Exploratory data analysis of labels

Gyuseung Baek

2017. 05. 11.

Labels

- 총 라벨의 개수 = 4716

- 거의 빈도수에 따라 정렬되어있다. (correlation : 0.9999)

label_id	label_name
0	Games
1	Vehicle
2	Video game
3	Concert
4	Car
5	Dance
6	Animation
7	Musician
8	Football
9	Music video
10	Animal

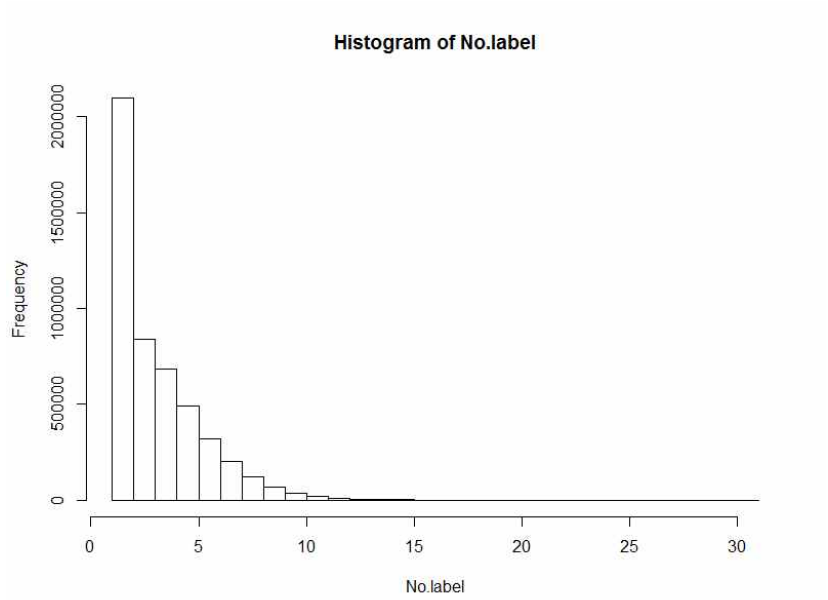
Distribution

- 전체 비디오 개수 : 4,906,660
- 전체 라벨의 수 : 16,639,392

누적 합 %	라벨 수 (%)
50	58 (1.23)
60	107 (2.27)
70	201 (4.26)
80	398 (8.44)
90	996 (21.1)
95	1870 (39.7)
99	3802 (80.6)

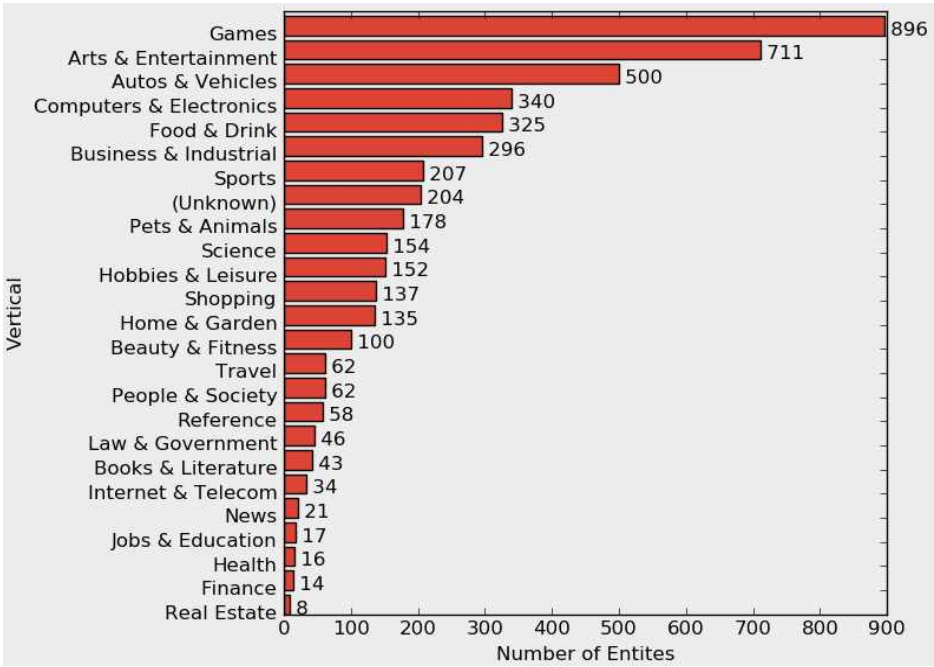
Length

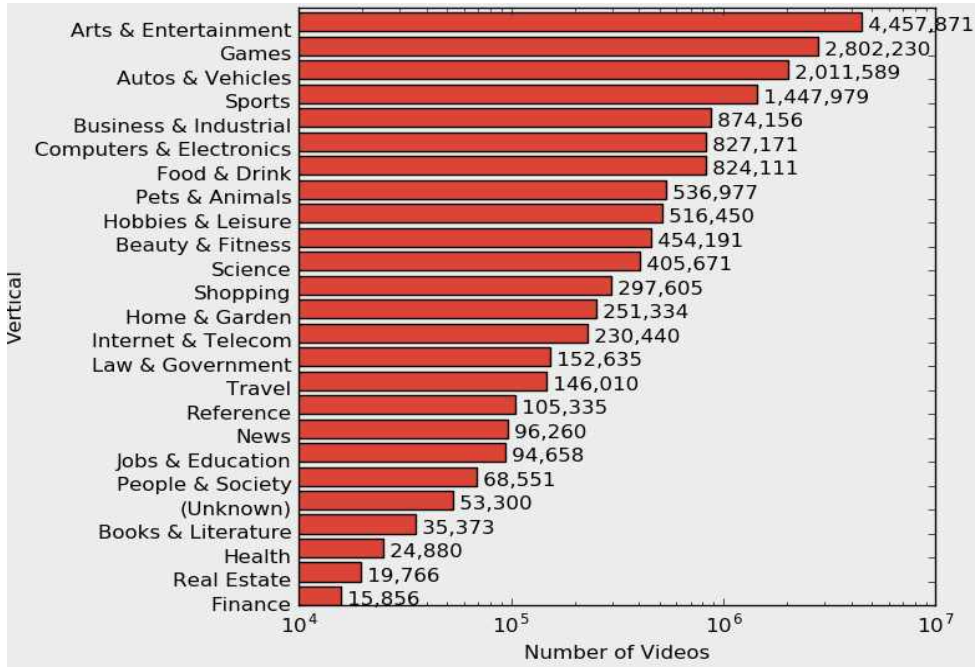
- 한 자료가 가지고 있는 라벨 개수의 평균: 3.39
- 최댓값 : 31
  - Vehicle, Nature, Aircraft, Airplane, Building, Aviation, Airport, … (항공기)



Topic

- https://research.google.com/youtube8m/index.html





Topic

- 24개의 상위 class 존재
- Vertical1, Vertical2, Vertical3, ….
- 상위 클래스 개수

0개	1개	2개	3개
204	3782	691	39

Topic

- 상위 클래스가 없는 label (204개)
  - 비디오 개수 : 53,300개
  - Erhu, The King of Fighters '98, Majorette
  - Erhu <- Arts & Entertainment에 넣을 수 있지 않을까?
  - https://www.youtube.com/watch?v=Nmb7gXd57Rk

Topic

- 상위 클래스가 여러 개인 label (730개 – 15.5%)
  - ex. Fish, Tree, Train station, Water, Parade,…

Fish	Business & Industrial	Pets & Animals	Science
Tree	Business & Industrial	Home & Garden	Science
Train station	Business & Industrial	Reference	Travel

- Tree video
  - https://www.youtube.com/watch?v=AOEadPNXSz8 (Business & Industrial?)
  - https://www.youtube.com/watch?v=XQziKp4NvGs (Science?)

Topic

- 각 자료별 topic의 개수

토픽 수	자료 수	토픽 수	자료 수
0	5,139	7	6,208
1	2,408,859 (49%)	8	936
2	1,439,749 (29%)	9	116
3	648,056 (13%)	10	29
4	284,780 (6%)	11	6
5	86,992 (2%)	12	
6	25,789	13	1

- label수 31개인 자료 : topic이 9개
- topic이 13개인 자료 : label이 12개

Confidence

- 신뢰도 = 1인 규칙 : 202개 (전체 규칙 : 11M)

신뢰도(≥) %	규칙 수
90	3144
95	2114
99	781
100	202

- 신뢰도 = 1인 규칙 중, 지목 대상이 자주 되는 라벨
  - Vehicle, Car, Animal, Motosport, …

- Car를 지목한 라벨 : Jeep Grand Cherokee, Nissan 240sx, ….
- Animal을 지목한 라벨 : Dressage, Dog agility, Bull riding, …