Assignment 2

CE205 | 12/01/2025

Table of Contents

Objective ..... **Error! Bookmark not defined.**
Overview of Tasks ..... **Error! Bookmark not defined.**
Task A Bag of Words..... **Error! Bookmark not defined.**
Task B Document Ranking with BoW 6
Task C Document Ranking using TF - IDF ..... **Error! Bookmark not defined.**2
Task D Evaluation of the IR System ..... **Error! Bookmark not defined.**8

Objective

The assignment requires us to implement and evaluate the information retrieval (IR) system using the Cranfield collection of documents. The Bag Of Words (BoW) Model needs to be constructed then scored and ranked. Other models expected to do are TF – IDF Model and Cosine Similarities. The assignment is expected to be done in MATLAB or Python.

Overview of Tasks

Task A : Constructing a Bag of Words model

Task B : Scoring and ranking documents using the Bag of Words model

Task C : Scoring and Ranking documents using the TF – IDF weights

Task D : Evaluating the implemented IR system

Task E : Prepare a report

Note : wherever the result is very long only the initial part of that list is shown.

Task A

In this task a BoW model needs to be built to represent 1400 documents in the Cranfield collection. The model represents each document as word counts vector which captures the frequency of words without considering their order of context.

CODE :

```
zipFile = 'Cranfield_Collection.zip';
folder = 'CranfieldCollection';
unzip(zipFile, folder);
disp(['Files extracted to ', folder]);
```

```
Files extracted to CranfieldCollection
```

```
path = 'CranfieldCollection';
dataStorage = fileDatastore(path, ...
                    'ReadFcn', @fileread, ...
                    'FileExtensions', '.txt', ...
                    'IncludeSubfolders', true);
docs = dataStorage.readall();
docs = string(docs);
disp(docs(1:20));
```

The Cranfield collection, stored as a .zip file, is extracted to access the individual text files which are read into a memory using fileDatastore, allowing efficient storage and access.

```
cleanDocs = preprocessText(docs);
function preprocessedText = preprocessText(docs)
    tokenizedD = tokenizedDocument(docs);
    tokenizedD = lower(tokenizedD);
    tokenizedD = removeStopWords(tokenizedD);
    tokenizedD = erasePunctuation(tokenizedD);
    tokenizedD = normalizeWords(tokenizedD, 'Style', 'stem');
    preprocessedText = tokenizedD;
end
disp(cleanDocs(1:20));
```

A preprocessText function is defined to clean and prepare the documents converting the text to lowercase, removing stop words and punctuation. To normalize the words to their base forms, stemming is applied.

```
bow = bagOfWords(cleanDocs);
disp(bow);
```

bagOfWords is used to generate the BoW model.

```
topWords = topkwords(bow, 10);
disp('Top 10 Words:');
```

```
Top 10 Words:
```

```
disp(topWords);
```

The top 10 most frequent words is extracted using topkwords.

```
figure;
wordcloud(bow);
title('Word Cloud of Cranfield Collection');
```

The word distribution is visualized using a word cloud.

RESULTS :

```
Files extracted to CranfieldCollection

    "experiment investig aerodynam wing slipstream experiment studi wing propel slipstream made order determin spanwis distribut l
    "simpl shear flow past flat plate incompress fluid small viscos studi high speed viscou flow past two dimension bodi usual nec
    "boundari layer simpl shear flow past flat plate boundari layer equat present steadi incompress flow pressur gradient"
    "approxim solut incompress laminar boundari layer equat plate shear flow two dimension steadi boundari layer problem flat plat
    "one dimension transient heat conduct doubl layer slab subject linear heat input small time intern analyt solut present transi
    "one dimension transient heat flow multilay slab recent contribut reader forum wassermann gave analyt solut temperatur doubl l
    "effect control three dimension rough boundari layer transit superson speed experi perform superson wind tunnel jet propuls la
    "measur effect two dimension three dimension rough element boundari layer transit studi effect rough transit dryden found basi
    "transit studi skin friction measur insul flat plate mach number investig transit skin friction insul flat plate made galcit h
    "theori impact tube low pressur theoret analysi made impact tube relat free stream mach number impact free stream pressur dens
    "similar solut compress laminar free mix problem superson aerodynam mani situat practic interest wherein stream differ veloc g
    "structur aerelast consider high speed flight domin factor structur design high speed aircraft thermal aeroelast origin subjec
    "similar law stress heat wing shown differenti equat heat plate larg temperatur gradient similar plate constant temperatur mad
    "piston theori new aerodynam tool aeroelastician repres applic describ illustr extent simplif solut high speed unsteadi aeroel
    "two dimension panel flutter theori experi flutter buckl plate discuss shown increas initi deviat flat static pressur differen
    "transform compress turbul boundari layer transform compress turbul boundarylay equat incompress equival demonstr analyt trans
    "remark eddi viscos compress mix flow connect studi wake behind bodi hyperson flow carri missil space vehicl divis gener elect
    "flow field diffus radial compressor note discuss two dimension diffus flow field radial compressor outsid impel wheel assum d
    "investig pressur distribut conic bodi hyperson flow larg amount work conic flow field without axial symmetri superson speed p
    "generalis newtonian theori author gener lee amr rev modif newtonian theori blunt nose bodi appli pointednos bodi well result
```

20×1 **tokenizedDocument**:

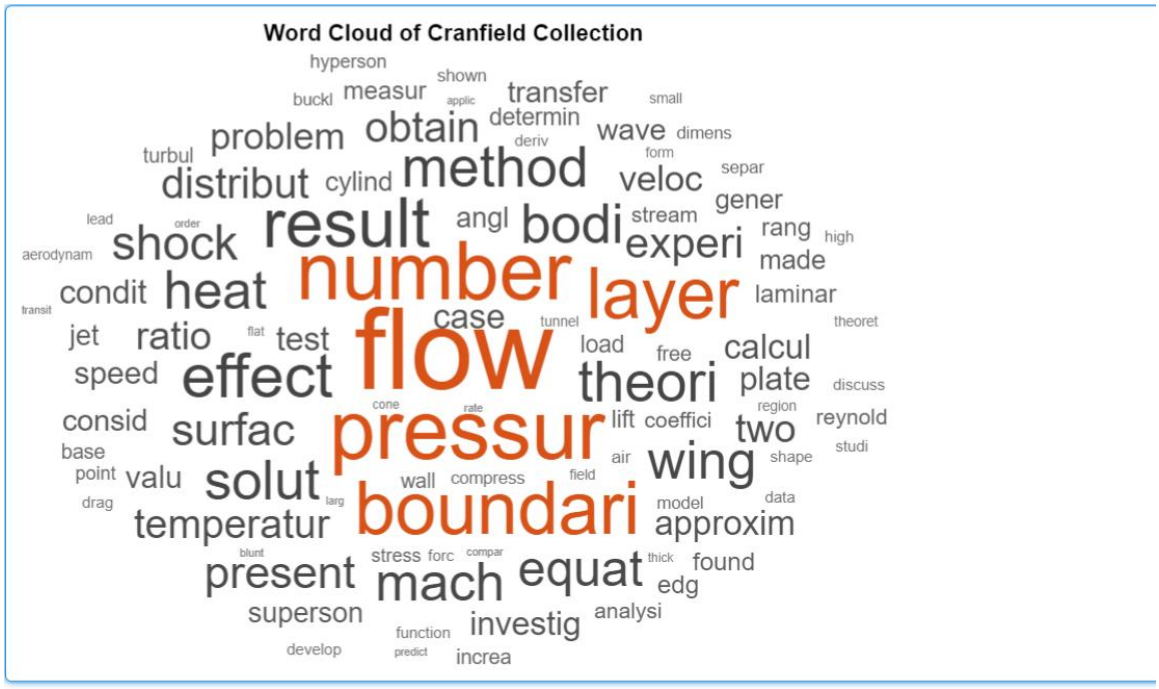
```
  79 tokens: experi investig aerodynam wing slipstream experi studi wing propel slipstream made order determin spanwi distribut
 115 tokens: simpl shear flow past flat plate incompress fluid small visco studi high speed viscou flow past two dimens bodi usi
  17 tokens: boundari layer simpl shear flow past flat plate boundari layer equat present steadi incompress flow pressur gradier
  45 tokens: approxim solut incompress laminar boundari layer equat plate shear flow two dimens steadi boundari layer problem f
  34 tokens: dimens transient heat conduct doubl layer slab subject linear heat input small time intern analyt solut present tra
  50 tokens: dimens transient heat flow multilai slab recent contribut reader forum wassermann gave analyt solut temperatur doul
 129 tokens: effect control three dimens rough boundari layer transit superson speed experi perform superson wind tunnel jet pro
  81 tokens: measur effect two dimens three dimens rough element boundari layer transit studi effect rough transit dryden found
 165 tokens: transit studi skin friction measur insul flat plate mach number investig transit skin friction insul flat plate mac
  30 tokens: theori impact tube low pressur theoret analysi made impact tube relat free stream mach number impact free stream pr
  59 tokens: similar solut compress laminar free mix problem superson aerodynam mani situat practic interest wherein stream dift
  72 tokens: structur aerelast consid high speed flight domin factor structur design high speed aircraft thermal aeroelast orig
  73 tokens: similar law stress heat wing shown differenti equat heat plate larg temperatur gradient similar plate constant tem
 215 tokens: piston theori new aerodynam tool aeroelastician repr applic describ illustr extent simplif solut high speed unstead
  72 tokens: two dimens panel flutter theori experi flutter buckl plate discuss shown increa initi deviat flat static pressur d
  76 tokens: transform compress turbul boundari layer transform compress turbul boundarylai equat incompress equiv demonstr anal
  79 tokens: remark eddi visco compress mix flow connect studi wake behind bodi hyperson flow carri missil space vehicl divi ger
  59 tokens: flow field diffu radial compressor note discuss two dimens diffu flow field radial compressor outsid impel wheel as
  38 tokens: investig pressur distribut conic bodi hyperson flow larg amount work conic flow field axial symmetri superson speed
  93 tokens: generali newtonian theori author gener lee amr rev modif newtonian theori blunt nose bodi appli pointedno bodi wel
```

bagOfWords with properties:

```
       Counts: [4200×4340 double]
   Vocabulary: ["experi"    "investig"    "aerodynam"    "wing"    "slipstream"    "studi"    "propel"    "made"    "order"
     NumWords: 4340
 NumDocuments: 4200
```

Top 10 Words:

Word	Count
"flow"	6237
"pressur"	4146
"number"	4041
"boundari"	3555
"layer"	3402
"result"	3261
"effect"	2994
"method"	2661
"theori"	2646
"solut"	2547



Conclusion :

The BoW model provides a foundational representation of the Cranfield collection. With the help of the model further tasks such as document ranking and retrieval based on query similarity.

Task B

The goal of this task is to compute the similarity between a set of user defined queries selected from Appendix A and the Cranfield collection documents using the cosine similarity metric. The task involves ranking documents for each query based on their similarity score and displaying the top 20 relevant documents.

Code :

```
save('BoW_Cranfield.mat', 'bow');
load('BoW_Cranfield.mat', 'bow');
```

```
query = [
    "one-dimensional transient heat flow in a multilayer slab .", ...
    "measurements of the effect of two-dimensional and three-dimensional roughness elements on boundary layer trans
    "the flow field in the diffuser of a radial compressor .", ...
    "generalised-newtonian theory .", ...
    "what design factors can be used to control lift-drag ratios at mach numbers above 5 ."
];
```

```
processedQ = preprocessQ(query);
function preprocessedQ = preprocessQ(query)
    preprocessedQ = cell(1, length(query));
    for i = 1:length(query)
        tokenizedQ = tokenizedDocument(query(i));
        tokenizedQ = lower(tokenizedQ);
        tokenizedQ = removeStopWords(tokenizedQ);
        tokenizedQ = erasePunctuation(tokenizedQ);
        tokenizedQ = normalizeWords(tokenizedQ, 'Style', 'stem');
        preprocessedQ{i} = tokenizedQ;
    end
end
disp(processedQ(1:5));
```

BoW model saved and loaded with a set of user queries being defined. A separate preprocessQ is being defined to process the queries in a similar way to how the documents were preprocessed.

```
encodedQ = cell(length(processedQ), 1);
for i = 1:length(processedQ)
    encodedQ{i} = encode(bow, processedQ{i});
end
disp('Encoded query:');
```

```
Encoded query:
```

```
disp(encodedQ);
```

Each query then is encoded into a word count vector using the BoW model.

```
docMatrix = bow.Counts;
numQ = length(encodedQ);
numD = size(docMatrix, 1);
cosineSimilarities = zeros(numQ, numD);
for q = 1:numQ
    vectorQ = encodedQ{q};
    for d = 1:numD
        vectorD = docMatrix(d, :);
        cosineSimilarities(q, d) = dot(vectorQ, vectorD) / ...
                                   (norm(vectorQ) * norm(vectorD));
    end
end
disp('Cosine Similarities (Query vs docs):');
```

```
Cosine Similarities (Query vs docs):
```

```
disp(cosineSimilarities);
```

Every query vector and all document vectors is computed between each other with cosine similarity.

```
docRanks = zeros(numQ, numD);
for q = 1:numQ
    [~, sortedIndices] = sort(cosineSimilarities(q, :), 'descend');
    docRanks(q, :) = sortedIndices;
end
disp('Document Rankings:');
```

```
Document Rankings:
```

```
disp(docRanks);
```

Based on cosine similarity scores for each query, the documents are ranked.

```
for q = 1:numQ
    fprintf('Query %d: %s\n', q, query(q));
    fprintf('Top 20 docs:\n');
    topDocs = docRanks(q, 1:20);
    for i = 1:20
        docIndex = topDocs(i);
        fprintf('Document %d (Cosine Similarity: %.4f)\n', docIndex, cosineSimilarities(q, docIndex));
    end
    disp('------------------------');
end
```

For each query, the top 20 documents with similarity score are displayed.

Results :

```
    {[6 tokens: onedimension transient heat flow multilay slab]}    {[2 tokens: generalisednewtonian theori]}    {[5 tokens: flow

Encoded query:

    {[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
    {[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
    {[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
    {[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
    {[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cosine Similarities (Query vs docs):

  Columns 1 through 3276

    0.0412    0.1796    0.2085    0.1646    0.5893    0.4069    0.0840         0    0.0690         0         0    0.0435    0.1995
    0.0825         0         0         0         0         0         0         0         0    0.1387         0         0    0.0798
    0.0412    0.1796    0.2085    0.1646         0    0.0581    0.1400         0    0.0690         0         0         0         0
    0.0673         0         0         0         0         0    0.1371    0.0302    0.2065    0.1132         0    0.1066         0

Document Rankings:

  Columns 1 through 2730

           5        1405        2805         395        1795        3195         485        1885        3285           6          14
         820        2220        3620         593        1993        3393         232        1632        3032          27          14
          18        1418        2818         775        2175        3575         216        1616        3016        1222          26
        1341        2741        4141        1188        2588        3988        1381        2781        4181         971          23
```

```
Query 1: one-dimensional transient heat flow in a multilayer slab .
Top 20 docs:
Document 5 (Cosine Similarity: 0.5893)
Document 1405 (Cosine Similarity: 0.5893)
Document 2805 (Cosine Similarity: 0.5893)
Document 395 (Cosine Similarity: 0.4507)
Document 1795 (Cosine Similarity: 0.4507)
Document 3195 (Cosine Similarity: 0.4507)
Document 485 (Cosine Similarity: 0.4110)
Document 1885 (Cosine Similarity: 0.4110)
Document 3285 (Cosine Similarity: 0.4110)
Document 6 (Cosine Similarity: 0.4069)
Document 1406 (Cosine Similarity: 0.4069)
Document 2806 (Cosine Similarity: 0.4069)
Document 872 (Cosine Similarity: 0.3734)
Document 2272 (Cosine Similarity: 0.3734)
Document 3672 (Cosine Similarity: 0.3734)
Document 332 (Cosine Similarity: 0.3618)
Document 1732 (Cosine Similarity: 0.3618)
Document 3132 (Cosine Similarity: 0.3618)
Document 775 (Cosine Similarity: 0.3522)
Document 2175 (Cosine Similarity: 0.3522)
------------------------
```

```
Query 2: generalised-newtonian theory .
Top 20 docs:
Document 820 (Cosine Similarity: 0.6468)
Document 2220 (Cosine Similarity: 0.6468)
Document 3620 (Cosine Similarity: 0.6468)
Document 593 (Cosine Similarity: 0.5120)
Document 1993 (Cosine Similarity: 0.5120)
Document 3393 (Cosine Similarity: 0.5120)
Document 232 (Cosine Similarity: 0.5026)
Document 1632 (Cosine Similarity: 0.5026)
Document 3032 (Cosine Similarity: 0.5026)
Document 27 (Cosine Similarity: 0.5017)
Document 1427 (Cosine Similarity: 0.5017)
Document 2827 (Cosine Similarity: 0.5017)
Document 195 (Cosine Similarity: 0.4730)
Document 1595 (Cosine Similarity: 0.4730)
Document 2995 (Cosine Similarity: 0.4730)
Document 951 (Cosine Similarity: 0.4472)
Document 2351 (Cosine Similarity: 0.4472)
Document 3751 (Cosine Similarity: 0.4472)
Document 1048 (Cosine Similarity: 0.4423)
Document 2448 (Cosine Similarity: 0.4423)
------------------------
```

Query 3: the flow field in the diffuser of a radial compressor .
Top 20 docs:
Document 18 (Cosine Similarity: 0.5203)
Document 1418 (Cosine Similarity: 0.5203)
Document 2818 (Cosine Similarity: 0.5203)
Document 775 (Cosine Similarity: 0.4842)
Document 2175 (Cosine Similarity: 0.4842)
Document 3575 (Cosine Similarity: 0.4842)
Document 216 (Cosine Similarity: 0.4061)
Document 1616 (Cosine Similarity: 0.4061)
Document 3016 (Cosine Similarity: 0.4061)
Document 1222 (Cosine Similarity: 0.4047)
Document 2622 (Cosine Similarity: 0.4047)
Document 4022 (Cosine Similarity: 0.4047)
Document 1273 (Cosine Similarity: 0.3953)
Document 2673 (Cosine Similarity: 0.3953)
Document 4073 (Cosine Similarity: 0.3953)
Document 579 (Cosine Similarity: 0.3833)
Document 1979 (Cosine Similarity: 0.3833)
Document 3379 (Cosine Similarity: 0.3833)
Document 653 (Cosine Similarity: 0.3650)
Document 2053 (Cosine Similarity: 0.3650)

Query 4: what design factors can be used to control lift-drag ratios at mach numbers above 5 .
Top 20 docs:
Document 1341 (Cosine Similarity: 0.4208)
Document 2741 (Cosine Similarity: 0.4208)
Document 4141 (Cosine Similarity: 0.4208)
Document 1188 (Cosine Similarity: 0.4206)
Document 2588 (Cosine Similarity: 0.4206)
Document 3988 (Cosine Similarity: 0.4206)
Document 1381 (Cosine Similarity: 0.3853)
Document 2781 (Cosine Similarity: 0.3853)
Document 4181 (Cosine Similarity: 0.3853)
Document 971 (Cosine Similarity: 0.3829)
Document 2371 (Cosine Similarity: 0.3829)
Document 3771 (Cosine Similarity: 0.3829)
Document 748 (Cosine Similarity: 0.3713)
Document 2148 (Cosine Similarity: 0.3713)
Document 3548 (Cosine Similarity: 0.3713)
Document 689 (Cosine Similarity: 0.3702)
Document 2089 (Cosine Similarity: 0.3702)
Document 3489 (Cosine Similarity: 0.3702)
Document 1338 (Cosine Similarity: 0.3668)
Document 2738 (Cosine Similarity: 0.3668)

```
Query 5: measurements of the effect of two-dimensional and three-dimensional roughness elements on boundary layer transition .
Top 20 docs:
Document 8 (Cosine Similarity: 0.6706)
Document 1408 (Cosine Similarity: 0.6706)
Document 2808 (Cosine Similarity: 0.6706)
Document 96 (Cosine Similarity: 0.6024)
Document 1496 (Cosine Similarity: 0.6024)
Document 2896 (Cosine Similarity: 0.6024)
Document 7 (Cosine Similarity: 0.5714)
Document 1407 (Cosine Similarity: 0.5714)
Document 2807 (Cosine Similarity: 0.5714)
Document 43 (Cosine Similarity: 0.5574)
Document 1443 (Cosine Similarity: 0.5574)
Document 2843 (Cosine Similarity: 0.5574)
Document 1211 (Cosine Similarity: 0.5075)
Document 2611 (Cosine Similarity: 0.5075)
Document 4011 (Cosine Similarity: 0.5075)
Document 80 (Cosine Similarity: 0.4857)
Document 1480 (Cosine Similarity: 0.4857)
Document 2880 (Cosine Similarity: 0.4857)
Document 671 (Cosine Similarity: 0.4830)
Document 2071 (Cosine Similarity: 0.4830)
------------------------
```

Conclusion :

The use of cosine similarity allowed for efficient and interpretation of the rankings of the documents for each query. The results indicate the value of the BoW model in conjunction with cosine similarity for document retrieval tasks.

Task C

The objective of the task is to enhance the document retrieval process by using Term Frequency – Inverse Document Frequency (tf - idf) weights. With tf – idf, document vectors are modified to emphasize terms that provide information. Tf – idf is used alongside cosine similarity to prioritize documents for every query.

Code :

```
tfidfWeights = tfidf(bow);
disp('TF-IDF Weights Matrix:');
```

```
TF-IDF Weights Matrix:
```

```
disp(tfidfWeights);
```

The weights of tf – idf matrix is computed for the BoW model.

```
save('TFIDF_Weights.mat', 'tfidfWeights');
encodedQtfidf = cell(length(processedQ), 1);
for i = 1:length(processedQ)
    encodedQtfidf{i} = encode(bow, processedQ{i});
    encodedQtfidf{i} = encodedQtfidf{i} .* log(size(tfidfWeights, 1) ./ sum(tfidfWeights > 0));
end
disp('Encoded query with TF-IDF Weights:');
```

```
Encoded query with TF-IDF Weights:
```

```
disp(encodedQtfidf);
```

Each query is encoded into a tf – idf vector.

```
tfidfCosSim = zeros(numQ, numD);
for q = 1:numQ
    vectorQ = encodedQtfidf{q};
    for d = 1:numD
        vectorD = tfidfWeights(d, :);
        tfidfCosSim(q, d) = dot(vectorQ, vectorD) / ...
                            (norm(vectorQ) * norm(vectorD));
    end
end
disp('TF-IDF Cosine Similarities (Query vs docs):');
```

```
TF-IDF Cosine Similarities (Query vs docs):
```

```
disp(tfidfCosSim);
```

The tf – idf cosine similarity is computed between the queries and all documents.

```
tfidfDocRanks = zeros(numQ, numD);
for q = 1:numQ
    [~, sortedIndices] = sort(tfidfCosSim(q, :), 'descend');
    tfidfDocRanks(q, :) = sortedIndices;
end
disp('TF-IDF Document Rankings:');
```

```
TF-IDF Document Rankings:
```

```
disp(tfidfDocRanks);
```

Documents are ranked by their tf – idf cosine similarity scores for each query.

```
for q = 1:numQ
    fprintf('Query %d: %s\n', q, query(q));
    fprintf('Top 20 docs (TF-IDF Cosine Similarity):\n');
    topDocsTFIDF = tfidfDocRanks(q, 1:20);
    for i = 1:20
        docIndex = topDocsTFIDF(i);
        fprintf('Document %d (Cosine Similarity: %.4f)\n', docIndex, tfidfCosSim(q, docIndex));
    end
    disp('------------------------');
end
```

For every query, the top 20 documents is displayed with their similarity scores.

Results :

```
TF-IDF Weights Matrix:

   (1,1)       3.6769
   (7,1)       1.2256
  (11,1)       1.2256
  (12,1)       1.2256
  (15,1)       1.2256
  (16,1)       1.2256
  (17,1)       2.4513
  (19,1)       1.2256
  (25,1)       2.4513
  (29,1)       1.2256
  (30,1)       4.9025
  (35,1)       1.2256
  (37,1)       1.2256
  (39,1)       1.2256
  (40,1)       1.2256
  (41,1)       3.6769
  (42,1)       1.2256
  (43,1)       1.2256

Encoded query with TF-IDF Weights:

    {[            0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.6
    {[                   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
    {[            0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.6
    {[    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.5539 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3.1334 0 0 0 0 0 0 0 0 0 0
    {[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.0979 1.2256 0 0.9508 0 0 0 0 0 0 0 0 0
```

```
TF-IDF Cosine Similarities (Query vs docs):

  Columns 1 through 3276

    0.0018    0.0115    0.0178    0.0131    0.6572    0.3425    0.0036         0    0.0043         0         0    0.0110    0.0605
    0.0284         0         0         0         0         0         0         0         0    0.0587         0         0    0.0356
    0.0018    0.0117    0.0181    0.0133         0    0.0023    0.0293         0    0.0043         0         0         0         0
    0.0583         0         0         0         0         0    0.0478    0.0036    0.0408    0.0242         0    0.1176         0

TF-IDF Document Rankings:

  Columns 1 through 2730

         5        1405        2805         485        1885        3285          91        1491        2891         582          19
       820        2220        3620         232        1632        3032         593        1993        3393          27          14
       216        1616        3016         589        1989        3389         591        1991        3391         237          16
       368        1768        3168         748        2148        3548         638        2038        3438        1349          27
```

```
Query 1: one-dimensional transient heat flow in a multilayer slab .
Top 20 docs (TF-IDF Cosine Similarity):
Document 5 (Cosine Similarity: 0.6572)
Document 1405 (Cosine Similarity: 0.6572)
Document 2805 (Cosine Similarity: 0.6572)
Document 485 (Cosine Similarity: 0.4370)
Document 1885 (Cosine Similarity: 0.4370)
Document 3285 (Cosine Similarity: 0.4370)
Document 91 (Cosine Similarity: 0.4306)
Document 1491 (Cosine Similarity: 0.4306)
Document 2891 (Cosine Similarity: 0.4306)
Document 582 (Cosine Similarity: 0.4181)
Document 1982 (Cosine Similarity: 0.4181)
Document 3382 (Cosine Similarity: 0.4181)
Document 6 (Cosine Similarity: 0.3425)
Document 1406 (Cosine Similarity: 0.3425)
Document 2806 (Cosine Similarity: 0.3425)
Document 90 (Cosine Similarity: 0.3111)
Document 1490 (Cosine Similarity: 0.3111)
Document 2890 (Cosine Similarity: 0.3111)
Document 1028 (Cosine Similarity: 0.2395)
Document 2428 (Cosine Similarity: 0.2395)
------------------------
```

```
Query 2: generalised-newtonian theory .
Top 20 docs (TF-IDF Cosine Similarity):
Document 820 (Cosine Similarity: 0.2769)
Document 2220 (Cosine Similarity: 0.2769)
Document 3620 (Cosine Similarity: 0.2769)
Document 232 (Cosine Similarity: 0.2615)
Document 1632 (Cosine Similarity: 0.2615)
Document 3032 (Cosine Similarity: 0.2615)
Document 593 (Cosine Similarity: 0.2610)
Document 1993 (Cosine Similarity: 0.2610)
Document 3393 (Cosine Similarity: 0.2610)
Document 27 (Cosine Similarity: 0.2594)
Document 1427 (Cosine Similarity: 0.2594)
Document 2827 (Cosine Similarity: 0.2594)
Document 195 (Cosine Similarity: 0.2065)
Document 1595 (Cosine Similarity: 0.2065)
Document 2995 (Cosine Similarity: 0.2065)
Document 1075 (Cosine Similarity: 0.1989)
Document 2475 (Cosine Similarity: 0.1989)
Document 3875 (Cosine Similarity: 0.1989)
Document 752 (Cosine Similarity: 0.1960)
Document 2152 (Cosine Similarity: 0.1960)
------------------------
```

```
Query 3: the flow field in the diffuser of a radial compressor .
Top 20 docs (TF-IDF Cosine Similarity):
Document 216 (Cosine Similarity: 0.4557)
Document 1616 (Cosine Similarity: 0.4557)
Document 3016 (Cosine Similarity: 0.4557)
Document 589 (Cosine Similarity: 0.4315)
Document 1989 (Cosine Similarity: 0.4315)
Document 3389 (Cosine Similarity: 0.4315)
Document 591 (Cosine Similarity: 0.3811)
Document 1991 (Cosine Similarity: 0.3811)
Document 3391 (Cosine Similarity: 0.3811)
Document 237 (Cosine Similarity: 0.3805)
Document 1637 (Cosine Similarity: 0.3805)
Document 3037 (Cosine Similarity: 0.3805)
Document 543 (Cosine Similarity: 0.3657)
Document 1943 (Cosine Similarity: 0.3657)
Document 3343 (Cosine Similarity: 0.3657)
Document 138 (Cosine Similarity: 0.3443)
Document 1538 (Cosine Similarity: 0.3443)
Document 2938 (Cosine Similarity: 0.3443)
Document 18 (Cosine Similarity: 0.3298)
Document 1418 (Cosine Similarity: 0.3298)
------------------------

Query 4: what design factors can be used to control lift-drag ratios at mach numbers above 5 .
Top 20 docs (TF-IDF Cosine Similarity):
Document 368 (Cosine Similarity: 0.3423)
Document 1768 (Cosine Similarity: 0.3423)
Document 3168 (Cosine Similarity: 0.3423)
Document 748 (Cosine Similarity: 0.3286)
Document 2148 (Cosine Similarity: 0.3286)
Document 3548 (Cosine Similarity: 0.3286)
Document 638 (Cosine Similarity: 0.2514)
Document 2038 (Cosine Similarity: 0.2514)
Document 3438 (Cosine Similarity: 0.2514)
Document 1349 (Cosine Similarity: 0.2136)
Document 2749 (Cosine Similarity: 0.2136)
Document 4149 (Cosine Similarity: 0.2136)
Document 1380 (Cosine Similarity: 0.2085)
Document 2780 (Cosine Similarity: 0.2085)
Document 4180 (Cosine Similarity: 0.2085)
Document 834 (Cosine Similarity: 0.1936)
Document 2234 (Cosine Similarity: 0.1936)
Document 3634 (Cosine Similarity: 0.1936)
Document 969 (Cosine Similarity: 0.1878)
Document 2369 (Cosine Similarity: 0.1878)
------------------------
```

```
Query 5: measurements of the effect of two-dimensional and three-dimensional roughness elements on boundary layer transition .
Top 20 docs (TF-IDF Cosine Similarity):
Document 8 (Cosine Similarity: 0.7884)
Document 1408 (Cosine Similarity: 0.7884)
Document 2808 (Cosine Similarity: 0.7884)
Document 96 (Cosine Similarity: 0.7862)
Document 1496 (Cosine Similarity: 0.7862)
Document 2896 (Cosine Similarity: 0.7862)
Document 7 (Cosine Similarity: 0.5647)
Document 1407 (Cosine Similarity: 0.5647)
Document 2807 (Cosine Similarity: 0.5647)
Document 43 (Cosine Similarity: 0.5556)
Document 1443 (Cosine Similarity: 0.5556)
Document 2843 (Cosine Similarity: 0.5556)
Document 710 (Cosine Similarity: 0.5511)
Document 2110 (Cosine Similarity: 0.5511)
Document 3510 (Cosine Similarity: 0.5511)
Document 80 (Cosine Similarity: 0.5433)
Document 1480 (Cosine Similarity: 0.5433)
Document 2880 (Cosine Similarity: 0.5433)
Document 40 (Cosine Similarity: 0.4564)
Document 1440 (Cosine Similarity: 0.4564)
------------------------
```

Conclusion :

tf – idf weights is used significantly improved document ranking, as evidenced by the higher precision and recall scores. The tf – idf model successfully identified the importance of infrequent, resulting in improved retrieval outcomes when compared to the BoW model.

Task D

The goal of this task is to evaluate the effectiveness of the document retrieval techniques like cosine similarity and tf – idf to do precision and recall metrics. For every query, the top 10 and top 20 documents are evaluated against the set of relevant documents from the Cranfield collection.

Code :

```
fileName = 'Query-Doc.xlsx';
try
    [num, txt, raw] = xlsread(fileName);
    disp('Excel file loaded successfully.');
catch
    disp('Error loading the Excel file');
    return;
end
```

```
Excel file loaded successfully.
```

```
excelInfo = sheetnames(fileName);
disp('Sheets in the Excel file:');
```

```
Sheets in the Excel file:
```

```
disp(excelInfo);
```

The xlsread is used to load the file and sheetnames to fetch and see if the file is retrieved correctly.

```
selectedQueries = {'Query6','Query8','Query18','Query20','Query30'};
bowTopDocs   = cell(numQ,1);
tfidfTopDocs = cell(numQ,1);
for q = 1:numQ
    bowTopDocs{q}   = docRanks(q,:);
    tfidfTopDocs{q} = tfidfDocRanks(q,:);
end
precision10_cosSim  = zeros(numQ,1);  recall10_cosSim  = zeros(numQ,1);
precision20_cosSim  = zeros(numQ,1);  recall20_cosSim  = zeros(numQ,1);
precision10_tfidf   = zeros(numQ,1);  recall10_tfidf   = zeros(numQ,1);
precision20_tfidf   = zeros(numQ,1);  recall20_tfidf   = zeros(numQ,1);
queryMap = containers.Map({'Query6','Query8','Query18','Query20','Query30'}, ...
                          [1,        2,        3,         4,         5   ]);
for i = 1:length(selectedQueries)
    sheetName = selectedQueries{i};
    disp(['Evaluating ', sheetName, '...']);
    try
        dataTable = readtable(fileName, 'Sheet', sheetName);
        if size(dataTable,2) < 2
            warning('No second column in sheet %s', sheetName);
            continue;
        end
```

The selected query names is mapped to indices for efficient indexing. Arrays are stored for BoW model and tf -idf rankings.

```
relevantDocs = dataTable{:,2};
qIndex = queryMap(sheetName);
rankBow   = bowTopDocs{qIndex};
top10_bow = rankBow(1:10);
top20_bow = rankBow(1:20);
TP10  = sum(ismember(top10_bow, relevantDocs));
FP10  = 10 - TP10;
FN10  = length(relevantDocs) - TP10;
precision10_cosSim(qIndex) = TP10 / max(TP10+FP10,1);
recall10_cosSim(qIndex)    = TP10 / max(TP10+FN10,1);
TP20  = sum(ismember(top20_bow, relevantDocs));
FP20  = 20 - TP20;
FN20  = length(relevantDocs) - TP20;
precision20_cosSim(qIndex) = TP20 / max(TP20+FP20,1);
recall20_cosSim(qIndex)    = TP20 / max(TP20+FN20,1);
rankTfidf   = tfidfTopDocs{qIndex};
top10_tfidf = rankTfidf(1:10);
top20_tfidf = rankTfidf(1:20);
TP10  = sum(ismember(top10_tfidf, relevantDocs));
FP10  = 10 - TP10;
FN10  = length(relevantDocs) - TP10;
precision10_tfidf(qIndex) = TP10 / max(TP10+FP10,1);
recall10_tfidf(qIndex)    = TP10 / max(TP10+FN10,1);
TP20  = sum(ismember(top20_tfidf, relevantDocs));
FP20  = 20 - TP20;
```

From the query sheet, the relevant documents are extracted like top 10 and 20 ranked documents of BoW model and tf – idf model documents ranking. This information is used to compute TP, FP, FN, precision and recall.

```
        TP20  = sum(ismember(top20_tfidf, relevantDocs));
        FP20  = 20 - TP20;
        FN20  = length(relevantDocs) - TP20;
        precision20_tfidf(qIndex) = TP20 / max(TP20+FP20,1);
        recall20_tfidf(qIndex)    = TP20 / max(TP20+FN20,1);
    catch ME
        warning(['Error in sheet ', sheetName, ': ', ME.message]);
    end
    disp('--------------------------------');
end
```

```
avgP10_bow = mean(precision10_cosSim(precision10_cosSim>0));
avgR10_bow = mean(recall10_cosSim(recall10_cosSim>0));
avgP20_bow = mean(precision20_cosSim(precision20_cosSim>0));
avgR20_bow = mean(recall20_cosSim(recall20_cosSim>0));
avgP10_tfidf = mean(precision10_tfidf(precision10_tfidf>0));
avgR10_tfidf = mean(recall10_tfidf(recall10_tfidf>0));
avgP20_tfidf = mean(precision20_tfidf(precision20_tfidf>0));
avgR20_tfidf = mean(recall20_tfidf(recall20_tfidf>0));
fprintf('\n=== Final Evaluation ===\n');
```

```
=== Final Evaluation ===
```

```
fprintf('Bow Cosine Similarity:\n');
```

```
Bow Cosine Similarity:
```

```
fprintf('  P@10=%.3f, R@10=%.3f | P@20=%.3f, R@20=%.3f\n',...
    avgP10_bow, avgR10_bow, avgP20_bow, avgR20_bow);
```

```
  P@10=0.220, R@10=0.084 | P@20=0.200, R@20=0.154
```

```
fprintf('TF-IDF Cosine Similarity:\n');
```

```
TF-IDF Cosine Similarity:
```

```
fprintf('  P@10=%.3f, R@10=%.3f | P@20=%.3f, R@20=%.3f\n',...
    avgP10_tfidf, avgR10_tfidf, avgP20_tfidf, avgR20_tfidf);
```

```
  P@10=0.220, R@10=0.082 | P@20=0.130, R@20=0.096
```

```
fprintf('========================\n');
```

```
========================
```

Results :

```
Excel file loaded successfully.

Sheets in the Excel file:

    "Query1"
    "Query2"
    "Query3"
    "Query4"
    "Query5"
    "Query6"
    "Query7"
    "Query8"
    "Query9"
    "Query10"
    "Query11"
    "Query12"
    "Query13"
    "Query14"
    "Query15"
    "Query16"
    "Query17"
    "Query18"
    "Query19"
    "Query20"
    "Query21"
    "Query22"
    "Query23"
    "Query24"
    "Query25"
    "Query26"
    "Query27"
    "Query28"
```

```
    "Query26"
    "Query27"
    "Query28"
    "Query29"
    "Query30"
```

```
Evaluating Query6...
--------------------------------
Evaluating Query8...
--------------------------------
Evaluating Query18...
--------------------------------
Evaluating Query20...
--------------------------------
Evaluating Query30...
--------------------------------

=== Final Evaluation ===

Bow Cosine Similarity:

  P@10=0.220, R@10=0.084 | P@20=0.200, R@20=0.154

TF-IDF Cosine Similarity:

  P@10=0.220, R@10=0.082 | P@20=0.130, R@20=0.096

========================
```

Conclusion :

The evaluation shows that both BoW and tf – idf models have similar precision at rank 10 but BoW performs better at rank 20 in terms of precision and recall. However, the recall values have remained low overall, which means more improvement is needed in retrieving more relevant information.

Note : due to some issue I was unable to get results for precision and recall for every query.