CT102 Information systems assignment 1

Maxwell Maia

21236277

Question 1 - Pre-processing and tf-idf weighting

i)

Given text:

The ethics of Artificial Intelligence is the branch of the ethics of technology specific to artificially intelligent systems. It is sometimes divided into a concern with the moral behaviour of humans as they design, make, use and treat, artificially intelligent systems, and a concern with the behaviour of machines, in machine ethics.

Standard pre-processing steps:  
1. Case folding (all to lower case)

2. Punctuation removal

3. Stop word removal (using the stop word list: stopwords2.txt)

4. Stemming (using Porter stemming algorithm)

After case folding:

the ethics of artificial intelligence is the branch of the ethics of technology specific to artificially intelligent systems. it is sometimes divided into a concern with the moral behaviour of humans as they design, make, use and treat, artificially intelligent systems, and a concern with the behaviour of machines, in machine ethics.

After punctuation removal:

the ethics of artificial intelligence is the branch of the ethics of technology specific to artificially intelligent systems it is sometimes divided into a concern with the moral behaviour of humans as they design make use and treat artificially intelligent systems and a concern with the behaviour of machines in machine ethics

After stop word removal (using the stop word list: stopwords2.txt)

ethics artificial intelligence branch ethics technology specific artificially intelligent systems sometimes divided into concern moral behaviour humans design make use treat artificially intelligent systems concern behaviour machines machine ethics

After stemming (using Porter stemming algorithm)

ethic artifici intellig branch ethic technolog specif artifici intellig system sometim divid into concern moral behaviour human design make use treat artifici intellig system concern behaviour machin machin ethic

Final result:

ethic artifici intellig branch ethic technolog specif artifici intellig system sometim divid into concern moral behaviour human design make use treat artifici intellig system concern behaviour machin machin ethic

ii)

Calculate tf-idf

First get term frequency (tf).

tf(ethic) = number of times term occurs in document / total terms in document

tf(ethic) = 3 / 29

tf = 3/29

Now for idf

idf = log10 ( 1 + N/df )

N is the number of documents in the collection of documents.

df is the number of documents term occurs in.

“there are 220 documents in the document collection and that the term occurs in 50 of them.”

Therefore

N = 220

df = 50

idf(ethic) = log10 ( 1 + 220/50 )

idf(ethic) = 0.73239…

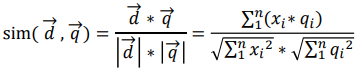
tf-idf = tf \* idf

tf-idf = 0.07576487171

Question 2 - Euclidean dot product (similarity)

Vector A < 0.30, 0.25, 0.1, 0.02, 0.00, 0.11 >

Vector B < 0.35, 0.00, 0.3, 0.11, 0.02, 0.20 >



= 0.744

Question 3 - Page Rank

General formula for page rank:

PR(page) = (1-d)/N + d (PR(T1)/C(T1) + … + PR(Tn)/C(Tn))

d = 0.85

I used a C program to calculate page rank.

//Maxwell Maia, Page Rank calculator

#include <stdio.h>

#include <math.h>

void main()

{

//this program works for 6 pages.

//for a different amount of pages change: the amount of formulas, contents of formula

double numPages = 6;

//variables to store the page ranks of each page

double prA = 1.0 / numPages;

double prB = 1.0 / numPages;

double prC = 1.0 / numPages;

double prD = 1.0 / numPages;

double prE = 1.0 / numPages;

double prF = 1.0 / numPages;

//initially the page ranks are even, this will change as the program runs

//variables to store the page rank of the previous iteration

double prevA = 0.0;

double prevB = 0.0;

double prevC = 0.0;

double prevD = 0.0;

double prevE = 0.0;

double prevF = 0.0;

//set number of outlinks

double cA = 2;

double cB = 3;

double cC = 3;

double cD = 2;

double cE = 4;

double cF = 2;

for (int i = 0; i < 100; ++i)

{

prevA = prA;

prevB = prB;

prevC = prC;

prevD = prD;

prevE = prE;

prevF = prF;

//PAGE RANK FORMULAS

prA = 0.15 / numPages + 0.85 \* (prevB / cB + prevC / cC + prevE / cE);

prB = 0.15 / numPages + 0.85 \* (prevA / cA + prevC / cC + prevE / cE + prevF / cF);

prC = 0.15 / numPages + 0.85 \* (prevA / cA + prevE / cE);

prD = 0.15 / numPages + 0.85 \* (prevB / cB + prevF / cF);

prE = 0.15 / numPages + 0.85 \* (prevC / cC + prevD / cD + prevF / cF);

prF = 0.15 / numPages + 0.85 \* (prevB / cB + prevD / cD + prevE / cE);

}

printf("Printing final page ranks. After 100 iterations...\n");

printf("A: %f\n", prA);

printf("B: %f\n", prB);

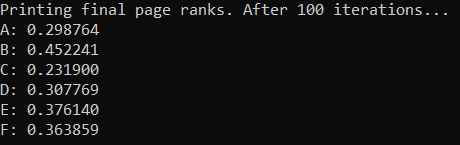
printf("C: %f\n", prC);

printf("D: %f\n", prD);

printf("E: %f\n", prE);

printf("F: %f\n", prF);

}

ii)

Question 4 - Plagiarism declaration

Plagiarism Declaration: “I am aware of what plagiarism is and include this here to confirm that this work is my own”