# IIR Homework 4 方郁文

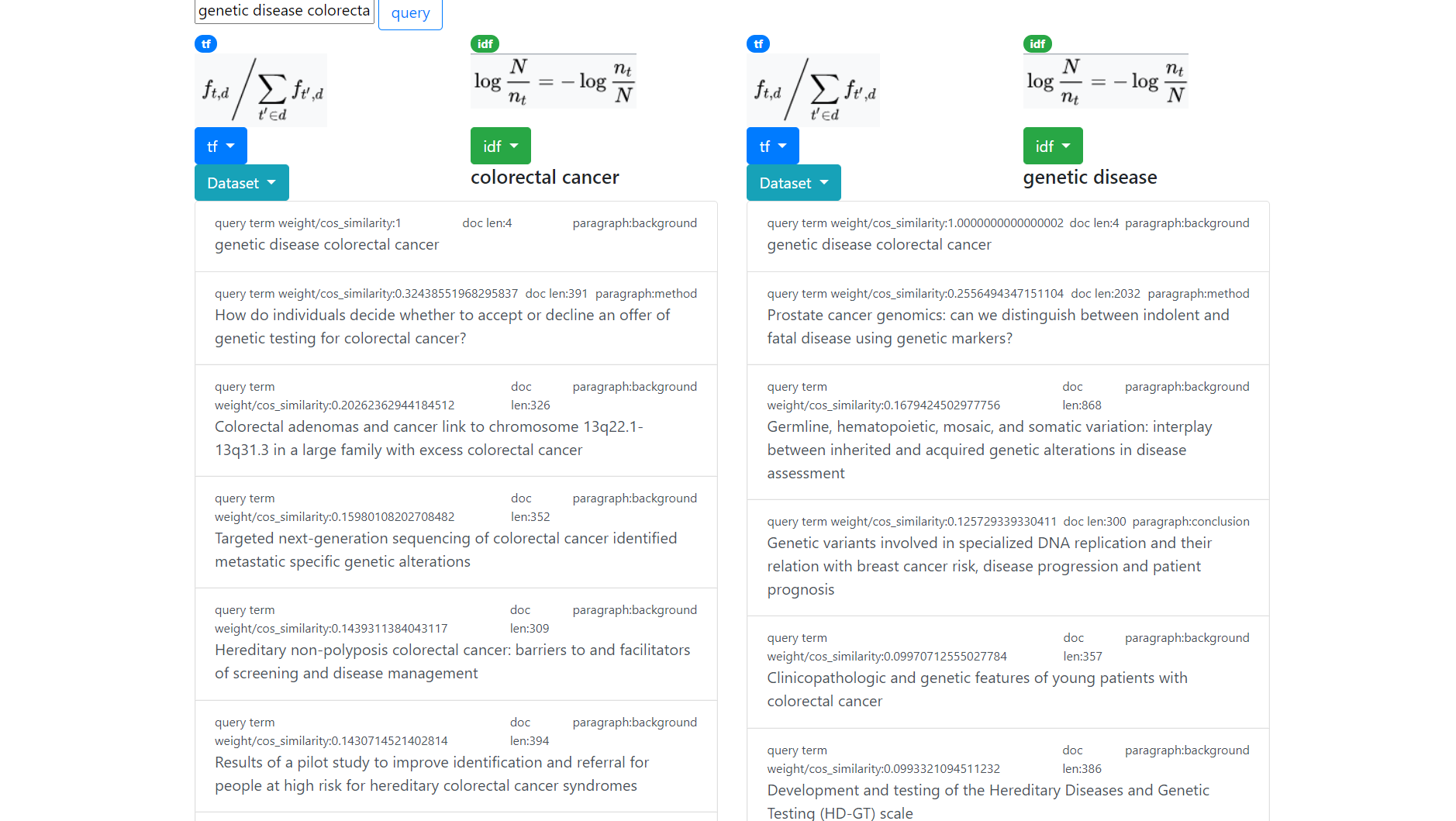
## 系統環境

* 作業系統 ： Ubuntu 20.04 LTS
* 程式語言 ： Python 3.7.9
* 網站架構 ： Django 3.1.1
* 讀取csv檔 ： Pandas 1.1.3
* 文件來源 : BMC journel

## 實作方法

* 利用python實作3種tf以及3種idf方法
* 從BMC journel抓取關於大腸癌以及遺傳疾病的document作為dataset
* Document具有以下paragraph: background, method, result, conclusio
* 透過tf idf，將每個document以每個term的weight表示成一個vector，以利後面計算cosine similarity
* 設計介面，讓使用者可以自由選擇tf以及idf的組合
* 提供兩種排序方法
  1. 將query term在各document作為指標進行排序
  2. 將query term當成小的document，和每個document比cosine similarity

## 介面展示



分成兩邊以利一次展示兩種組合

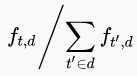
Dataset按鈕: 決定使用的dataset

tf及idf按鈕: 分別決定使用的tf-idf演算法，並預覽其數學式

## 實驗

1. 將query term在各document作為指標進行排序，會使idf影響排序結果

2. 將query term在各document作為指標進行排序，可使用tf為

來減少長的文件在排序上的優勢

3. 將query term當成小的document，把query term的vector拿去和各document中的background, method, result, conclusion比cosine similarity，發現較多情況中，query term與background最為相近