Exploratory data analysis of labels

Gyuseung Baek

2017. 05. 04.

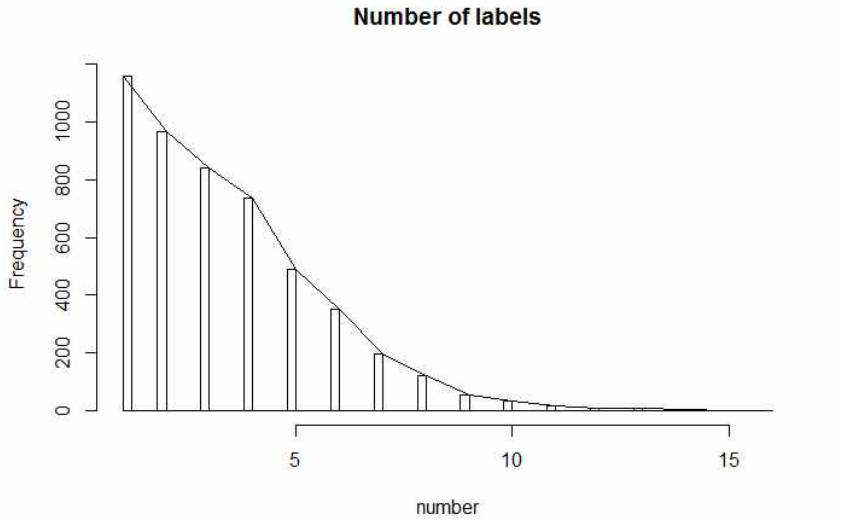
Labels

- 총 라벨의 개수 = 4717
- 빈도수에 따라 라벨이 정렬된 것으로 추측됨
- 분석을 위해 5,000개의 자료만 추출

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

Length

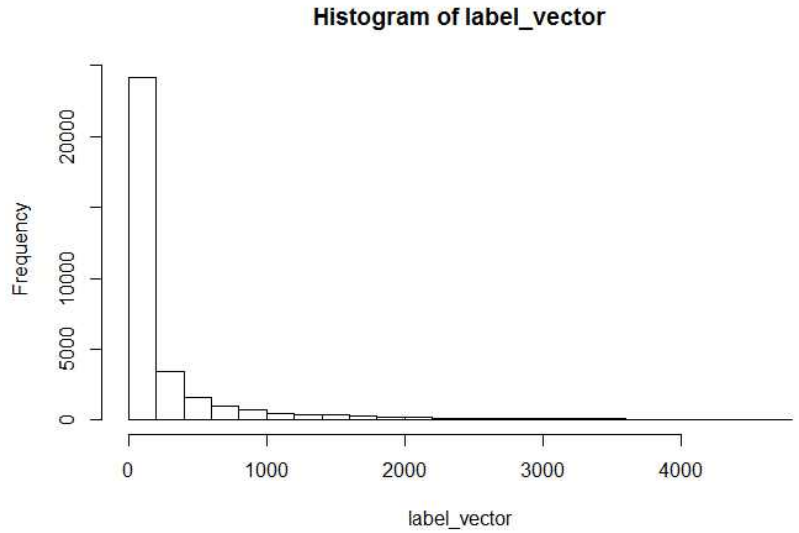
- 한 자료가 가지고 있는 라벨 개수의 평균: 3.39
- 최댓값 : 16
  - Food, Cooking, Recipe, Cuisine, Cooking show kitchen, Eating, Baking, Meal, Bread, Cheese, Grilling, Pizza, Breakfast, Sandwich



Distribution

- 전체 라벨의 수 : 34124

누적 합 %	라벨 수 (%)
50	56 (1.19)
60	103 (2.18)
70	191 (4.05)
80	372 (7.89)
90	877 (18.6)



duplicated rabels

- 80 개의 라벨이 중복되어있다.
- 모든 중복된 라벨은 두번씩만 나와있다.
- 중복된 라벨의 종류
  - Angry Birds, Asphalt, Drums, Gothic, etc.

39	Drums
44	Drums

Confidence

- 서로 다른 라벨 간의 관계를 확인하고자 함

- Label A -> Label B 의 신뢰도 :

  (Label A, B를 동시에 포함하는 자료의 수) / (Label A를 포함하는 자료의 수)

- 신뢰도 = 1인 규칙 : 5442개

  -> 라벨들 간에 강한 상관관계가 존재한다!

- 이 중 지지도가 높은 규칙
  - Motorsport => Vehicle (4%) , PC game => Games (1%)

Confidence

- 신뢰도 = 1인 규칙 중, 지목 대상이 자주 되는 라벨 (대분류일 확률이 높은 라벨)
  - Vehicle(1), Games(0), Video game(2), Car(4), Food(12), ….

- Car를 지목한 라벨 : Dump truck, Car wash, Audi, Driving, …
  McDonald's, Air force, Nintendo 3Ds, ….

- 지지도 : 0.001 이상인 규칙에서 Car를 지목한 라벨 (17개):
  Television, Warcraft, Manga, Disc jockey, Basketball, Engine, Motorcycling,
  Wedding, Tablet computer, Cooking show, Photography, Rallying, Basketball moves,
  Horse, Track, Album, Skateboard