**System Requirements:**

Windows with Python 3.9.2 installed.

**Coding Language and version:**

Python 3.9.2

**List of files in the application:**

1. porter\_stemmer.py (taken the python version from the link given in assignment)
2. pre\_process.py
3. inverted\_index.py
4. tfidf.py
5. cosine\_sim.py
6. ir\_system.py
7. input.txt (sample input file for checking)
8. stopword.txt (small file of stop words created manually )

**Instructions to Run Program**

**1. PREPROCESSING:**

**python .\pre\_process.py .\input.txt .\pre\_processed.txt .\stopword.txt**

**Expected file format:**

**input.txt**

D1 This is a sample document, it has less document then the first one.

D2 This is a name of second sample document. This one has more sentences than a sample document first.

D3 The name of the sample document is spaced by a TAB character! all docs are on a single line

**pre\_processed.txt**

D1 sampl document less document first

D2 name second sampl document more sentenc than sampl document first

D3 name sampl document space tab charact doc singl line

**2. INVERTED INDEX:**

**python .\inverted\_index.py .\pre\_processed.txt .\inverted\_indexed.txt**

**Expected file format:**

**pre\_processed.txt**

D1 sampl document less document first

D2 name second sampl document more sentenc than sampl document first

D3 name sampl document space tab charact doc singl line

**inverted\_indexed.txt**

sampl D1,1 D2,2 D3,1

document D1,2 D2,2 D3,1

less D1,1

first D1,1 D2,1

name D2,1 D3,1

second D2,1

more D2,1

sentenc D2,1

than D2,1

space D3,1

tab D3,1

charact D3,1

doc D3,1

singl D3,1

line D3,1 assign D3,1

challeng D4,1 D6,1

barrier D5,1

**3. TF-IDF:**

**python .\tfidf.py .\inverted\_indexed.txt .\tfidf.txt**

**Expected file format:**

**inverted\_indexed.txt**

sampl D1,1 D2,2 D3,1

document D1,2 D2,2 D3,1

less D1,1

first D1,1 D2,1

name D2,1 D3,1

second D2,1

more D2,1

sentenc D2,1

than D2,1

space D3,1

tab D3,1

charact D3,1

doc D3,1

singl D3,1

line D3,1

**tfidf.txt**

D1 D2 D3

sampl 0.0 0.0 0.0

document 0.0 0.0 0.0

less 0.239 0.0 0.0

first 0.088 0.088 0.0

name 0.0 0.088 0.176

second 0.0 0.239 0.0

more 0.0 0.239 0.0

sentenc 0.0 0.239 0.0

than 0.0 0.239 0.0

space 0.0 0.0 0.477

tab 0.0 0.0 0.477

charact 0.0 0.0 0.477

doc 0.0 0.0 0.477

singl 0.0 0.0 0.477

line 0.0 0.0 0.477

**4. COSINE:**

**python .\cosine\_sim.py .\tfidf.txt D1 D2**

**Expected file format:**

**tfidf.txt**

D1 D2 D3

sampl 0.0 0.0 0.0

document 0.0 0.0 0.0

less 0.239 0.0 0.0

first 0.088 0.088 0.0

name 0.0 0.088 0.176

second 0.0 0.239 0.0

more 0.0 0.239 0.0

sentenc 0.0 0.239 0.0

than 0.0 0.239 0.0

space 0.0 0.0 0.477

tab 0.0 0.0 0.477

charact 0.0 0.0 0.477

doc 0.0 0.0 0.477

singl 0.0 0.0 0.477

line 0.0 0.0 0.477

**Output on cmd:**

0.062

**5. IR SYSTEM:**

**python .\ir\_system.py .\input.txt "first"**

**Expected input file format:**

**input.txt**

D1 This is a sample document, it has less document then the first one.

D2 This is a name of second sample document. This one has more sentences than a sample document first.

D3 The name of the sample document is spaced by a TAB character! all docs are on a single line

**Output on cmd:**

D1 0.346  
D2 0.178  
D3 0.0