Annotation guideline

Mục tiêu:

Gán nhãn data vào 1 trong 4 lĩnh vực:

1. Khoa học tự nhiên (Địa chất học- Địa lý học, Hóa học, Khoa học máy tính, Logic, Sinh học, Thiên văn học, Toán học, Vật lý học, Y học).
2. Khoa học xã hội (Chính trị học, Giáo dục, Kinh tế học, Lịch sử, Luật pháp, Ngôn ngữ học, Nhân chủng học, Tâm lý học, Thần học, Triết học, Xã hội học, Địa lý hành chính).
3. Kỹ thuật (Công nghiệp, Cơ học, Điện tử học, Giao thông, Kiến trúc, Năng lượng, Người máy, Nông nghiệp, Quân sự, Y tế).
4. Văn hóa (Âm nhạc, Chính trị, Du lịch, Điện ảnh, Giải trí, Vũ đạo, Nghệ thuật, Phong tục tập quán, Thần thoại, Thể thao, Thời trang, Tôn giáo, Văn học).
5. Khác (Các mẫu không thuộc lĩnh vực nào trong các lĩnh vực trên)

Ví dụ:

title	content	label
Melocalamus scandens	Melocalamus scandens là một loài thực vật có hoa trong họ Hòa thảo . Loài này được Hsueh & C.M.Hui mô tả khoa học đầu tiên năm 1992. [1]	0
Quốc kỳ Argentina	Quốc kỳ Argentina (tiếng Tây Ban Nha : Bandera de la Argentina) là biểu tượng của nhà nước và quốc gia Argentina . Nó có ba dải màu nằm ngang kích thước bang nhau : màu xanh biển nhạt, màu trắng , màu xanh biển nhạt. Có nhiều cách giải thích về lý do cho những màu sắc.	3

Dundahera	Dundahera là một thị trấn thống kê (census town) của quận Gurgaon thuộc bang Haryana , Ấn Độ .	1
Nạn khan hiếm nhu yếu phẩm tại Venezuela	Nạn khan hiếm nhu yếu phẩm tại Venezuela (tiếng Tây Ban Nha : Escasez en Venezuela ; tiếng Anh : Shortages in Venezuela, 2016) được cho là hậu quả trực tiếp của sự sụt giảm giá dầu hỏa trong năm 2015 và gián tiếp là các quyết sách kinh tế tập trung của tổng thống Hugo Chávez cũng như người kế nhiệm Nicolás Maduro	1
Samsung S8000	Samsung Jet (còn được gọi là Samsung S8000), là điện thoại cảm ứng phát hành vào tháng 6 năm 2009 bởi Samsung . [1] [2] [3] Nó có vi xử lý 800 MHz và trình duyệt Web mới gọi là Dolfin, nó sẽ có sẵn trên điện thoại Samsung vào tương lai. [4]	2
4 tháng 2	Ngày 4 tháng 2 là ngày thứ 35 trong lịch Gregory . Còn 330 ngày trong năm (331 ngày trong năm nhuận).	4

Gán nhãn:

- Đa phần các mẫu đều khá dễ nhận biết. Ví dụ: các loài sinh vật → 0, thiết bị điện tử → 2, ...
- Các mẫu về thị trấn, xã, tỉnh, ... là các mẫu thuộc lĩnh vực KHXH (cụ thể là địa lý hành chính, chứ không phải địa lý học - KHTN)
- Có nhiều mẫu khó phân vào KHTN hay Kỹ thuật. Ví dụ:
    - Moxisylyte → 0. Bài viết thuộc vào thể loại Y học (KHTN), cần phân biệt rõ với Y tế (Kỹ thuật)
    - Liệu pháp ánh sáng → 2. Bài viết thuộc thể loại Y tế

- Những dữ liệu như 4 tháng 2 sẽ được xếp vào loại **Khác.**
- Những mẫu wiki về nhân vật có đóng góp trong nhiều lĩnh vực khác nhau (toán học, vật lý, hóa học, triết học, ...) sẽ được xếp vào nhãn 3. Những wiki về nhân vật hoạt động thể thao, nghệ thuật, chính trị cũng thuộc nhẫn 3.
- Những wiki về nhân vật lịch sử và hoạt động liên quan tới lịch sử thời kỳ tương ứng của họ sẽ được xếp vào nhãn 2.
- Các chú ý và quy tắc khác sẽ được bổ sung sau.

Bổ sung data thủ công:

Sau khi gán nhãn hơn 900 mẫu data, có tới hơn 60% là thuộc vào label 0. Lí do chủ yếu là ở việc có quá nhiều wiki về các giống, loài sinh vật. Những wiki này có số lượng quá lớn khiến tỉ lệ gặp phải bài viết thuộc loại nội dung này quá cao (dù trên thực tế thì những wiki này lại thuộc vào những wiki ít được truy cập nhất do nó chỉ nói về 1 loài sinh vật cụ thể).

Các mẫu ở label 1 chiếm khoảng hơn 20%, label 3 chiếm khoảng 10%, còn lại là label 2. Có 2 mẫu thuộc label 4 (1 trong đó là 4 tháng 2). Rất nhiều mẫu thuộc label 1 là các wiki về 1 xã, 1 thị trấn, ... nào đó. Những wiki này có tính chất tương tự với các wiki về 1 loài sinh vật cụ thể đã nêu ở trên.

⇒ Thực hiện bổ sung các mẫu dữ liệu bằng cách chọn các wiki đa dạng về nội dung hơn cho label 0 và 1 để phủ rộng đủ lĩnh vực. Với các label 2 và 3, bổ sung cả về mặt số lượng lẫn độ phủ rộng lĩnh vực.

Quality of life:

Gán nhãn bằng GPT và check lại kết quả label. Có nhiều nhãn thuộc các lĩnh vực như điện tử, vật liệu, y tế, ... (Kỹ thuật) vẫn bị đánh nhãn thành KHTN do những mẫu nãy có nội dung gần với vật lý, y học, ...