7506-2020-2 Parcialito de Hashing y LSH

Santiago Locatelli

TOTAL POINTS

90 / 100

QUESTION 1

1 Hashing y LSH 90 / 100

Item 2)

✓ **- 5 pts** **No hay desarrollo sobre como obtiene las funciones de minhash**

✓ **- 5 pts** **No hay desarrollo sobre como obtiene la funcion de hashing universal o como se parametriza la misma.**

1 Se está aproximando. Además, se pide que la segunda probabilidad sea menor a 0,01 ...

2 Falta explicación. ¿Se hashea todo el tweet? ¿Sobre qué se calcula el minhash?

Santiago Locatelli, 104107

“Queremos que si la semejanza entre tweets es mayor o igual a 0.55 tengamos más de un 70% de probabilidad de que sean candidatos. Por otro lado, si la semejanza es menor o igual a 0.05 queremos tener menos de un 1% de probabilidad de que sean candidatos a ser comparados. “

A)

Teniendo que la Distancia de Jaccard = 1 - Semejanza de Jaccard, podemos decir que:

- d1 = 0.45
- d2 = 0.95

Y además:

- p1 > 0.70
- p2 < 0.01

También, teniendo la fórmula: p = 1 - (1 - (1 - d)^r)^b, debemos pedir que se cumpla:

- 0.70 < 1 - (1 - (1 - d1)^r)^b, es decir, 0.70 < 1 - (1 - 0.55^r)^b
- 0.01 > 1 - (1 - (1 - d2)^r)^b, es decir, 0.01 > 1 - (1 - 0.05^r)^b

Probando números, llegó a que b = 4 y r = 2 se cumplen estas condiciones:

- 1 - (1 - 0.55^2)^4 = 0.76 1
- 1 - (1 - 0.05^2)^4 = 0.01

Entonces si se tiene una base de 10.000 nuevos tweets vamos a tener:

- Falsos negativos = 10.000 * (1 - p1) = 10.000 * 0.24 = **2400**
- Falsos positivos = 10.000 * p2 = 10.000 * 0.01 = **100**

B)

Primero, como tenemos b = 4 y r =2, vamos a necesitar 8 minhashes. Esto quiere decir que vamos a usar 4 grupos y 2 minhashes por grupo.

Habiendo hecho 4 pares de minhashes, lo que sigue es calcular los 8 minhashes de un tweet, y suponiendo que mh1 y mh2, son los minhashes de la tabla1, entonces se podría pensar a estos minhashes como un vector, por ejemplo, [mh1, mh2], y a este resultado aplicarle un función universal de hashing, más específicamente una función de hash de vectores. Esto se aplica para los 4 pares de minhashes, es decir, que habría 4 vectores para cada tweet. De esta forma queda una única tabla. Cabe aclarar que esto se debe calcular para cada tweet. 2

C)

Ya teniendo toda la base tweets procesada, se debe realizar el procesamiento descrito en el punto B, con el tweet que queremos clasificar. Entonces debemos fijarnos con qué otros elementos de la tabla colisionó. Y utilizando la Semejanza de Jaccard, con el criterio

definido en el enunciado, es decir, si la semejanza es mayor o igual a 0.73 debemos definir si es fake news. Entonces, si la semejanza de jaccard entre el elemento query y el elemento con el que colisionó en la tabla es mayor a la definida, decimos que es fake news.

1 Hashing y LSH 90 / 100

Item 2)

✓ **- 5 pts** **No hay desarrollo sobre como obtiene las funciones de minhash**

✓ **- 5 pts** **No hay desarrollo sobre como obtiene la funcion de hashing universal o como se parametriza la misma.**

1. Se está aproximando. Además, se pide que la segunda probabilidad sea menor a 0,01 ...
2. Falta explicación. ¿Se hashea todo el tweet? ¿Sobre qué se calcula el minhash?