Guide to get youtube links

phamvandan.cntt2

November 2020

1 Yêu cầu các link audio

- Tiếng việt
- Một người nói vào một thời điểm
- Không chứa **nhiều tạp âm** như tiếng nhạc, tiếng ồn xe cộ,...
- Là âm thanh do người thực phát ra, không phải từ băng đĩa hoặc từ một bộ phát âm thanh tự động như text to speech của google, fpt,...
- Không thu quá nhiều link cùng một chủ đề, các chủ đề hiện tại đã thu được mô tả cuối trang.

2 Cách thu và lưu các link audio

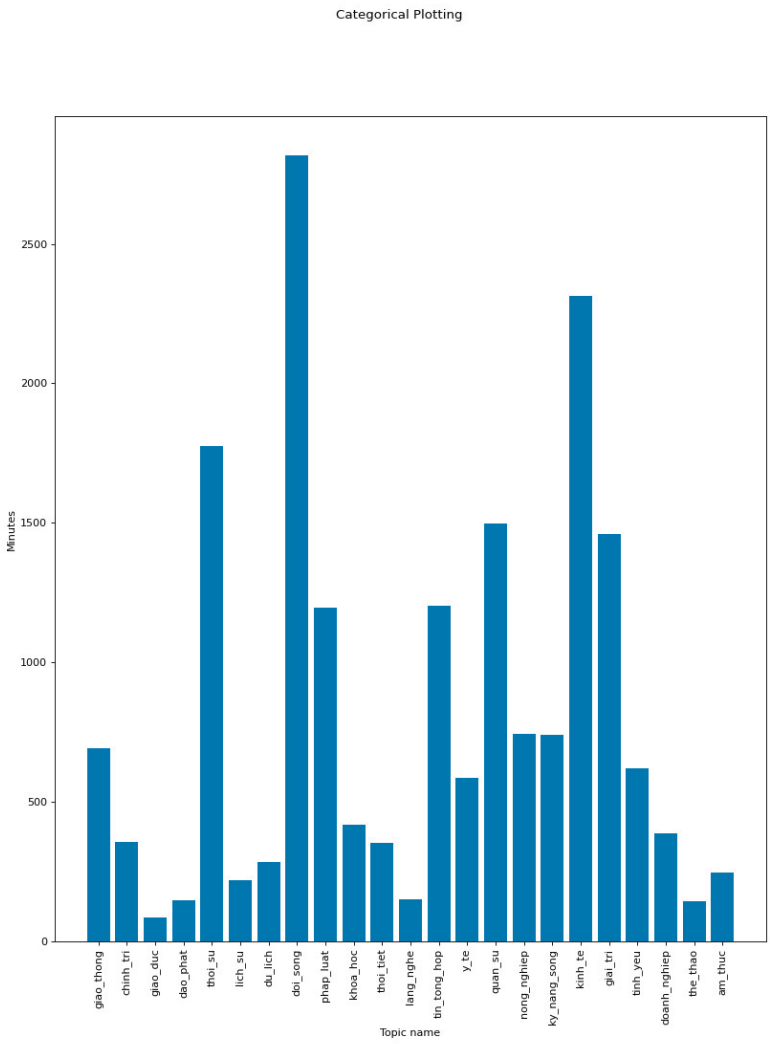
- Cách thu
  - Thu link audio rời rạc, thực hiện **thủ công**.
  - Thu từ link của một playlist hoặc query search. Sử dụng tools ở link github sau: github.com/phamvandan/preparespeech. **Đặc biệt chú ý:** khi sử dụng **query hoặc playlist urls** phải chắc chắn các urls trong đấy đều tốt, nếu không **phải có bước lọc bỏ** sau khi thu được các urls từ tools trên.
- Cách thức lưu trữ
  - Click vào link sau https://docs.google.com/spreadsheets.
  - Duplicate và tạo một sheet cho mình.
  - Ví dụ có thể xem ở sheet Dan
  - Các chủ đề và thời gian đã collect được xem ở sheet Topic, có thể bổ sung thêm các chủ đề mới

3 Số lượng audio theo chủ đề hiện tại, các bạn có thể căn cứ để lựa chọn chủ đề phù hợp đảm bảo sự cân bằng

fx | Topic names

	A	B	C	D	E
1	Topic names	Link saved	Total durations	307,0333333	hours
2	giao_thong	69	692		
3	chinh_tri	52	355		
4	giao_duc	49	84		
5	dao_phat	30	147		
6	thoi_su	263	1775		
7	lich_su	15	220		
8	du_lich	135	284		
9	doi_song	91	2816		
10	phap_luat	56	1195		
11	khoa_hoc	39	417		
12	thoi_tiet	130	353		
13	lang_nghe	17	150		
14	tin_tong_hop	236	1202		
15	y_te	70	585		
16	quan_su	137	1498		
17	nong_nghiep	73	742		
18	ky_nang_song	54	739		
19	kinh_te	307	2314		
20	giai_tri	58	1459		
21	tinh_yeu	83	619		
22	doanh_nghiep	37	386		
23	the_thao	55	145		
24	am_thuc	40	245		

Hình 1: Các chủ đề nhóm đã thu



Hình 2: Biểu đồ thống kê thời lượng audio đã thu của các chủ đề