Information Retrieval

Homework 2

1. **Introduction**

The objective of this assignment is to calculate the term weights for the tokens that occur in each document in the collection. To execute the program, it should be sufficient to type:

python3 calcwts.py input-directory output-directory

where, the input-directory contains the original files (i.e., the unprocessed files directory) and the output-directory contains all generated output files.

The program also needs a temporary folder with the name "temp" to store intermediate results after preprocessing

2. **Input**

The input directory contains a set of HTML documents that have to be processed

3. **Output**

The goal is to build one output file per input file. In the output file, there would be one line per token that survives preprocessing. That line would contain the token and the token weight.

4. **Preprocessing**

I have used the processor from Homework1 to process the documents. After loading the file contents, BeautifulSoup library is used to parse the HTML documents and extract information in pure text format. Any special characters are removed using the concept of regular expressions. NLTK library is used to tokenize the file content which results in a list containing individual words from each particular document.

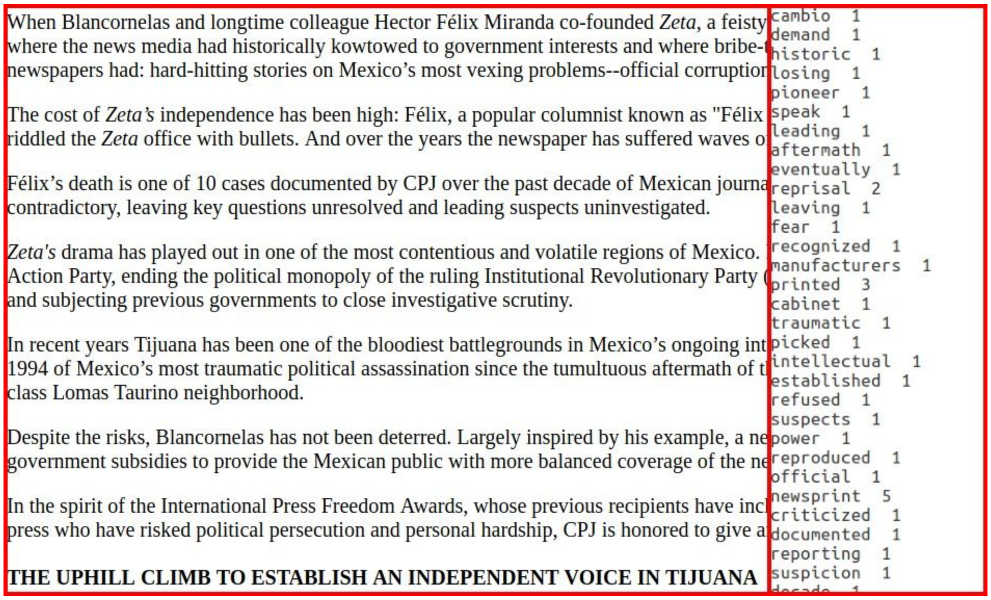
The initial processor is further extended to do the following:

a) Remove words of length 1

b) Remove all the words occurring in the file stopwords.txt

c) Remove words occurring only once in the entire corpus

Figure below shows two HTML files before and after preprocessing.



When Blancornelas and longtime colleague Hector Félix Miranda co-founded *Zeta*, a feisty
where the news media had historically kowtowed to government interests and where bribe-
newspapers had: hard-hitting stories on Mexico’s most vexing problems--official corruptio

The cost of *Zeta’s* independence has been high: Félix, a popular columnist known as "Félix
riddled the *Zeta* office with bullets. And over the years the newspaper has suffered waves o

Félix’s death is one of 10 cases documented by CPJ over the past decade of Mexican journa
contradictory, leaving key questions unresolved and leading suspects uninvestigated.

Zeta's drama has played out in one of the most contentious and volatile regions of Mexico.
Action Party, ending the political monopoly of the ruling Institutional Revolutionary Party
and subjecting previous governments to close investigative scrutiny.

In recent years Tijuana has been one of the bloodiest battlegrounds in Mexico’s ongoing int
1994 of Mexico’s most traumatic political assassination since the tumultuous aftermath of t
class Lomas Taurino neighborhood.

Despite the risks, Blancornelas has not been deterred. Largely inspired by his example, a ne
government subsidies to provide the Mexican public with more balanced coverage of the ne

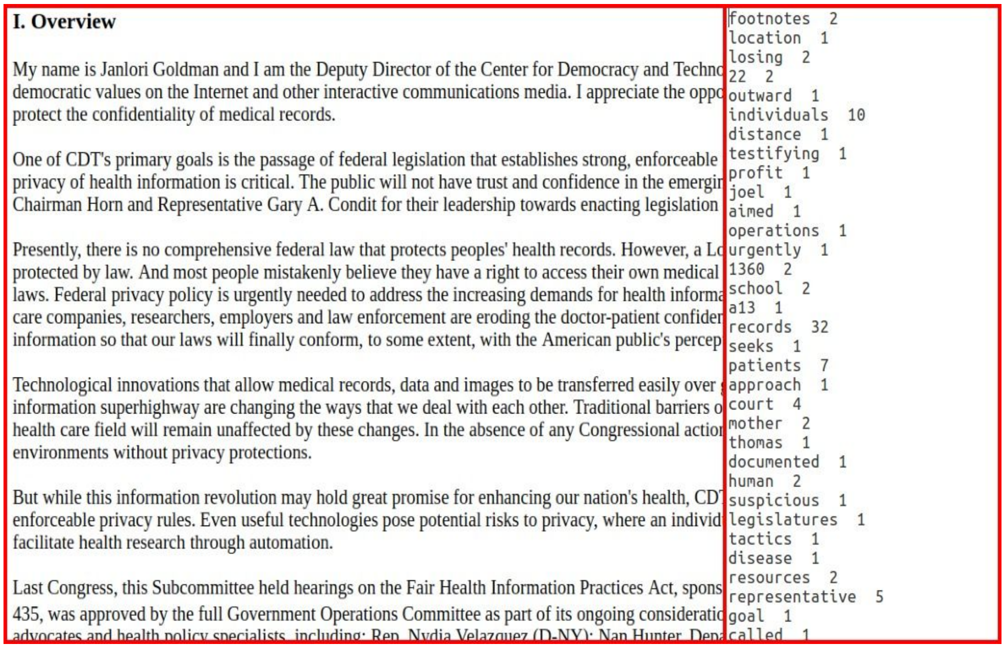
In the spirit of the International Press Freedom Awards, whose previous recipients have incl
press who have risked political persecution and personal hardship, CPJ is honored to give a

THE UPHILL CLIMB TO ESTABLISH AN INDEPENDENT VOICE IN TIJUANA

```
cambio  1
demand  1
historic  1
losing  1
pioneer  1
speak  1
leading  1
aftermath  1
eventually  1
reprisal  2
leaving  1
fear  1
recognized  1
manufacturers  1
printed  3
cabinet  1
traumatic  1
picked  1
intellectual  1
established  1
refused  1
suspects  1
power  1
reproduced  1
official  1
newsprint  5
criticized  1
documented  1
reporting  1
suspicion  1
decade  1
```

Fig 1 : Left box has part of HTMl1 before preprocessing

Right box has part of tokenized words and their frequencies



I. Overview

My name is Janlori Goldman and I am the Deputy Director of the Center for Democracy and Techn
democratic values on the Internet and other interactive communications media. I appreciate the opp
protect the confidentiality of medical records.

One of CDT's primary goals is the passage of federal legislation that establishes strong, enforceable
privacy of health information is critical. The public will not have trust and confidence in the emergi
Chairman Horn and Representative Gary A. Condit for their leadership towards enacting legislation

Presently, there is no comprehensive federal law that protects peoples' health records. However, a L
protected by law. And most people mistakenly believe they have a right to access their own medical
laws. Federal privacy policy is urgently needed to address the increasing demands for health inform
care companies, researchers, employers and law enforcement are eroding the doctor-patient confide
information so that our laws will finally conform, to some extent, with the American public's percep

Technological innovations that allow medical records, data and images to be transferred easily over
information superhighway are changing the ways that we deal with each other. Traditional barriers o
health care field will remain unaffected by these changes. In the absence of any Congressional actio
environments without privacy protections.

But while this information revolution may hold great promise for enhancing our nation's health, CD
enforceable privacy rules. Even useful technologies pose potential risks to privacy, where an individ
facilitate health research through automation.

Last Congress, this Subcommittee held hearings on the Fair Health Information Practices Act, spons
435, was approved by the full Government Operations Committee as part of its ongoing consideratio
advocates and health policy specialists, including: Rep. Nydia Velazquez (D-NY); Nan Hunter, Dep

```
footnotes  2
location  1
losing  2
22  2
outward  1
individuals  10
distance  1
testifying  1
profit  1
joel  1
aimed  1
operations  1
urgently  1
1360  2
school  2
a13  1
records  32
seeks  1
patients  7
approach  1
court  4
mother  2
thomas  1
documented  1
human  2
suspicious  1
legislatures  1
tactics  1
disease  1
resources  2
representative  5
goal  1
called  1
```

Fig 2: Left box has part of HTML2 before preprocessing

Right box has part of tokenized words and their frequencies

5. Term Weighting

Okapi BM25 function is used to calculate term weights.It is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. Given a query Q, containing keywords q_1, ..., q_n, the BM25 score of a document D is

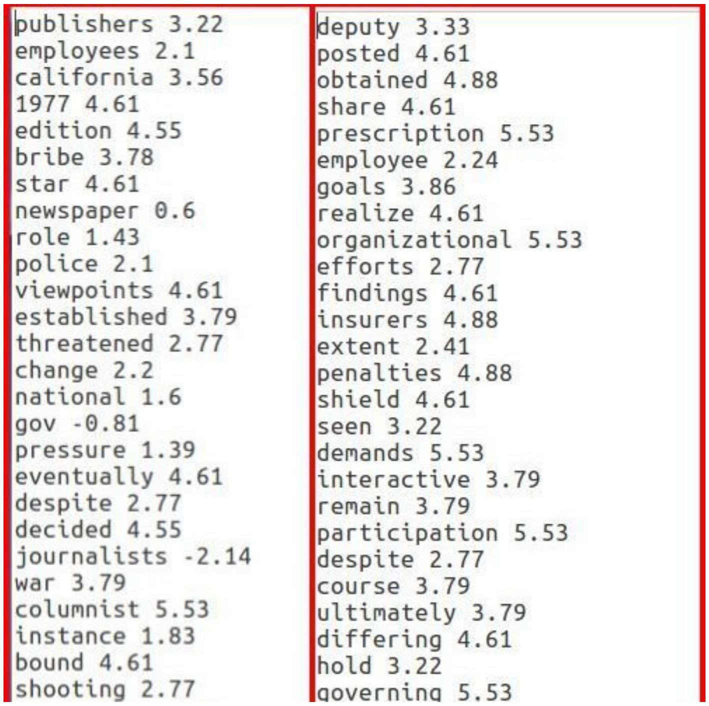
$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

where f(q{i},D)) is term frequency in the document D, |D| is the length of the document D in words, and avgdl is the average document length in the text collection from which documents are drawn.

K1 and b are free parameters whose values have been chosen as 1.2 and 0.75 respectively.

Inverse document frequency(idf) is calculated using the formula log(N/n(q)),

Where N is the total number of documents in the corpus and n(q) is the number of documents containing the word q.Since the above BM25 formula uses length of document D as well as the average length of documents, it also takes the normalization factor into account. In the final output files, the term weights of some words are shown to be negative. This is because the BM25 function calculates term weights of words that occur in more than half of the document collection to be negative.



```
publishers 3.22
employees 2.1
california 3.56
1977 4.61
edition 4.55
bribe 3.78
star 4.61
newspaper 0.6
role 1.43
police 2.1
viewpoints 4.61
established 3.79
threatened 2.77
change 2.2
national 1.6
gov -0.81
pressure 1.39
eventually 4.61
despite 2.77
decided 4.55
journalists -2.14
war 3.79
columnist 5.53
instance 1.83
bound 4.61
shooting 2.77
```

```
deputy 3.33
posted 4.61
obtained 4.88
share 4.61
prescription 5.53
employee 2.24
goals 3.86
realize 4.61
organizational 5.53
efforts 2.77
findings 4.61
insurers 4.88
extent 2.41
penalties 4.88
shield 4.61
seen 3.22
demands 5.53
interactive 3.79
remain 3.79
participation 5.53
despite 2.77
course 3.79
ultimately 3.79
differing 4.61
hold 3.22
governing 5.53
```

Fig 3 : Term weights of a set of words from documents HTML1 and HTML2

6. Evaluation

The program has been tested on different number of files and the graph shown below has been plotted by indexing time as a function of the number of documents. The timings shown is for the entire process of reading from files , preprocessing , calculating term weights and writing to output files.

Tokenization and calculating term weights are the two functions that utilized most of the time.

Also, since I have used a temporary folder to store intermediate results from tokenization, this has added an extra step of reading and writing to files which contributes to the significant increase in time as the number of files increases.

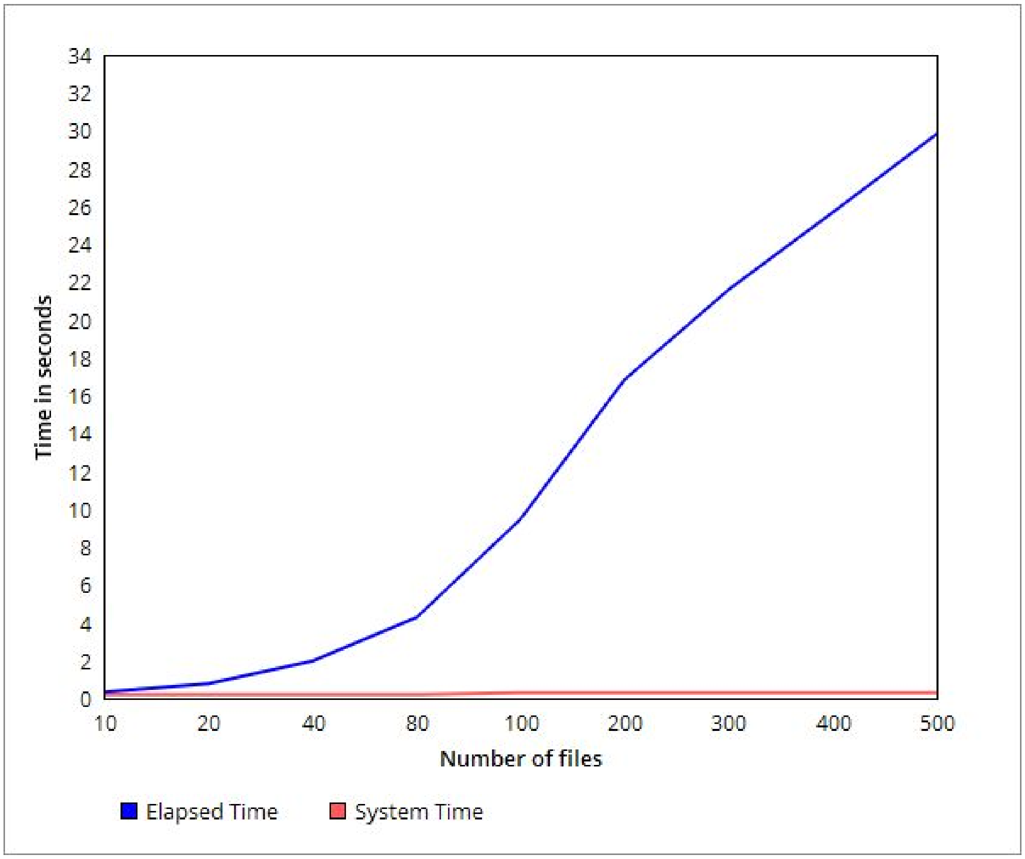


Fig 4: Execution time vs number of files