Jaccard

Jordy Joaquin Cuan Robledo [1], Eduardo Cantorán Flores [1],
Rafael Pérez Aguirre [1]
[1] Facultad de Ciencias de la Computación, Benemérita Universidad
Autónoma de Puebla, México

Jaccard nos sirve para calcular la distancia entre dos conjuntos, se le conoce como la distancia de Jaccard. Un forma de representar esto, es usar la distancia de Jaccard para determinar qué tan parecido son dos documentos. Este método utiliza lo que comúnmente se conoce como "bag of words" (bolsa de palabras, en español), la cual es simple, pero es suficiente para muchas aplicaciones.

Algunos ejemplos en los que podemos utilizar Jaccard son los siguientes:

· Dadas dos tareas (reportes), podemos detectar si una es plagio de la otra.
· Google utiliza para poder eliminar listas duplicadas o alternativas.

Conjuntos y distancias

Un conjunto es una colección de objetos, estos elementos están separados por comas y dentro de los llaves { }. No importa el orden los elementos, por ejemplo {a, b} = {b, a}.
Para calcular la distancia d(A, B) entre dos conjuntos, tenemos las siguientes propiedades:

· Es pequeña si A y B están cerca
· Es grande si están alejados.
· Usualmente es 0 si son iguales.
· Tiene una representación entre [0, ∞].

Para calcular la similaridad s(A, B) entre dos conjuntos, tenemos las siguientes propiedades:

· Es grande si los objetos de A y B están cerca
· Es pequeña si están lejos.
· Usualmente es 1 si son iguales
· Está en el rango de [0, 1]

También las podemos convertir de dos formas:

· d(A, B) = 1 - s(A, B)
· d(A, B) = $\sqrt{s(A,\ A) + s(B,\ B) - 2s(A,\ B)}$.

Similitud Jaccard

Para calcular la similitud de dos conjuntos con Jaccard, basta con aplicar la siguiente fórmula:

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Se desarrolló un programa en el lenguaje de programación Python, el programa toma un conjunto de texto en diferentes idiomas (Ingles, Italiano, Frances) y muestra la similitud que existe entre estos lenguajes.

Primero importamos la libreria string, está la ocuparemos para poder eliminar los caracteres no deseados. Posteriormente se abre el archivo con los textos en los diferentes idiomas.

```
import string

idiomas = open('./Idiomas.txt', 'r')
```

Se declaran los conjuntos para cada idioma y también la lista con el nombre de los idiomas que contiene el texto.

```
lista_idiomas = ['Italian', 'French', 'English']
Italian = set()
French = set()
English = set()
```

Leemos linea por linea el archivo y verificamos de que idioma se trata esa línea, una vez detectada, se procede a guardar su contenido en su respectiva lista.

```
for linea in idiomas:

    #Si es idioma Italiano
    if linea.split()[0] == lista_idiomas[0]:
        Italian = Italian | set(linea.split()[1:])
    #Si es idioma Frances
    elif linea.split()[0] == lista_idiomas[1]:
        French = French | set(linea.split()[1:])
    #Entonces es Ingles
    else:
        English = English | set(linea.split()[1:])
```

Aplicamos la fórmula de Jaccard entre cada idioma y multiplicamos por 100 para poder mostrar el porcentaje de similitud.

```
#Calculamos la similitud entre Frances e Italiano
print "<<<<<<<Porcentaje de Similitud Frances | Italiano>>>>>>> "
similitud = len(Italian & French)/ float(len(Italian | French))
print similitud*100,"%"
#Calculamos la similitud entre Frances e Ingles
print "<<<<<<<Porcentaje de Similitud Frances | Ingles>>>>>>> "
similitud = len(English & French)/ float(len(English | French))
print similitud * 100, "%"
#Calculamos la similitud entre Ingles e Italiano
print "<<<<<<<Porcentaje de Similitud Ingles | Italiano>>>>>>> "
similitud = len(Italian & English)/ float(len(Italian | English))
print similitud *100, "%"
```