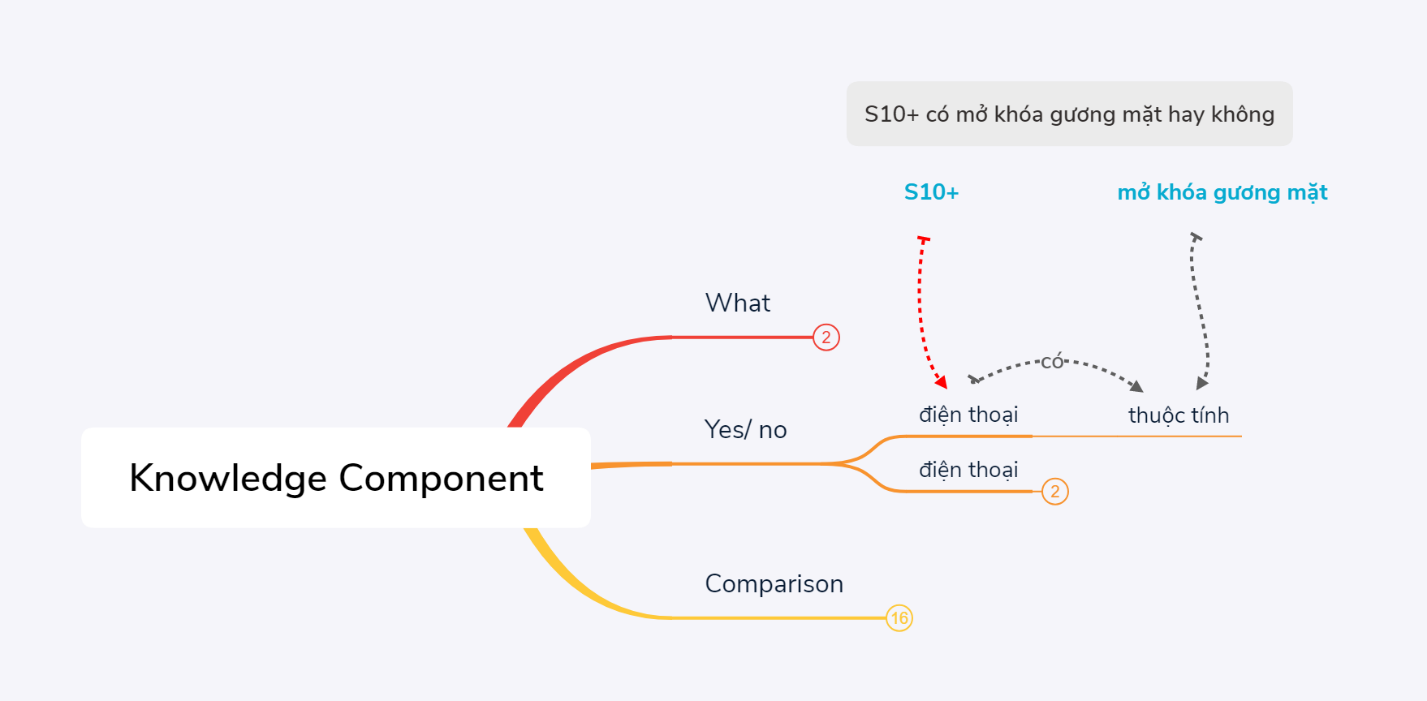


* Chi tiết cho câu hỏi What:

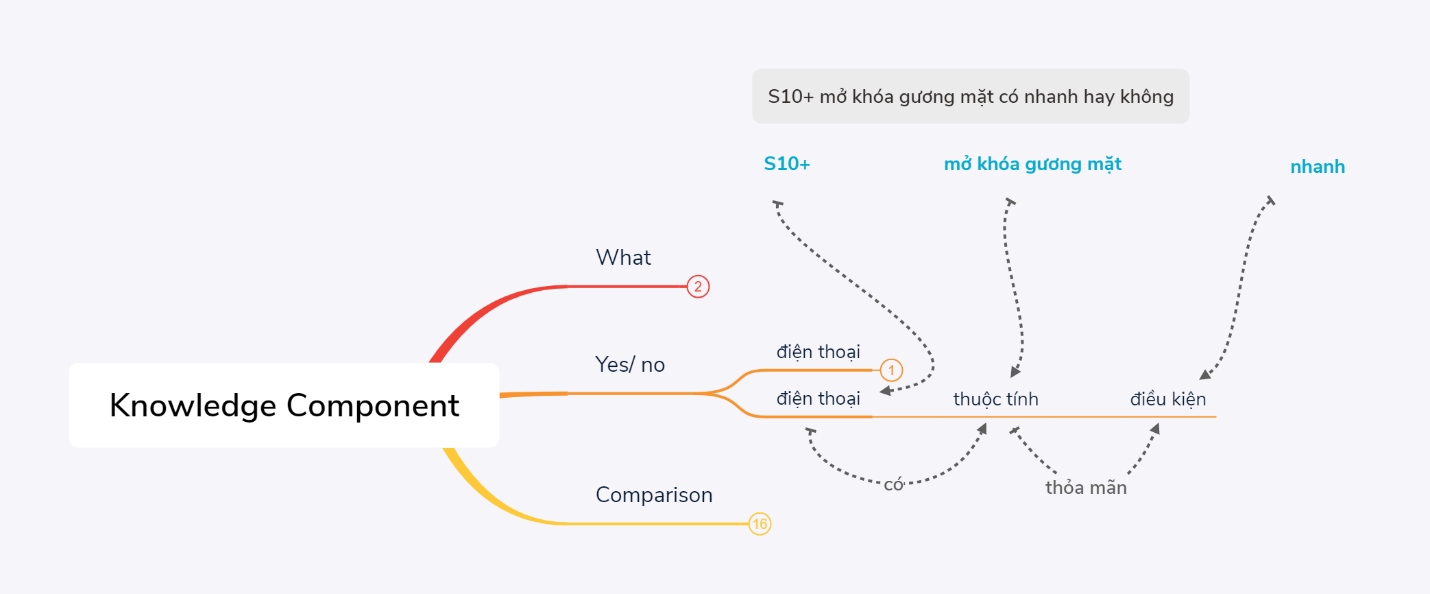


Khi người dùng hỏi: “Kích thước iPhone 11 Pro Max 512GB thế nào nhỉ”, đây là dạng câu hỏi What, entities có thể rút trích là “iPhone 11 Pro Max 512GB” và “Kích thước”, sau khi đã có entities cần thiết, ta thực hiện truy vấn tìm object có tên là “iPhone 11 Pro Max 512GB” và thuộc tính “kích thước” của object đó để trả lời câu hỏi

* Chi tiết cho câu hỏi dạng Yes/no – loại 1:

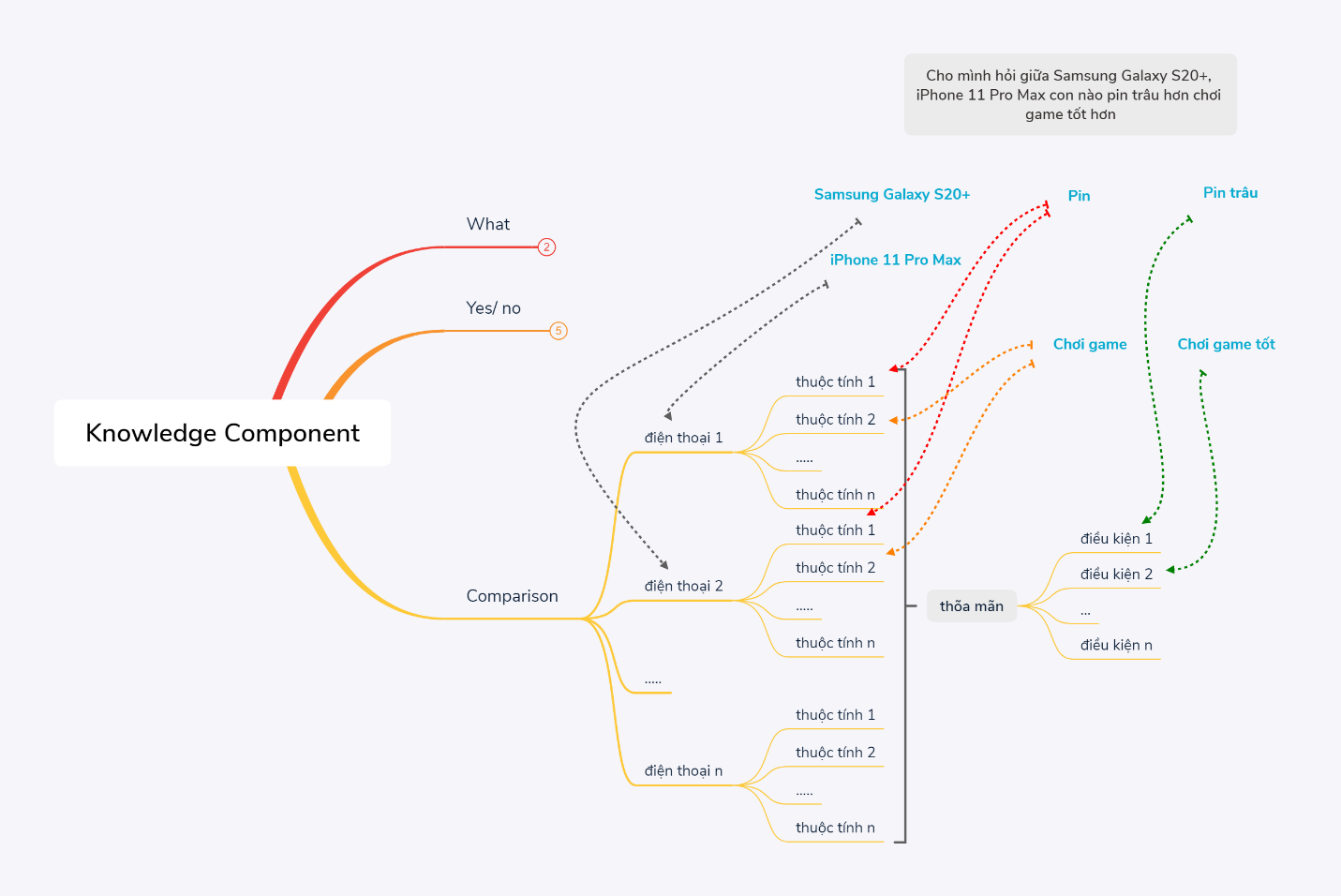


* Chi tiết cho câu hỏi dạng Yes/no – loại 2:



Cũng giống như loại 1, nhưng lần này câu hỏi yes/no loại 2 có thêm điều kiện.

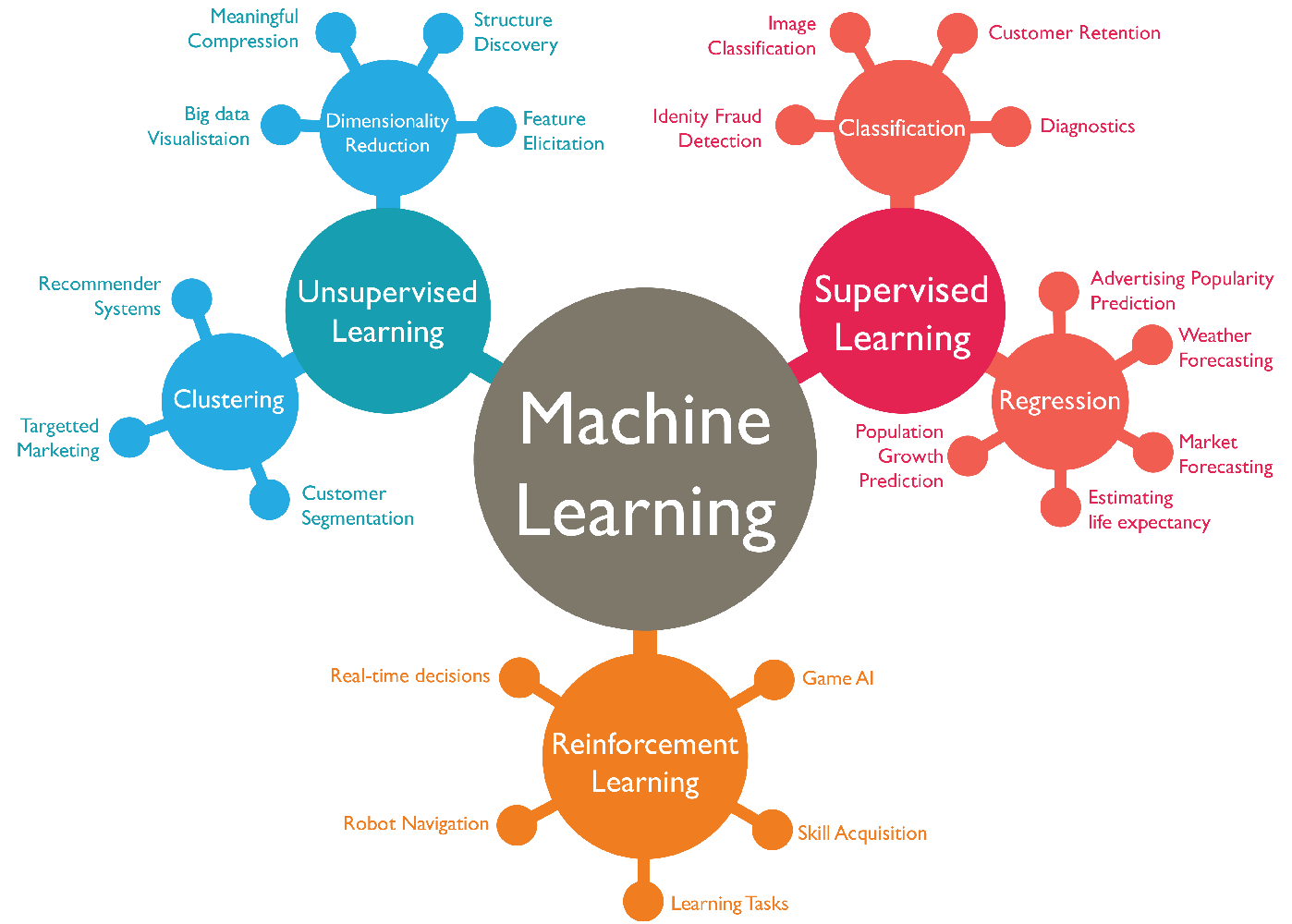
* Chi tiết cho câu hỏi dạng so sánh:



Với câu hỏi so sánh, thường ta sẽ thấy yêu cầu so sánh 2 hoặc nhiều điện thoại với 2 hoặc nhiều điều kiện cần so sánh. Lấy ví dụ: “Cho mình hỏi giữa Samsung Galaxy S20+, iPhone 11 Pro Max con nào pin trâu hơn chơi game tốt hơn”, những entities mong muốn được rút trích là tên điện thoại, thuộc tính cần so sánh, điều kiện so sánh.

# **Kiến thức đã tìm hiểu thêm**

## 3.1 Kiến thức cơ bản về machine learning:



Machine Learning thường chia làm 3 lĩnh vực: Supervised Learning, Unsupervised Learning, Reinforcement Learning.

* **Supervised Learning (Học có giám sát):** là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning. Trong supervised learning lại chia làm hai loại:
* Classification (Phân loại):

Một bài toán được gọi là classification nếu các label của input data được chia thành một số hữu hạn nhóm. Ví dụ: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không. Ba ví dụ phía trên được chia vào loại này.

* Regression (Hồi quy):

Nếu label không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ: một căn nhà rộng x m2, có y phòng ngủ và cách trung tâm thành phố z kmz km sẽ có giá là bao nhiêu?

* **Unsupervised Learning (Học không giám sát):**

Trong thuật toán này, chúng ta không biết được outcome hay nhãn mà chỉ có

dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ

liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc

giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ

và tính toán.

* **Reinforcement Learning (Học Củng Cố):**

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

Đối với phạm vi luận văn chatbot cho thương mai điện tử, thuật toán machine learning sẽ sử dụng là Supervised Learning – Classification. Input cho thuật toán là các câu hỏi được Crawled từ các trang bán điện thoại khác nhau như thegioididong.com, fptshop.com.vn, cellphones.com.vn. Các câu hỏi sẽ được phân loại cũng như đánh nhãn để làm dữ liệu training cho model. Kết quả mong muốn là khi đưa vào 1 câu hỏi chưa đánh nhãn, RASA sẽ phân loại câu hỏi (classify) và tìm ra được có những entity nào trong câu hỏi đó.

## 3.2 Kiến thức cơ bản về Python:

Những kiến thức cơ bản về lập trình Python (biến, hàm, kiểu dữ liệu, loop, module, virtual environment,…) đã được nhóm tìm hiểu.

Nhóm đã tham khảo course Python trên Udemy:

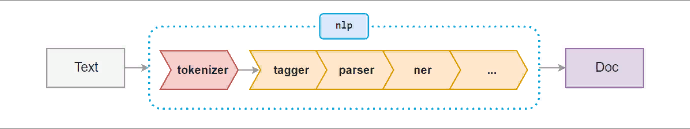
<https://bosch-etm.udemy.com/course/complete-python-developer-zero-to-mastery>

## 3.3 Kiến thức chung về NLP (xử lí ngôn ngữ tự nhiên):

Nhóm đã tham khảo course NLP trên Udemy:

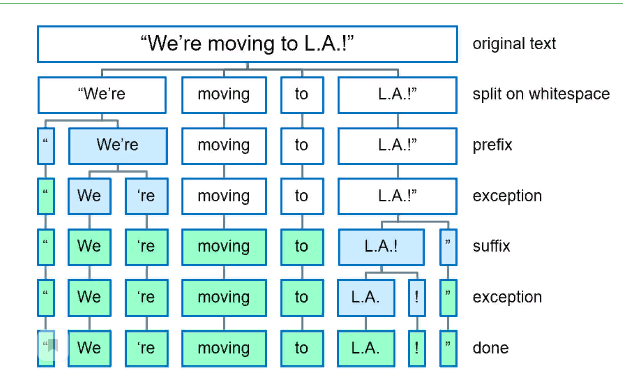
<https://bosch-etm.udemy.com/course/nlp-natural-language-processing-with-python>

Course này cung cấp kiến thức cho việc sử dụng thư viện Spacy để xử lí ngôn ngữ tự nhiên. Quy trình khi chạy câu lệnh “nlp” trên Spacy sẽ theo thứ tự như sau:

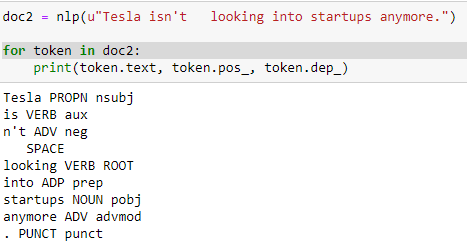


Những kiến thức chung thu được:

* Regular Expression: tạo pattern và match pattern.
* Tokenization: nhận vào 1 câu và tách những từ trong câu ra thành các token:

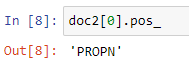


Mỗi token sau khi phân tách sẽ được gán nhãn và có thể truy suất thông tin của token đó qua các phương thức “.text”, “.pos\_”, “.dep\_”:



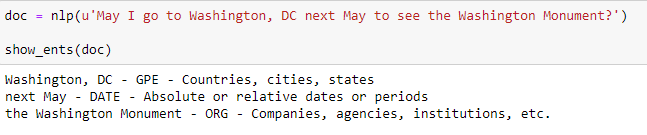
* Part-of-Speech Tagging (POS): gắn nhãn cho token và cho biết token đó là thành phần gì của câu (danh từ, động từ, trạng từ,..). Truy suất qua phương thức “.pos\_”

Ví dụ:



* Named Entity Recognition (NER): nhận dạng entity trong câu.

Ví dụ:

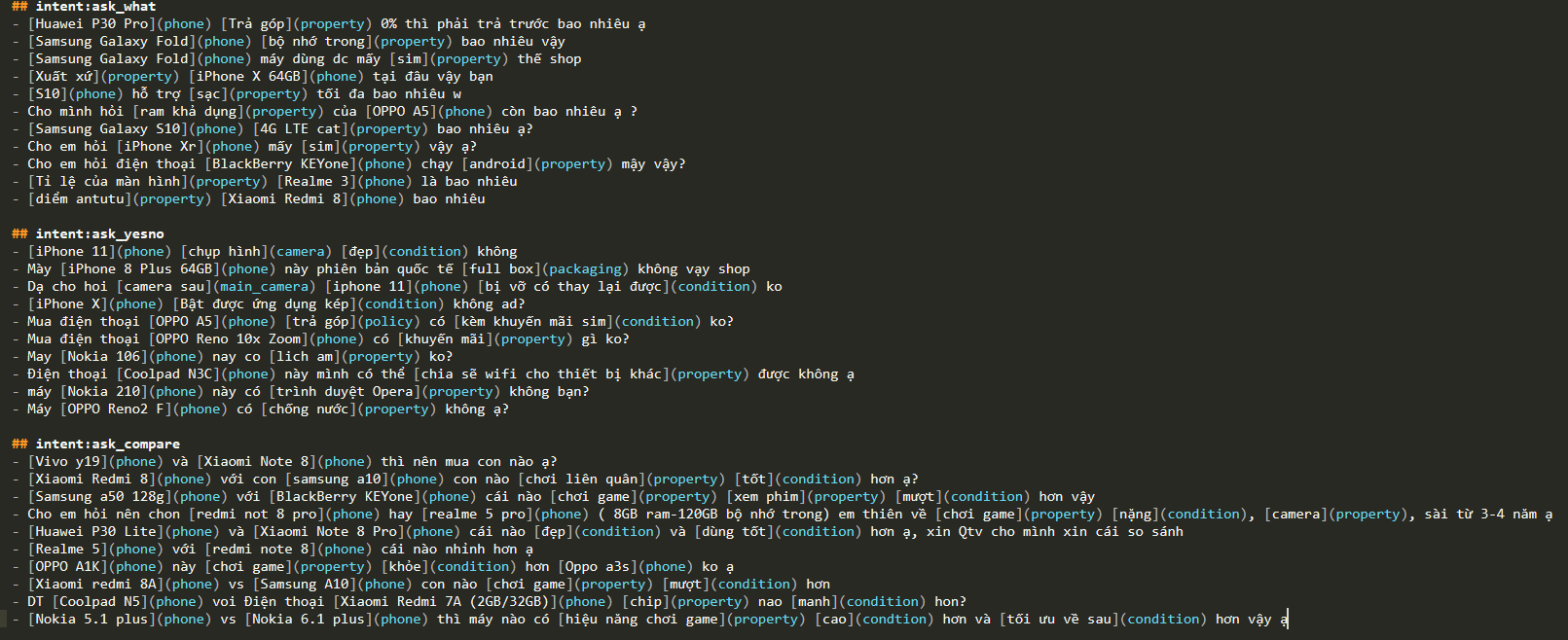


Một số entity phổ biến như địa danh, ngày tháng sẽ có sẵn trong Spacy, ta cũng có thể tự thêm entity riêng vào.

# **Cài đặt thử RASA:**

Nhóm đã setup môi trường chạy RASA thành công, tuy nhiên vẫn chưa tinh chỉnh configuration cho phù hợp nhất với bài toán chatbot bán hàng online (cần thời gian tìm hiểu) mà vẫn để cấu hình mặc định của RASA chạy thử với dữ liệu đã đánh nhãn số lượng nhỏ (3 intent: what, yes/no, so sánh, mỗi intent 10 câu đã đánh nhãn). Kết quả được trình bày như bên dưới.

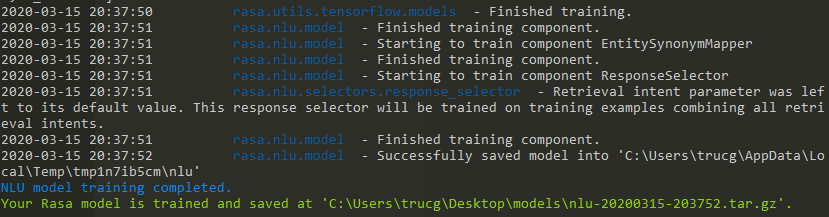
* + Chuẩn bị và đánh nhãn intent:



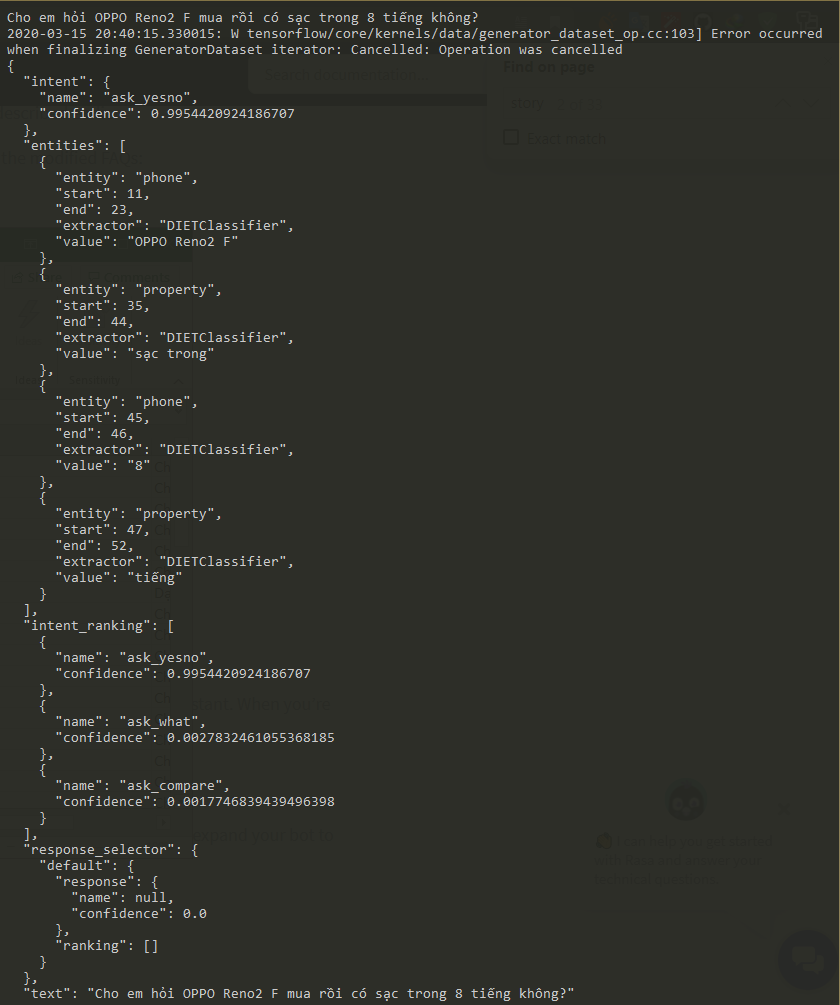
* Cấu hình mặc định của RASA:



* Hoàn tất training



Thử hỏi: Cho em hỏi OPPO Reno2 F mua rồi có sạc trong 8 tiếng không? Response của RASA:



Kết quả: phân loại được thành intent loại yes/no nhưng vẫn extract dư entity.

Thử hỏi: Giữa oppo k3 với realme 3 pro cái nào sẽ chiến game tốt hơn ạ? 