# Question a)

```
disp("Question A)")
load 'wordVecV.mat' % goes into a variable V

% disp(V)

% Euclidiean Distance
disp("Euclidean Distance")
euclidean_distances = zeros(10,10);
max_coord = zeros(1,2);
max = 0;
for i = 1:10
    for j = 1:10
        euclidean_distances(i,j) = euclidean_distance(V(:, i), V(:, j));

        if euclidean_distances(i,j) > max
            max = euclidean_distances(i,j);
            max_coord(1) = i;
            max_coord(2) = j;
        end
    end
end

disp(euclidean_distances)
disp(max)
disp(max_coord)

disp("Closest euclidean distance: Article: " + max_coord(1) + " and Article:
" + max_coord(2))

% Angle Distance
disp("Angle Distance")
angle_distances = zeros(10,10);
max_coord = zeros(1,2);
max = 0;
for i = 1:10
    for j = 1:10
        angle_distances(i,j) = angle_distance(V(:, i), V(:, j));

        if angle_distances(i,j) > max
            max = angle_distances(i,j);
            max_coord(1) = i;
            max_coord(2) = j;
        end
    end
end

disp(angle_distances)
disp(max)
disp(max_coord)
```

```
disp("Closest angle distance: Article: " + max_coord(1) + " and Article: " +
max_coord(2))

% Question b)
disp("Question b")

% Euclidiean Distance
disp("Euclidean Distance")
euclidean_distances = zeros(10,10);
max_coord = zeros(1,2);
max = 0;
for i = 1:10
    for j = 1:10
        euclidean_distances(i,j) = euclidean_distance(V(:, i)/sum(V(:, i)),
V(:, j)/sum(V(:, j)));

        if euclidean_distances(i,j) > max
            max = euclidean_distances(i,j);
            max_coord(1) = i;
            max_coord(2) = j;
        end
    end
end

disp(euclidean_distances)
disp(max)
disp(max_coord)

disp("Closest euclidean distance: Article: " + max_coord(1) + " and Article:
" + max_coord(2))

% Angle Distance
disp("Angle Distance")
angle_distances = zeros(10,10);
max_coord = zeros(1,2);
max = 0;
for i = 1:10
    for j = 1:10
        angle_distances(i,j) = angle_distance(V(:, i)/sum(V(:, i)), V(:, j)/
sum(V(:, j)));

        if angle_distances(i,j) > max
            max = angle_distances(i,j);
            max_coord(1) = i;
            max_coord(2) = j;
        end
    end
end

disp(angle_distances)
disp(max)
disp(max_coord)

disp("Closest angle distance: Article: " + max_coord(1) + " and Article: " +
```

```
max_coord(2))

% Question C
disp("Question C")

% Get f_doc
f_doc = zeros(1,1651);

for i = 1:length(f_doc)
    num_articles = 0;
    for j = 1:10
        if V(i, j) ~= 0
            num_articles = num_articles +1;
        end
    end
    f_doc(i) = num_articles;
end

% get w(t,d)
w_t_d = zeros(1651,10);
for i = 1:10
    for j = 1:1651
        w_t_d(j,i) = (V(j,i) / sum(V(:,i))) * sqrt(log(10/f_doc(j)));
    end
end

% Euclidiean Distance
disp("Euclidean Distance")
euclidean_distances = zeros(10,10);
max_coord = zeros(1,2);
max = 0;
for i = 1:10
    for j = 1:10
        euclidean_distances(i,j) = euclidean_distance(w_t_d(:, i), w_t_d(:,
j));

        if euclidean_distances(i,j) > max
            max = euclidean_distances(i,j);
            max_coord(1) = i;
            max_coord(2) = j;
        end
    end
end

disp(euclidean_distances)
disp(max)
disp(max_coord)

disp("TF-IDF: Closest euclidean distance: Article: " + max_coord(1) + " and
Article: " + max_coord(2))

% Angle Distance
disp("Angle Distance")
angle_distances = zeros(10,10);
```

```
max_coord = zeros(1,2);
max = 0;
for i = 1:10
    for j = 1:10
        angle_distances(i,j) = angle_distance(w_t_d(:, i), w_t_d(:, j));

        if angle_distances(i,j) > max
            max = angle_distances(i,j);
            max_coord(1) = i;
            max_coord(2) = j;
        end
    end
end

disp(angle_distances)
disp(max)
disp(max_coord)

disp("TF_IDF: Closest angle distance: Article: " + max_coord(1) + " and
Article: " + max_coord(2))
```

# Why use TF-IDF?

TF-IDF scales the data by how much each word shows up in all the documents. This means that similar to normlization, words from larger documents aren't disproportionaly weighted. It also makes sure that rare words that are only in a few documents aren't unfairly biased against jut because they are a rare word.

```
function euc_dist = euclidean_distance(x,y)
    euc_dist = sqrt(sum((x-y).^2));
end

function ang = angle_distance(x,y)
    ang = acos((sum(x.*y))/(norm(x)*norm(y)));
end

function norm = norm(x)
    norm = sqrt(sum(x.^2));
end
```

```
Question A)
Euclidean Distance
  Columns 1 through 7

         0   32.2955   35.8050   44.3959   32.2025   55.2540   29.0000
   32.2955         0   34.7707   42.1426   35.3553   53.3292   33.1361
   35.8050   34.7707         0   46.6154   33.6601   56.4535   37.2961
   44.3959   42.1426   46.6154         0   50.2792   51.3809   46.3681
   32.2025   35.3553   33.6601   50.2792         0   59.7662   29.0517
   55.2540   53.3292   56.4535   51.3809   59.7662         0   57.6541
   29.0000   33.1361   37.2961   46.3681   29.0517   57.6541         0
   31.0805   34.9428   37.9737   50.1498   30.9031   62.5700   24.7184
   48.1560   45.9347   45.2217   51.7301   48.5592   50.5569   50.5371
   40.3485   39.7366   40.1497   48.0729   41.9881   51.9326   42.2493
```

```
  Columns 8 through 10

   31.0805   48.1560   40.3485
   34.9428   45.9347   39.7366
   37.9737   45.2217   40.1497
   50.1498   51.7301   48.0729
   30.9031   48.5592   41.9881
   62.5700   50.5569   51.9326
   24.7184   50.5371   42.2493
         0   53.3573   43.2435
   53.3573         0   34.3948
   43.2435   34.3948         0

   62.5700

     6     8

Closest euclidean distance: Article: 6 and Article: 8
Angle Distance
  Columns 1 through 4

   0.0000 + 0.0000i   0.7779 + 0.0000i   0.8375 + 0.0000i   0.8407 + 0.0000i
   0.7779 + 0.0000i   0.0000 + 0.0000i   0.7762 + 0.0000i   0.7839 + 0.0000i
   0.8375 + 0.0000i   0.7762 + 0.0000i   0.0000 + 0.0000i   0.8760 + 0.0000i
   0.8407 + 0.0000i   0.7839 + 0.0000i   0.8760 + 0.0000i   0.0000 + 0.0000i
   0.8835 + 0.0000i   0.8853 + 0.0000i   0.7958 + 0.0000i   1.0011 + 0.0000i
   0.7751 + 0.0000i   0.7583 + 0.0000i   0.8337 + 0.0000i   0.7399 + 0.0000i
   0.8143 + 0.0000i   0.8331 + 0.0000i   0.9138 + 0.0000i   0.8792 + 0.0000i
   0.9034 + 0.0000i   0.8957 + 0.0000i   0.9411 + 0.0000i   0.9905 + 0.0000i
   0.7736 + 0.0000i   0.7396 + 0.0000i   0.7287 + 0.0000i   0.8278 + 0.0000i
   0.7483 + 0.0000i   0.7381 + 0.0000i   0.7443 + 0.0000i   0.8345 + 0.0000i

  Columns 5 through 8

   0.8835 + 0.0000i   0.7751 + 0.0000i   0.8143 + 0.0000i   0.9034 + 0.0000i
   0.8853 + 0.0000i   0.7583 + 0.0000i   0.8331 + 0.0000i   0.8957 + 0.0000i
   0.7958 + 0.0000i   0.8337 + 0.0000i   0.9138 + 0.0000i   0.9411 + 0.0000i
   1.0011 + 0.0000i   0.7399 + 0.0000i   0.8792 + 0.0000i   0.9905 + 0.0000i
   0.0000 + 0.0000i   0.8721 + 0.0000i   0.8925 + 0.0000i   1.0032 + 0.0000i
   0.8721 + 0.0000i   0.0000 + 0.0000i   0.7648 + 0.0000i   0.8864 + 0.0000i
   0.8925 + 0.0000i   0.7648 + 0.0000i   0.0000 + 0.0000i   0.8625 + 0.0000i
   1.0032 + 0.0000i   0.8864 + 0.0000i   0.8625 + 0.0000i   0.0000 + 0.0000i
   0.7587 + 0.0000i   0.7098 + 0.0000i   0.7819 + 0.0000i   0.8305 + 0.0000i
   0.7788 + 0.0000i   0.7498 + 0.0000i   0.7602 + 0.0000i   0.7447 + 0.0000i

  Columns 9 through 10

   0.7736 + 0.0000i   0.7483 + 0.0000i
   0.7396 + 0.0000i   0.7381 + 0.0000i
   0.7287 + 0.0000i   0.7443 + 0.0000i
   0.8278 + 0.0000i   0.8345 + 0.0000i
   0.7587 + 0.0000i   0.7788 + 0.0000i
   0.7098 + 0.0000i   0.7498 + 0.0000i
   0.7819 + 0.0000i   0.7602 + 0.0000i
```

```
   0.8305 + 0.0000i   0.7447 + 0.0000i
   0.0000 + 0.0000i   0.5327 + 0.0000i
   0.5327 + 0.0000i   0.0000 + 0.0000i

    1.0032

     5     8

Closest angle distance: Article: 5 and Article: 8
Question b
Euclidean Distance
  Columns 1 through 7

         0    0.0852    0.0977    0.0857    0.1081    0.0860    0.0988
    0.0852         0    0.0930    0.0841    0.1101    0.0866    0.1025
    0.0977    0.0930         0    0.1002    0.1031    0.1000    0.1157
    0.0857    0.0841    0.1002         0    0.1188    0.0809    0.1045
    0.1081    0.1101    0.1031    0.1188         0    0.1092    0.1171
    0.0860    0.0866    0.1000    0.0809    0.1092         0    0.0952
    0.0988    0.1025    0.1157    0.1045    0.1171    0.0952         0
    0.1060    0.1075    0.1172    0.1131    0.1289    0.1071    0.1106
    0.0895    0.0877    0.0903    0.0938    0.0979    0.0849    0.0992
    0.0826    0.0840    0.0896    0.0896    0.0983    0.0861    0.0945

  Columns 8 through 10

    0.1060    0.0895    0.0826
    0.1075    0.0877    0.0840
    0.1172    0.0903    0.0896
    0.1131    0.0938    0.0896
    0.1289    0.0979    0.0983
    0.1071    0.0849    0.0861
    0.1106    0.0992    0.0945
         0    0.1033    0.0909
    0.1033         0    0.0643
    0.0909    0.0643         0

    0.1289

     5     8

Closest euclidean distance: Article: 5 and Article: 8
Angle Distance
  Columns 1 through 4

   0.0000 + 0.0000i   0.7779 + 0.0000i   0.8375 + 0.0000i   0.8407 + 0.0000i
   0.7779 + 0.0000i   0.0000 + 0.0000i   0.7762 + 0.0000i   0.7839 + 0.0000i
   0.8375 + 0.0000i   0.7762 + 0.0000i   0.0000 + 0.0000i   0.8760 + 0.0000i
   0.8407 + 0.0000i   0.7839 + 0.0000i   0.8760 + 0.0000i   0.0000 + 0.0000i
   0.8835 + 0.0000i   0.8853 + 0.0000i   0.7958 + 0.0000i   1.0011 + 0.0000i
   0.7751 + 0.0000i   0.7583 + 0.0000i   0.8337 + 0.0000i   0.7399 + 0.0000i
   0.8143 + 0.0000i   0.8331 + 0.0000i   0.9138 + 0.0000i   0.8792 + 0.0000i
   0.9034 + 0.0000i   0.8957 + 0.0000i   0.9411 + 0.0000i   0.9905 + 0.0000i
   0.7736 + 0.0000i   0.7396 + 0.0000i   0.7287 + 0.0000i   0.8278 + 0.0000i
```

```
   0.7483 + 0.0000i   0.7381 + 0.0000i   0.7443 + 0.0000i   0.8345 + 0.0000i

  Columns 5 through 8

   0.8835 + 0.0000i   0.7751 + 0.0000i   0.8143 + 0.0000i   0.9034 + 0.0000i
   0.8853 + 0.0000i   0.7583 + 0.0000i   0.8331 + 0.0000i   0.8957 + 0.0000i
   0.7958 + 0.0000i   0.8337 + 0.0000i   0.9138 + 0.0000i   0.9411 + 0.0000i
   1.0011 + 0.0000i   0.7399 + 0.0000i   0.8792 + 0.0000i   0.9905 + 0.0000i
   0.0000 + 0.0000i   0.8721 + 0.0000i   0.8925 + 0.0000i   1.0032 + 0.0000i
   0.8721 + 0.0000i   0.0000 + 0.0000i   0.7648 + 0.0000i   0.8864 + 0.0000i
   0.8925 + 0.0000i   0.7648 + 0.0000i   0.0000 + 0.0000i   0.8625 + 0.0000i
   1.0032 + 0.0000i   0.8864 + 0.0000i   0.8625 + 0.0000i   0.0000 + 0.0000i
   0.7587 + 0.0000i   0.7098 + 0.0000i   0.7819 + 0.0000i   0.8305 + 0.0000i
   0.7788 + 0.0000i   0.7498 + 0.0000i   0.7602 + 0.0000i   0.7447 + 0.0000i

  Columns 9 through 10

   0.7736 + 0.0000i   0.7483 + 0.0000i
   0.7396 + 0.0000i   0.7381 + 0.0000i
   0.7287 + 0.0000i   0.7443 + 0.0000i
   0.8278 + 0.0000i   0.8345 + 0.0000i
   0.7587 + 0.0000i   0.7788 + 0.0000i
   0.7098 + 0.0000i   0.7498 + 0.0000i
   0.7819 + 0.0000i   0.7602 + 0.0000i
   0.8305 + 0.0000i   0.7447 + 0.0000i
   0.0000 + 0.0000i   0.5327 + 0.0000i
   0.5327 + 0.0000i   0.0000 + 0.0000i

    1.0032

     5     8

Closest angle distance: Article: 5 and Article: 8
Question C
Euclidean Distance
  Columns 1 through 7

         0    0.1075    0.1169    0.1043    0.1176    0.0978    0.1131
    0.1075         0    0.1201    0.1039    0.1251    0.1050    0.1207
    0.1169    0.1201         0    0.1148    0.1312    0.1167    0.1265
    0.1043    0.1039    0.1148         0    0.1217    0.0983    0.1166
    0.1176    0.1251    0.1312    0.1217         0    0.1167    0.1246
    0.0978    0.1050    0.1167    0.0983    0.1167         0    0.1100
    0.1131    0.1207    0.1265    0.1166    0.1246    0.1100         0
    0.1210    0.1254    0.1327    0.1246    0.1344    0.1201    0.1303
    0.0990    0.1071    0.1143    0.1027    0.1122    0.0972    0.1097
    0.0968    0.1051    0.1130    0.1016    0.1118    0.0960    0.1073

  Columns 8 through 10

    0.1210    0.0990    0.0968
    0.1254    0.1071    0.1051
    0.1327    0.1143    0.1130
    0.1246    0.1027    0.1016
```

```
    0.1344    0.1122    0.1118
    0.1201    0.0972    0.0960
    0.1303    0.1097    0.1073
         0    0.1136    0.0965
    0.1136         0    0.0821
    0.0965    0.0821         0

    0.1344

     5     8

TF-IDF: Closest euclidean distance: Article: 5 and Article: 8
Angle Distance
  Columns 1 through 4

   0.0000 + 0.0000i   1.4965 + 0.0000i   1.5440 + 0.0000i   1.5177 + 0.0000i
   1.4965 + 0.0000i   0.0000 + 0.0000i   1.4857 + 0.0000i   1.3853 + 0.0000i
   1.5440 + 0.0000i   1.4857 + 0.0000i   0.0000 + 0.0000i   1.4603 + 0.0000i
   1.5177 + 0.0000i   1.3853 + 0.0000i   1.4603 + 0.0000i   0.0000 + 0.0000i
   1.5269 + 0.0000i   1.5389 + 0.0000i   1.5422 + 0.0000i   1.5470 + 0.0000i
   1.4762 + 0.0000i   1.4659 + 0.0000i   1.5532 + 0.0000i   1.4237 + 0.0000i
   1.5354 + 0.0000i   1.5431 + 0.0000i   1.5378 + 0.0000i   1.5424 + 0.0000i
   1.5464 + 0.0000i   1.5104 + 0.0000i   1.5346 + 0.0000i   1.5575 + 0.0000i
   1.5344 + 0.0000i   1.5342 + 0.0000i   1.5417 + 0.0000i   1.5382 + 0.0000i
   1.5121 + 0.0000i   1.5149 + 0.0000i   1.5327 + 0.0000i   1.5362 + 0.0000i

  Columns 5 through 8

   1.5269 + 0.0000i   1.4762 + 0.0000i   1.5354 + 0.0000i   1.5464 + 0.0000i
   1.5389 + 0.0000i   1.4659 + 0.0000i   1.5431 + 0.0000i   1.5104 + 0.0000i
   1.5422 + 0.0000i   1.5532 + 0.0000i   1.5378 + 0.0000i   1.5346 + 0.0000i
   1.5470 + 0.0000i   1.4237 + 0.0000i   1.5424 + 0.0000i   1.5575 + 0.0000i
   0.0000 + 0.0000i   1.5250 + 0.0000i   1.4864 + 0.0000i   1.5389 + 0.0000i
   1.5250 + 0.0000i   0.0000 + 0.0000i   1.4940 + 0.0000i   1.5437 + 0.0000i
   1.4864 + 0.0000i   1.4940 + 0.0000i   0.0000 + 0.0000i   1.5409 + 0.0000i
   1.5389 + 0.0000i   1.5437 + 0.0000i   1.5409 + 0.0000i   0.0000 + 0.0000i
   1.4754 + 0.0000i   1.5170 + 0.0000i   1.5194 + 0.0000i   1.4607 + 0.0000i
   1.4825 + 0.0000i   1.5139 + 0.0000i   1.4907 + 0.0000i   1.1948 + 0.0000i

  Columns 9 through 10

   1.5344 + 0.0000i   1.5121 + 0.0000i
   1.5342 + 0.0000i   1.5149 + 0.0000i
   1.5417 + 0.0000i   1.5327 + 0.0000i
   1.5382 + 0.0000i   1.5362 + 0.0000i
   1.4754 + 0.0000i   1.4825 + 0.0000i
   1.5170 + 0.0000i   1.5139 + 0.0000i
   1.5194 + 0.0000i   1.4907 + 0.0000i
   1.4607 + 0.0000i   1.1948 + 0.0000i
   0.0000 + 0.0000i   1.2867 + 0.0000i
   1.2867 + 0.0000i   0.0000 + 0.0000i

    1.5575
```

```
     4     8

TF_IDF: Closest angle distance: Article: 4 and Article: 8
```

*Published with MATLAB® R2023b*