Домашнее задание №2

Виталий Емельянов

10 мая 2016 г.

1 Использованные алгоритмы

1.1 Simhash

Для вычисления симхеша используется следующий алгоритм:

- каждый документ разбивается на слова
- от каждого слова берется питоновский метод __hash__
- вспомогательный вектор v длины 64 изначально равен $[0, 0, \ldots, 0]$
- в цикле по всем словам документа:
  - v[i] увеличивается на 1, если i-й бит хеша слова равен 1
  - v[i] уменьшается на 1, иначе
- значения битов симхеша получаются в соответствии со знаками элементов v

Реализация функции simhash() находится в файле simhash.py, вычисление симхешей коллекции документов реализовано в файле preparing_data.py

1.2 Расстояние в битах

Для того, чтобы измерить расстояние в битах между двумя симхешами используется расстояние Хэмминга

```
def distance(a, b, hashbits=64):
    x = (a ^ b) & ((1 << hashbits) - 1)
    total = 0
    while x:
        total += 1
        x &= x-1
    return total
```

Реализация функции distance() находится в файле simhash.py

1.3 Кластеризация

В качестве алгоритма кластеризации был выбран алгоритм, предложенный в условии: все документы сортируются по количеству слов, отсекаются “короткий” и “длинный” хвосты документов, которые заведомо не пройдут проверку по длине. Реализация кластеризации находится в файле clustering.py

2 Результаты

Код построения гистограмм и нахождения топ-10 находится в файле report.py

2.1 Распределения по размерам групп

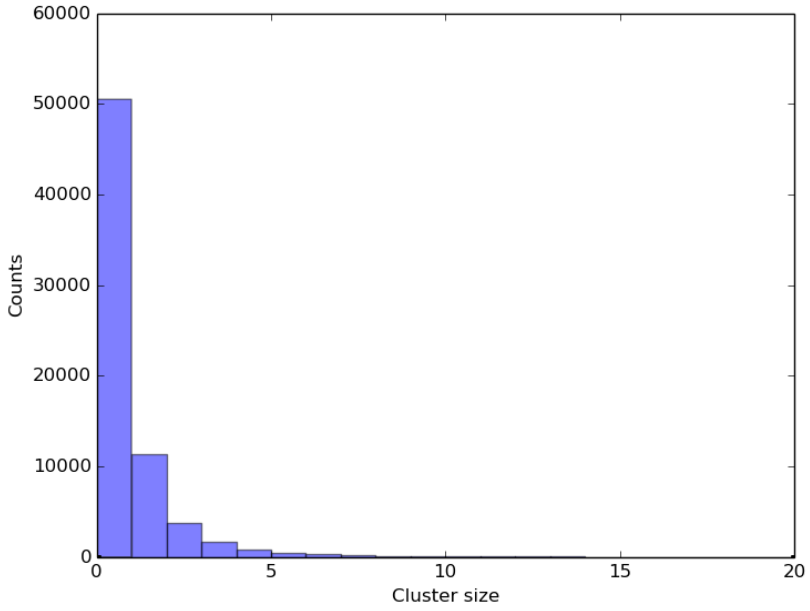

Рис. 1: Гистограмма распределения по размерам групп для n=5

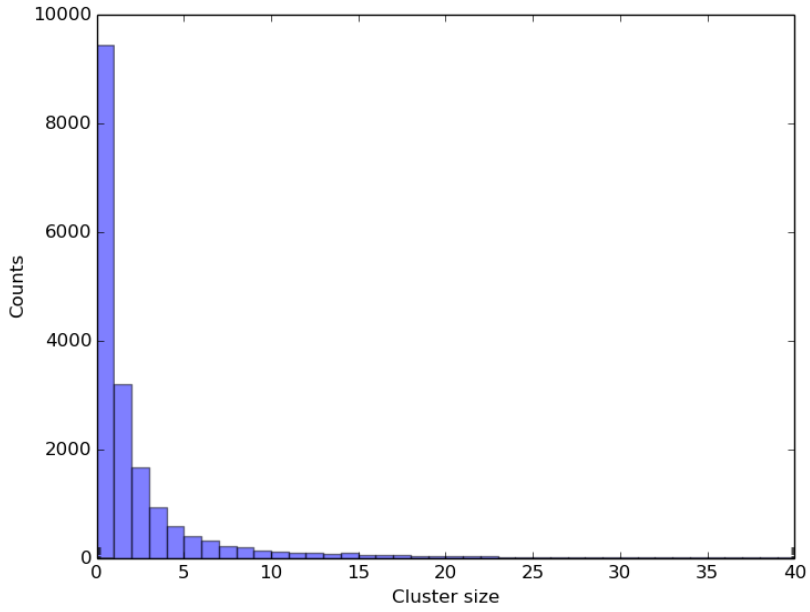

Рис. 2: Гистограмма распределения по размерам групп для n=10

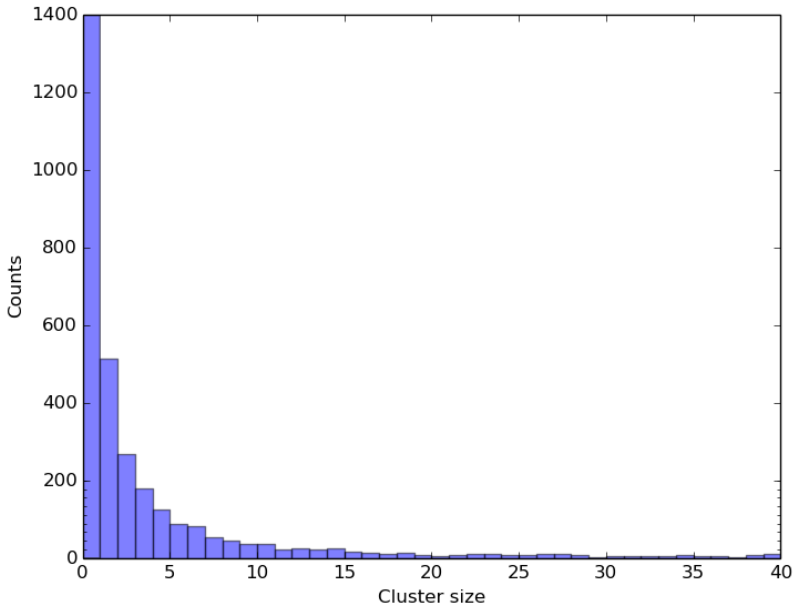

Рис. 3: Гистограмма распределения по размерам групп для n=15

Очевидно, что чем больше n, тем больше размеры кластеров. Форма распределений подтверждает это.

2.2 Топ-10 главных с самыми большими группами

n=5

http://simple.wikipedia.org/wiki/Polincove 1772
http://simple.wikipedia.org/wiki/Barou-en-Auge 817
http://simple.wikipedia.org/wiki/24_September 686
http://simple.wikipedia.org/wiki/Ascros 630
http://simple.wikipedia.org/wiki/Nebo,_Kentucky 520
http://simple.wikipedia.org/wiki/Washington,_Arkansas 509
http://simple.wikipedia.org/wiki/Warminster 439
http://simple.wikipedia.org/wiki/Sainte-Vertu 410
http://simple.wikipedia.org/wiki/Peyzieux-sur-Sa%C3%B4ne 405
http://simple.wikipedia.org/wiki/Dognen 392
http://simple.wikipedia.org/wiki/Saint-Pierre-Saint-Jean 332

Видно, что в топ-10 попали короткие статьи. В основном это кластеры однотипных коротких статей про французские коммуны, два кластера, объединяющих города США, кластер из дат (статья про 24 сентября в качестве главной). Причем наличие в топе большого количества статей про французские коммуны можно объяснить тем, что они отличаются, во-первых, содержанием, во-вторых, тем, что у части статей есть таблица справа с некоторыми географическими данными, а у части ее нет.

n=10

http://simple.wikipedia.org/wiki/La_Fert%C3%A9-Milon 3156
http://simple.wikipedia.org/wiki/Drakesboro,_Kentucky 1469
http://simple.wikipedia.org/wiki/Oxbow_lake 1387

http://simple.wikipedia.org/wiki/Saint-Victor,_Ard%C3%A8che 1293
http://simple.wikipedia.org/wiki/Heterodontosaur 1112
http://simple.wikipedia.org/wiki/Royston_Drenthe 1052
http://simple.wikipedia.org/wiki/Decilitre 812
http://simple.wikipedia.org/wiki/Guild 769
http://simple.wikipedia.org/wiki/Wim_Jansen 763
http://simple.wikipedia.org/wiki/Besan%C3%A7on_R.C. 749
http://simple.wikipedia.org/wiki/Gram_(mythology) 738

В топ-10 снова попало несколько кластеров французских коммун, пара кластеров коротких статей о футболистах (Royston Drenthe и Wim Jansen - главные), отличающихся друг от друга в оформлении (наличие таблицы со статистикой игрока). Кластер, где главной является статья о футбольном клубе Besançon R.C. содержит в себе еще несколько статей о футбольных клубах, но большинство из них отличается по тематике.

n=15

http://simple.wikipedia.org/wiki/Triple-Zero 5862
http://simple.wikipedia.org/wiki/Aign%C3%A9 3476
http://simple.wikipedia.org/wiki/La_Scala 3070
http://simple.wikipedia.org/wiki/Weblogs 2708
http://simple.wikipedia.org/wiki/Reforesting 2522
http://simple.wikipedia.org/wiki/Roman_%C4%8Cechm%C3%A1nek 2522
http://simple.wikipedia.org/wiki/The_Daily_Pennsylvanian 2483
http://simple.wikipedia.org/wiki/Parallelism_(grammar) 2469
http://simple.wikipedia.org/wiki/Taufiq_Rafat 2461
http://simple.wikipedia.org/wiki/Cartographer 2415
http://simple.wikipedia.org/wiki/Sark 2321

В топ-10 опять же попал кластер статей о французских коммунах во главе со статьей об Aigné. При n=15 состав остальных кластеров довольно трудно объяснить.

2.3 Распределение расстояний между симхешами

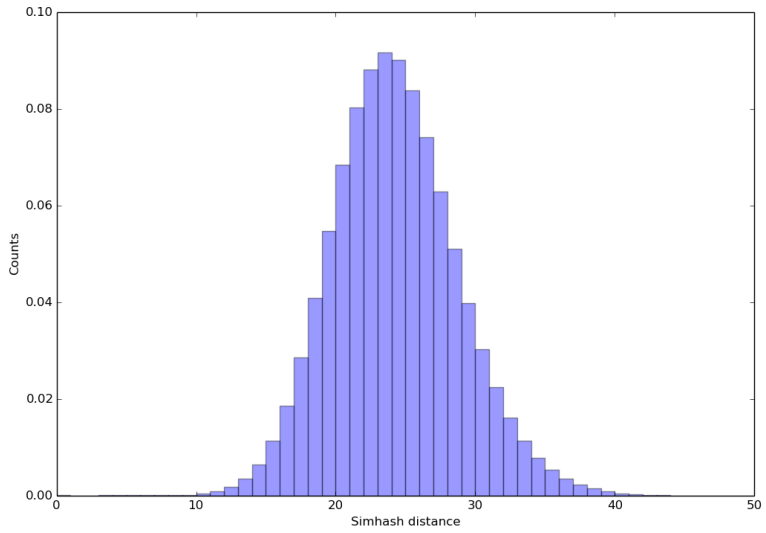

Рис. 4: Гистограмма распределения расстояний между симхешами для n=15

Распределение похоже на нормальное. Посчитаем выборочное среднее и выборочную дисперсию:

$$\mu = 23.7484577666$$

$$\sigma = 4.50027996533$$

Построим на том же графике теоретическую плотность нормального распределения с найденными параметрами μ и σ:

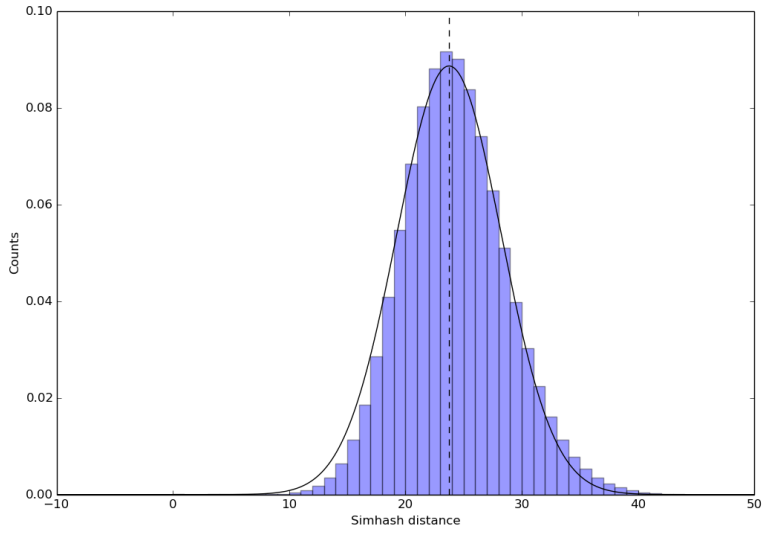

Рис. 5: Гистограмма распределения расстояний между симхешами для n=15 с наложенной теоретической плотностью $\mathcal{N}(23.75, 4.5)$

Теоретическая плотность довольно неплохо ложится на гистограмму, хотя и видно, что распределение ассиметрично с коэффициентом ассиметрии равным примерно 0.25.