_

Inna Williams
###

Four documents:

d1: Shipment of gold damaged in a fire.
d2: Delivery of silver arrived in a silver truck.
d3: Shipment of gold arrived in a truck.
d4: Shipment of gold arrived in a silver truck
###

1.

Repeat the rank-2 analysis for the query q="gold silver truck" and report your findings. A cosine similarity test threshold of 0.5 is considered relevant (this means that the angle between the vectors is < 60 degreees), while a 0.8 threshold is very good.

```
> restart( ) :
> with(LinearAlgebra) : with(plots) : interface(rtablesize = 15) :
> A := Matrix(11, 4, [⟨1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0⟩, ⟨1, 1, 0, 1, 0, 0, 1, 1, 0, 2, 1⟩, ⟨1, 1, 0, 0, 0, 1,
      1, 1, 1, 0, 1⟩, ⟨1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1⟩ ]);
```

$$A := \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad (1)$$

New query vector

```
> q_1 := ⟨0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1⟩
```

(2)

$$q\_1 := \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (2)$$

> $U, S, Vt := SingularValues(A, output = ['U','S','Vt'])$:

#Create the rank 2 approximation

> $Sm2 := DiagonalMatrix\left(S_{1..2}, 2, 2\right)$

$$Sm2 := \begin{bmatrix} 4.93115367652474 & 0. \\ 0. & 2.36166782730918 \end{bmatrix} \quad (3)$$

The rows of Um2 provide the coordinates of the 11 words

> $Um2 := SubMatrix(U, 1..11, 1..2)$

$$Um2 := \begin{bmatrix} -0.402466375103054 & -0.0770159571941109 \\ -0.321939474302626 & 0.199720711331478 \\ -0.0805269008004283 & -0.276736668525589 \\ -0.107635642500810 & 0.302469353467979 \\ -0.0805269008004284 & -0.276736668525589 \\ -0.294830732602244 & -0.379485310662090 \\ -0.402466375103054 & -0.0770159571941109 \\ -0.402466375103054 & -0.0770159571941109 \\ -0.294830732602244 & -0.379485310662090 \\ -0.329192168331150 & 0.608076165234006 \\ -0.321939474302626 & 0.199720711331478 \end{bmatrix} \quad (4)$$

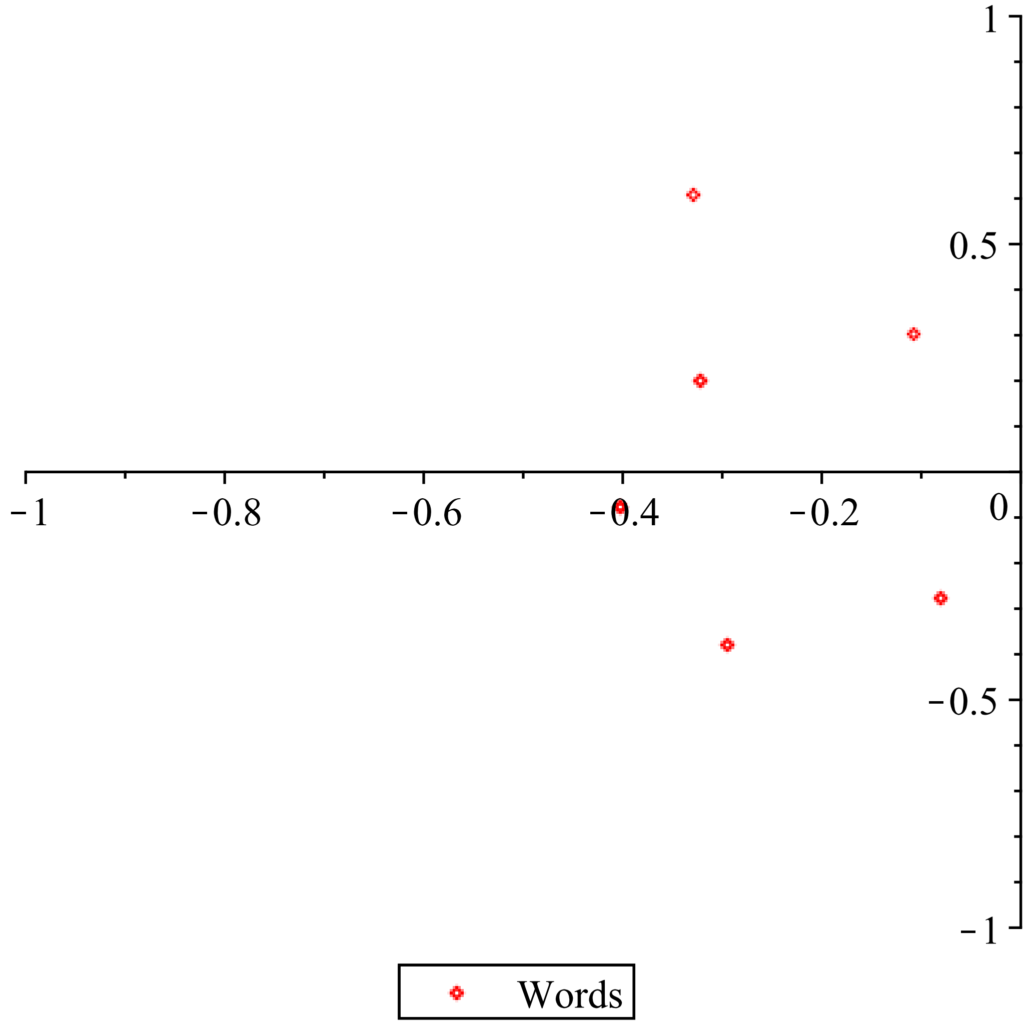
The columns of Vtm2 provide the coordinates of the 4 documents

> $Vtm2 := SubMatrix(Vt, 1..2, 1..4)$

$Vtm2 := [[-0.397090522941175, -0.530767894242970, -0.495003745419503, -0.561761382663360], [-0.653560086693608, 0.714332140832334, -0.250067796755502, 0.00740963432202280]]$ **(5)**

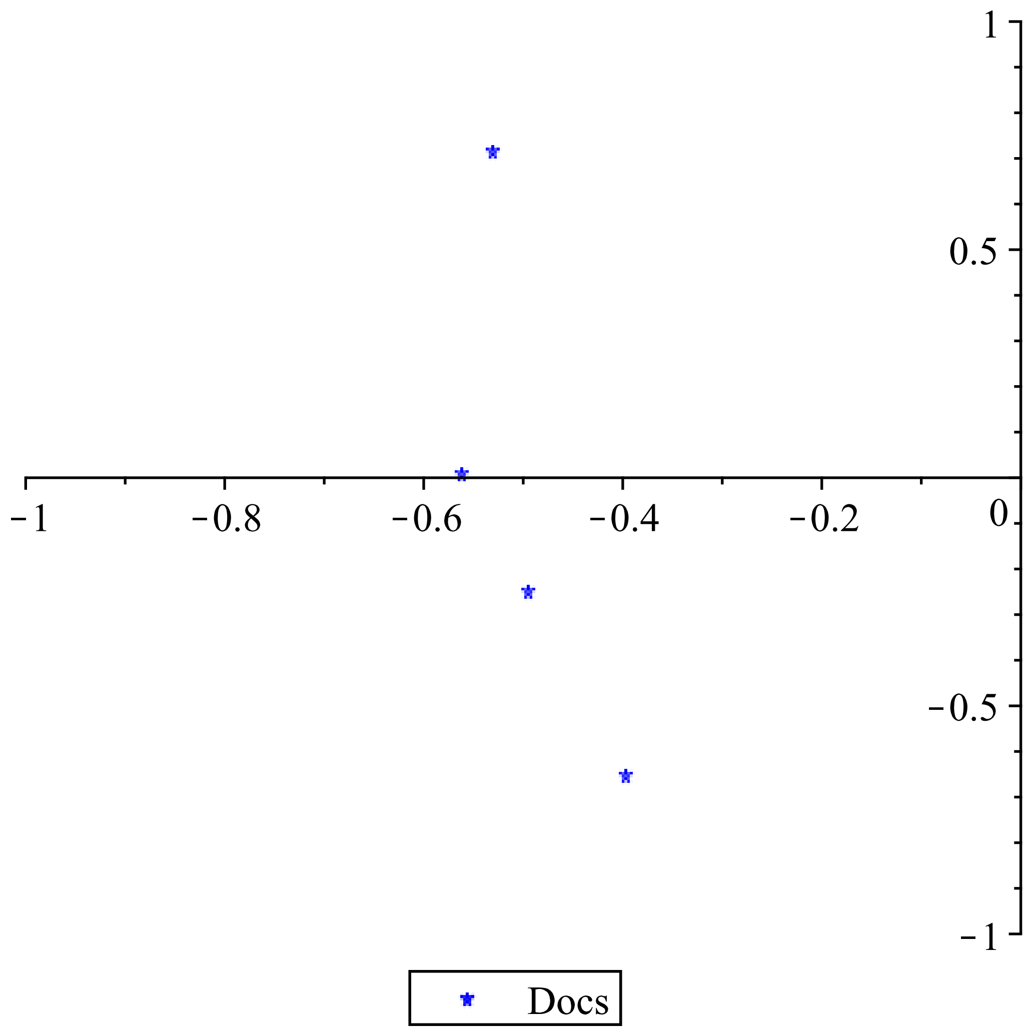
Plot the 11 words, some do overlap

```
> plot1_2d := pointplot(Um2, legend = "Words", color = red, axes = normal, view = [-1 ..0, -1 ..1]) : display(plot1_2d)
```



Plot the 4 documents

```
> plot2_2d := pointplot(Vtm2, legend = "Docs", color = blue, symbol = asterisk, axes = normal, view = [-1 ..0, -1 ..1]) : display(plot2_2d)
```



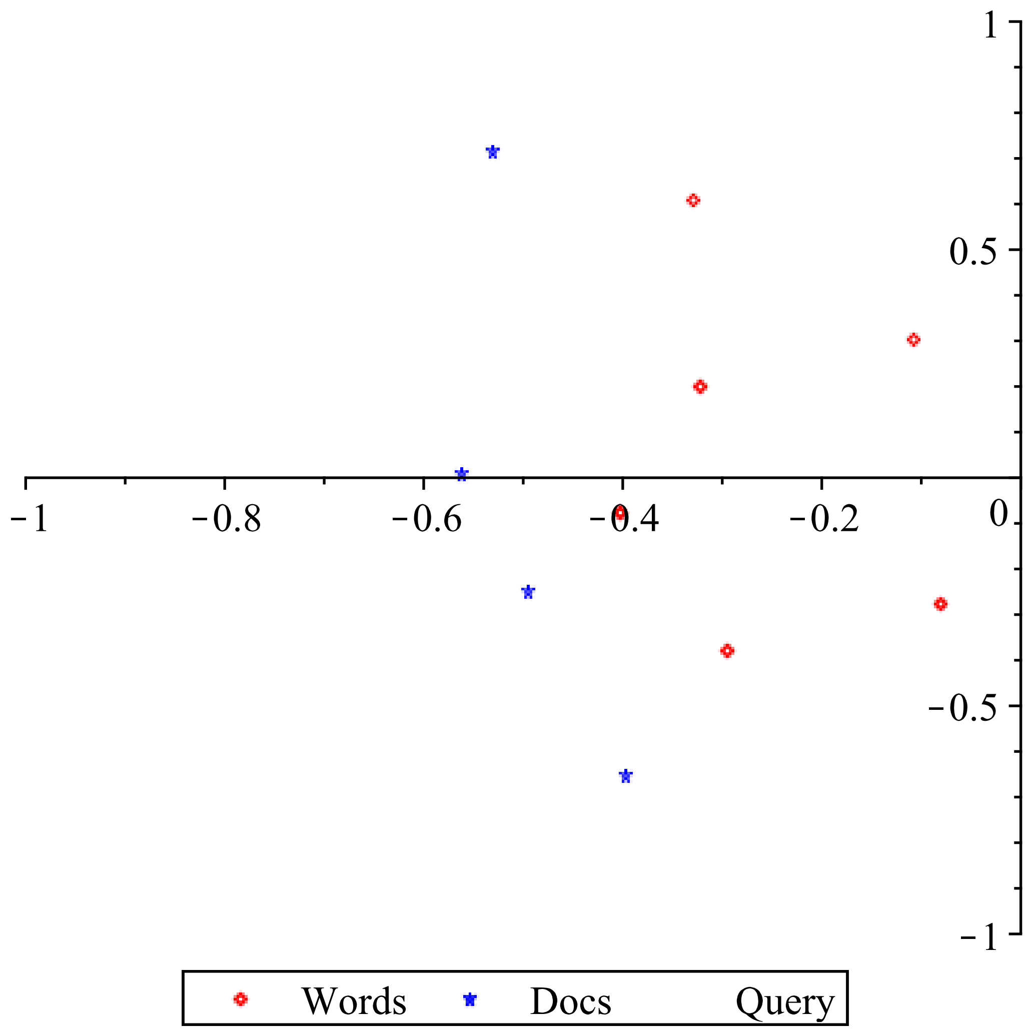
#We compute the coordinates of the new query

> $qcoord\_2 := Transpose(q\_1).(Um2.MatrixInverse(Sm2))$

$$qcoord\_2 := \begin{bmatrix} -0.191833886609409 & 0.181359783518498 \end{bmatrix} \quad (6)$$

Combine the plots

> $plot3\_2d := pointplot(qcoord\_2, symbol = box, legend = \text{"Query"}, axes = normal) :$
> $display(plot1\_2d, plot2\_2d, plot3\_2d)$



We run some similarity tests between the four documents and the new query.

> $doc1 := Column(Vtm2, 1) : test1 := \dfrac{qcoord\_2 \cdot doc1}{Norm(qcoord\_2, 2) \cdot Norm(doc1, 2)}$

$$test1 := -0.2097941873 \tag{7}$$

> $doc2 := Column(Vtm2, 2) : test2 := \dfrac{qcoord\_2 \cdot doc2}{Norm(qcoord\_2, 2) \cdot Norm(doc2, 2)}$

$$test2 := 0.9848249865 \tag{8}$$

> $doc3 := Column(Vtm2, 3) : test3 := \dfrac{qcoord\_2 \cdot doc3}{Norm(qcoord\_2, 2) \cdot Norm(doc3, 2)}$

$$test3 := 0.3388280712 \tag{9}$$

> $doc4 := Column(Vtm2, 4) : test4 := \dfrac{qcoord\_2 \cdot doc4}{Norm(qcoord\_2, 2) \cdot Norm(doc4, 2)}$

$$test4 := 0.7356638059 \tag{10}$$

Answer: for Rank 2 analysis cosine similarity test
for query q="gold silver truck"shows :
Very good for
doc2 =0.9848

relevant for doc4=0.7357

##

2

Perform a rank-3 SVD analysis for the same query q="gold silver truck" and report your findings

****

> $Sm3 := DiagonalMatrix(S_{1..3}, 3, 3);$

$$Sm3 := \begin{bmatrix} 4.93115367652474 & 0. & 0. \\ 0. & 2.36166782730918 & 0. \\ 0. & 0. & 1.37385875035970 \end{bmatrix} \quad (11)$$

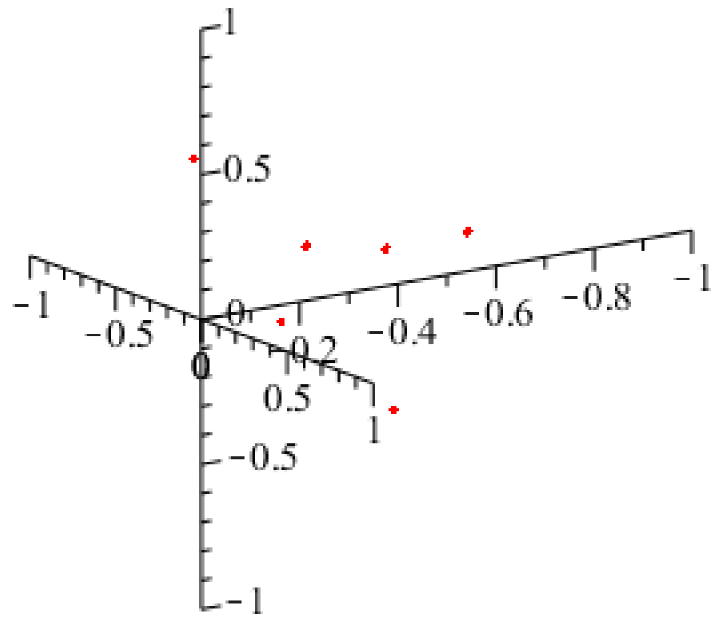
> $Um3 := SubMatrix(U, 1..11, 1..3);$

$$Um3 := \begin{bmatrix} -0.402466375103054 & -0.0770159571941109 & 0.103979350121338 \\ -0.321939474302626 & 0.199720711331478 & -0.364947747899966 \\ -0.0805269008004283 & -0.276736668525589 & 0.468927098021304 \\ -0.107635642500810 & 0.302469353467979 & 0.286328369561831 \\ -0.0805269008004284 & -0.276736668525589 & 0.468927098021304 \\ -0.294830732602244 & -0.379485310662090 & -0.182349019440493 \\ -0.402466375103054 & -0.0770159571941109 & 0.103979350121338 \\ -0.402466375103054 & -0.0770159571941109 & 0.103979350121338 \\ -0.294830732602244 & -0.379485310662090 & -0.182349019440493 \\ -0.329192168331150 & 0.608076165234006 & 0.336035199548203 \\ -0.321939474302626 & 0.199720711331478 & -0.364947747899966 \end{bmatrix} \quad (12)$$

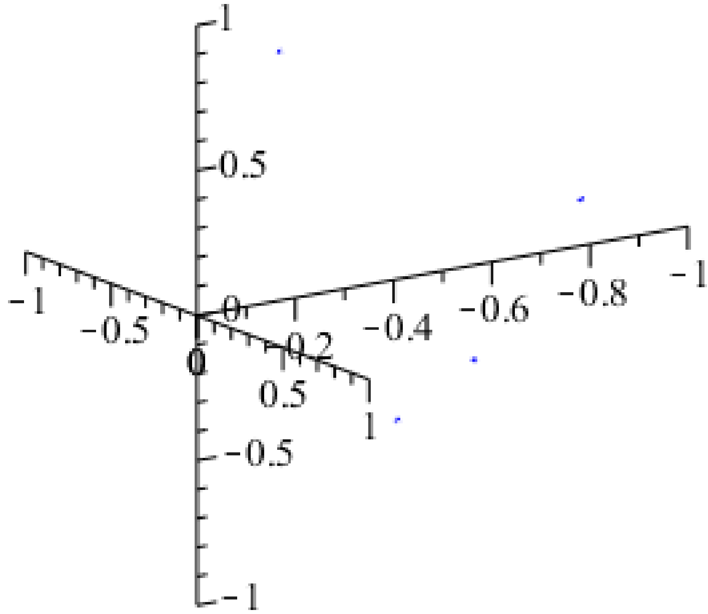
> $Vtm3 := SubMatrix(Vt, 1..3, 1..4)$

$Vtm3 := [[-0.397090522941175, -0.530767894242970, -0.495003745419503, -0.561761382663360],$
$[-0.653560086693608, 0.714332140832334, -0.250067796755502, 0.00740963432202280],$
$[0.644239596897351, 0.393374735998748, -0.569676820205855, -0.325084572669329]]$ **(13)**

> $plot1\_3d := pointplot3d(Um3, color = red, axes = normal, view = [-1..0, -1..1, -1..1]):$
> $display(plot1\_3d);$



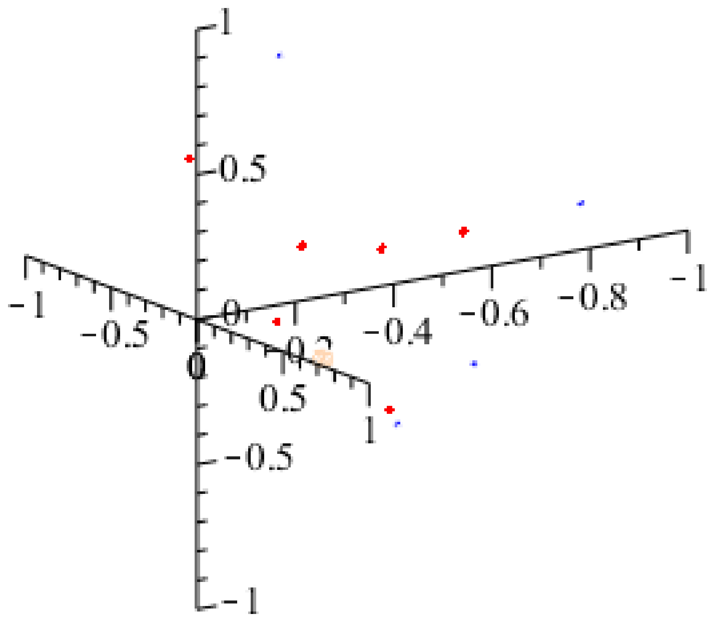
> $plot2\_3d := pointplot3d(Vtm3, color = blue, symbol = asterisk, axes = normal, view = [-1 ..0, -1 ..1, -1 ..1]) : display(plot2\_3d);$



> $qcoord\_3 := Transpose(q\_1).(Um3 \cdot MatrixInverse(Sm3));$

$$qcoord\_3 := \begin{bmatrix} -0.191833886609409 & 0.181359783518498 & -0.153772407634296 \end{bmatrix} \qquad \textbf{(14)}$$

> $plot3\_3d := pointplot3d(qcoord\_3, symbol = box, axes = normal) : display(plot1\_3d, plot2\_3d, plot3\_3d);$



> $doc1 := Column(Vtm3, 1) : test1 := \dfrac{qcoord\_3 \cdot doc1}{Norm(qcoord\_3, 2) \cdot Norm(doc1, 2)}$

$$test1 := -0.4629272750 \quad (15)$$

> $doc2 := Column(Vtm3, 2) : test2 := \dfrac{qcoord\_3 \cdot doc2}{Norm(qcoord\_3, 2) \cdot Norm(doc2, 2)}$

$$test2 := 0.5748450260 \quad (16)$$

> $doc3 := Column(Vtm3, 3) : test3 := \dfrac{qcoord\_3 \cdot doc3}{Norm(qcoord\_3, 2) \cdot Norm(doc3, 2)}$

$$test3 := 0.5648804643 \quad (17)$$

> $doc4 := Column(Vtm3, 4) : test4 := \dfrac{qcoord\_3 \cdot doc4}{Norm(qcoord\_3, 2) \cdot Norm(doc4, 2)}$

$$test4 := 0.8022956779 \quad (18)$$

Answer: for Rank 3 analysis cosine similarity test
for query q="gold silver truck"shows :
Very good for
doc4=0.8022
relevant:
doc2=0.5748
doc3=0.5649

We can see that rank 3 has better approximation
###
3
Repeat steps 1 and 2 for a new query q="shipment truck" and report your findings.
****

****
Analysis for rank 2
****

New query vector

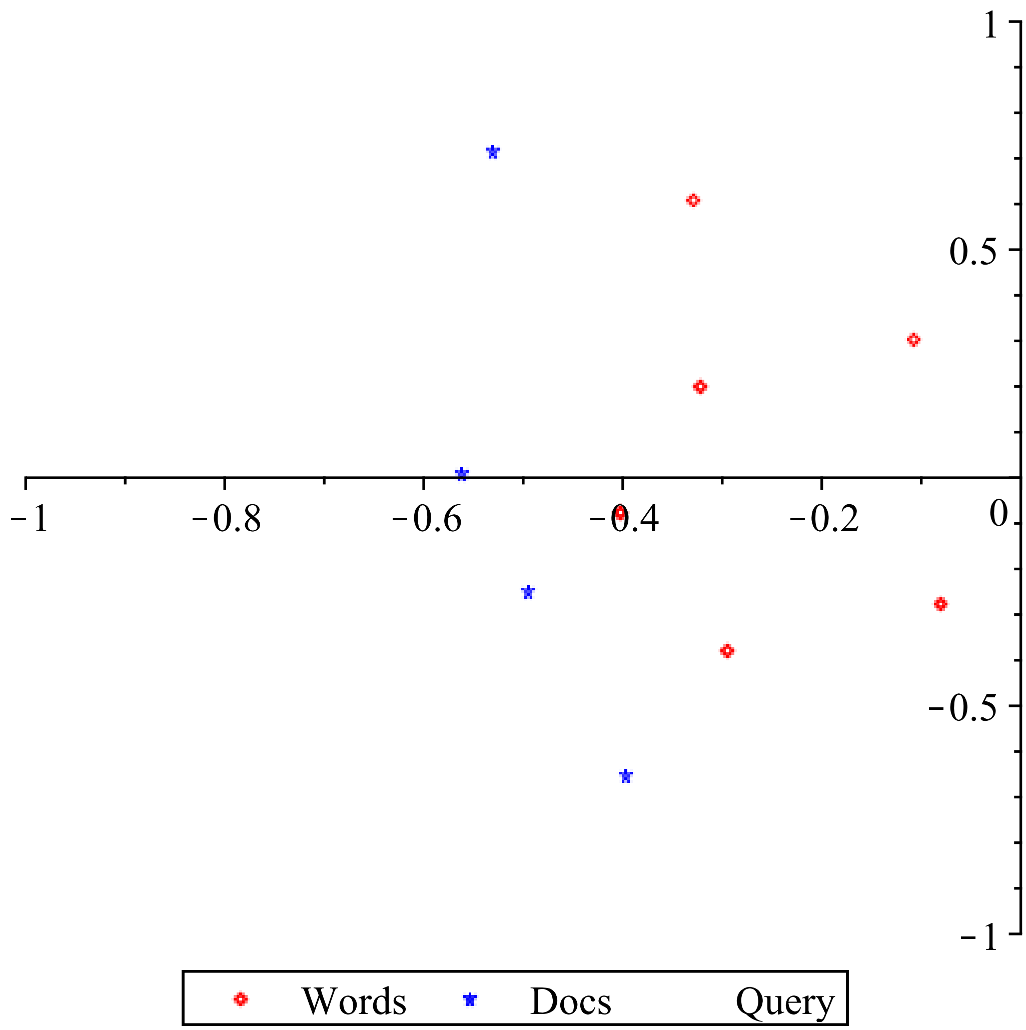
> $q\_2 := \langle 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1 \rangle$

$$q\_2 := \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \tag{19}$$

> $qcoord\_2 := Transpose(q\_2).(Um2.MatrixInverse(Sm2))$

$$qcoord\_2 := \begin{bmatrix} -0.125076249365553 & -0.0761176475590267 \end{bmatrix} \tag{20}$$

> $plot3\_2d := pointplot(qcoord\_2, symbol = box, legend = \text{"Query"}, axes = normal) :$
> $display(plot1\_2d, plot2\_2d, plot3\_2d)$



>

> $doc1 := Column(Vtm2, 1) : test1 := \dfrac{qcoord\_2 \cdot doc1}{Norm(qcoord\_2, 2) \cdot Norm(doc1, 2)}$

$$test1 := 0.8878595363 \quad \textbf{(21)}$$

> $doc2 := Column(Vtm2, 2) : test2 := \dfrac{qcoord\_2 \cdot doc2}{Norm(qcoord\_2, 2) \cdot Norm(doc2, 2)}$

$$test2 := 0.09219504386 \quad \textbf{(22)}$$

> $doc3 := Column(Vtm2, 3) : test3 := \dfrac{qcoord\_2 \cdot doc3}{Norm(qcoord\_2, 2) \cdot Norm(doc3, 2)}$

$$test3 := 0.9968881269 \quad \textbf{(23)}$$

> $doc4 := Column(Vtm2, 4) : test4 := \dfrac{qcoord\_2 \cdot doc4}{Norm(qcoord\_2, 2) \cdot Norm(doc4, 2)}$

$$test4 := 0.8473154702 \quad \textbf{(24)}$$

****

Answer: for Rank 2 analysis cosine similarity test
for query q="shipment truck"shows :
Very good for
doc3=0.9969

doc1=0.8879
doc4= 0.8473

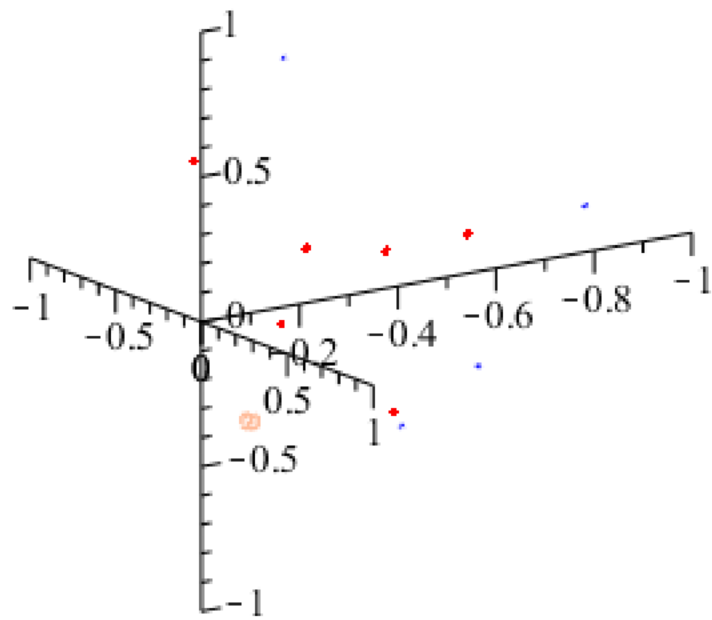
****
Analysis for rank 3
****

> $qcoord\_3 := Transpose(q\_2).(Um3 \cdot MatrixInverse(Sm3))$

$$qcoord\_3 := \begin{bmatrix} -0.125076249365553 & -0.0761176475590267 & -0.398364655170822 \end{bmatrix} \quad \textbf{(25)}$$

Combine the plots

> $plot3\_3d := pointplot3d(qcoord\_3, symbol = box, axes = normal) : display(plot1\_3d, plot2\_3d, plot3\_3d)$



> $doc1 := Column(Vtm3, 1) : test1 := \dfrac{qcoord\_3 \cdot doc1}{Norm(qcoord\_3, 2) \cdot Norm(doc1, 2)}$

$$test1 := -0.3704790850 \quad \textbf{(26)}$$

> $doc2 := Column(Vtm3, 2) : test2 := \dfrac{qcoord\_3 \cdot doc2}{Norm(qcoord\_3, 2) \cdot Norm(doc2, 2)}$

$$test2 := -0.3503806681 \quad \textbf{(27)}$$

> $doc3 := Column(Vtm3, 3) : test3 := \dfrac{qcoord\_3 \cdot doc3}{Norm(qcoord\_3, 2) \cdot Norm(doc3, 2)}$

$$test3 := 0.9124401690 \qquad (28)$$

> $doc4 := Column(Vtm3, 4) : test4 := \dfrac{qcoord\_3 \cdot doc4}{Norm(qcoord\_3, 2) \cdot Norm(doc4, 2)}$

$$test4 := 0.7230941199 \qquad (29)$$

Answer: for Rank 2 analysis cosine similarity test
for query q="shipment truck"
Very good for documents are
doc3 =0.9124
doc4=0.7231
Rank 3 has better approximation
##

II. Consider the following term-document example :

The $t = 6$ **terms**:

T1: bak(e,ing)
T2: recipes
T3: bread
T4: cake
T5: pastr(y,ies)
T6: pie

The $d = 5$ **document** titles:

D1: How to Bake Bread Without Recipes
D2: The Classic Art of Viennese Pastry
D3: Numerical Recipes: The Art of Scientific Computing
D4: Breads, Pastries, Pies and Cakes: Quantity Baking Recipes
D5: Pastry: A Book of Best French Recipes

and the query: **q=baking bread**

****

> $A := Matrix(6, 5, [\langle 1, 1, 1, 0, 0, 0\rangle, \langle 0, 0, 0, 0, 1, 0\rangle, \langle 0, 1, 0, 0, 0, 0\rangle, \langle 1, 1, 1, 1, 1, 1\rangle, \langle 0, 1, 0, 0, 1, 0\rangle]);$

(30)

$$A := \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{30}$$

> $q := \langle 1, 0, 1, 0, 0, 0 \rangle$

$$q := \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{31}$$

> $U, S, Vt := \mathit{SingularValues}(A, \mathit{output} = ['U','S','Vt'])$

$U, S, Vt := [[-0.415139716033301, 0.408248290463863, 0.162729146205422,$ **(32)**
$-0.366758196200682, -0.707106781186548, -3.88923546639465\ 10^{-15}],$
$[-0.590723295429479, 4.44089209850063\ 10^{-16}, -0.729386490672964,$
$0.345023673768503, -2.22044604925031\ 10^{-16}, 3.42334571175266\ 10^{-18}],$
$[-0.415139716033301, 0.408248290463863, 0.162729146205422,$
$-0.366758196200682, 0.707106781186548, 3.88581212068290\ 10^{-15}],$
$[-0.258865830839363, 0., 0.440768092169533, 0.488581590472949,$
$3.63598040564739\ 10^{-15}, -0.707106781186548],$
$[-0.415139716033301, -0.816496580927726, 0.162729146205422,$
$-0.366758196200682, 3.60822483003176\ 10^{-16}, 4.37348322807471\ 10^{-18}],$
$[-0.258865830839363, 5.55111512312578\ 10^{-17}, 0.440768092169533,$
$0.488581590472949, -4.13558076672871\ 10^{-15}, 0.707106781186548]],$

$$\begin{bmatrix} 3.01546471512117 \\ 1.41421356237309 \\ 1.20531141460241 \\ 0.673941351816550 \\ 1.76346875900021\ 10^{-17} \\ 0. \end{bmatrix}, [[-0.471238386697216, -0.137670228390193,$$

$-0.195897929916829, -0.780600778846626, -0.333568158307022],$
$[0.577350269189626, -0.577350269189626, 9.71445146547012\ 10^{-17},$
$-1.11022302462516\ 10^{-16}, -0.577350269189626],$

$[-0.335123515274566, 0.135010043241896, -0.605143601758358,$
$0.531262812684466, -0.470133558516463],$
$[-0.576448852093306, -0.544198979944052, 0.511949107794798,$
$0.329275337556019, -0.0322498721492537],$
$[0., 0.577350269189626, 0.577350269189626, -1.11022302462516\ 10^{-16},$
$-0.577350269189625]]$

> $Sm2 := DiagonalMatrix(S_{1..2}, 2, 2)$

$$Sm2 := \begin{bmatrix} 3.01546471512117 & 0. \\ 0. & 1.41421356237309 \end{bmatrix} \quad \textbf{(33)}$$

The rows of Um2 provide the coordinates of the 6 words

> $Um2 := SubMatrix(U, 1..6, 1..2)$

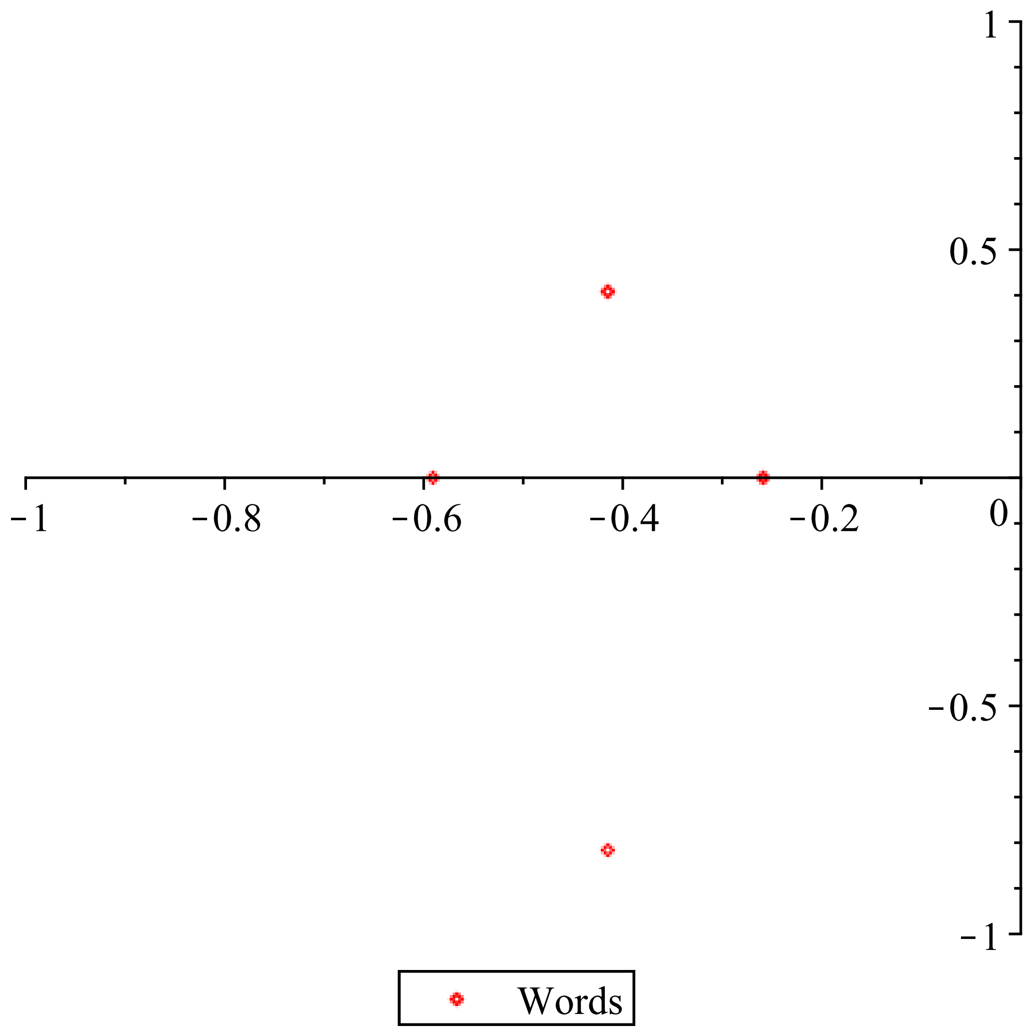
$$Um2 := \begin{bmatrix} -0.415139716033301 & 0.408248290463863 \\ -0.590723295429479 & 4.44089209850063\ 10^{-16} \\ -0.415139716033301 & 0.408248290463863 \\ -0.258865830839363 & 0. \\ -0.415139716033301 & -0.816496580927726 \\ -0.258865830839363 & 5.55111512312578\ 10^{-17} \end{bmatrix} \quad \textbf{(34)}$$

The columns of Vtm2 provide the coordinates of the 5 documents

> $Vtm2 := SubMatrix(Vt, 1..2, 1..5)$

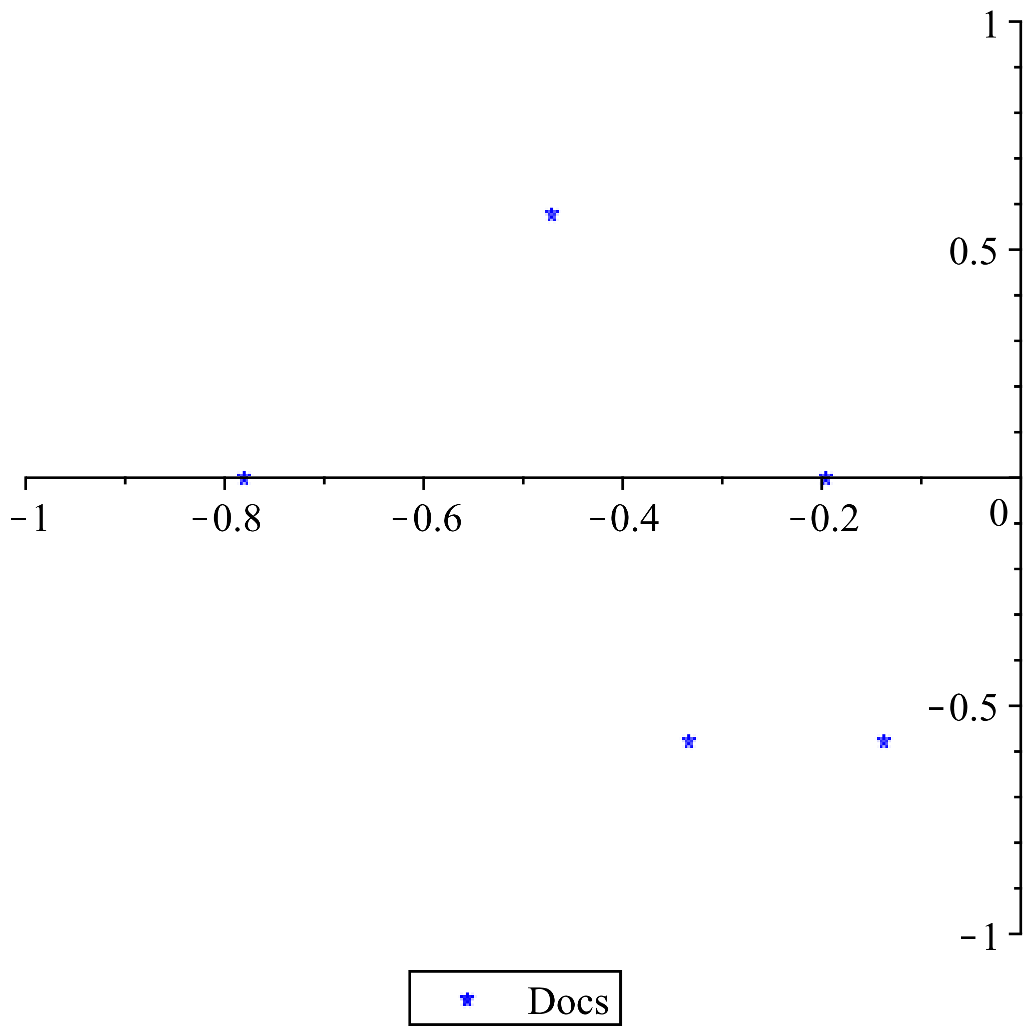
$Vtm2 := [[-0.471238386697216, -0.137670228390193, -0.195897929916829,$ **(35)**
$-0.780600778846626, -0.333568158307022],$
$[0.577350269189626, -0.577350269189626, 9.71445146547012\ 10^{-17},$
$-1.11022302462516\ 10^{-16}, -0.577350269189626]]$

> $plot1\_2d := pointplot(Um2, legend = "Words", color = red, axes = normal, view = [-1..0, -1..1]) : display(plot1\_2d)$



Plot the 5 documents

> $plot2\_2d := pointplot(Vtm2, legend = \text{"Docs"}, color = blue, symbol = asterisk, axes = normal,$
> $view = [-1 ..0, -1 ..1]) : display(plot2\_2d)$

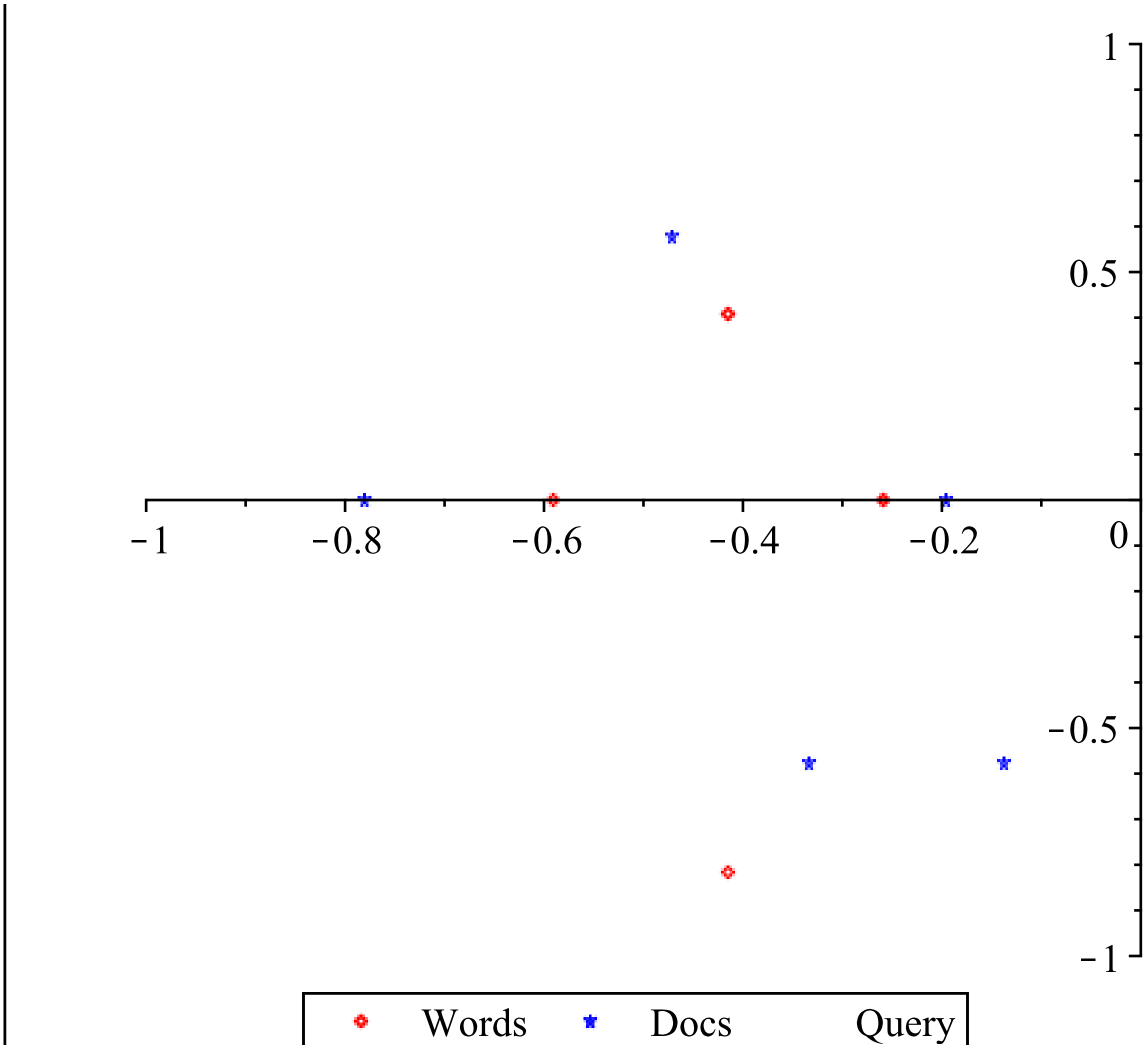


> $qcoord\_2 := Transpose(q).(Um2.MatrixInverse(Sm2))$

$$qcoord\_2 := \begin{bmatrix} -0.275340456780387 & 0.577350269189626 \end{bmatrix} \quad \textbf{(36)}$$

Combine the plots

> $plot3\_2d := pointplot(qcoord\_2, symbol = box, legend = \text{"Query"}, axes = normal)$:
> $display(plot1\_2d, plot2\_2d, plot3\_2d)$



> $doc1 := Column(Vtm2, 1) : test1 := \frac{qcoord\_2 \cdot doc1}{Norm(qcoord\_2, 2) \cdot Norm(doc1, 2)}$

$$test1 := 0.9714457632 \quad (37)$$

> $doc2 := Column(Vtm2, 2) : test2 := \frac{qcoord\_2 \cdot doc2}{Norm(qcoord\_2, 2) \cdot Norm(doc2, 2)}$

$$test2 := -0.7781502377 \quad (38)$$

> $doc3 := Column(Vtm2, 3) : test3 := \frac{qcoord\_2 \cdot doc3}{Norm(qcoord\_2, 2) \cdot Norm(doc3, 2)}$

$$test3 := 0.4304582471 \quad (39)$$

> $doc4 := Column(Vtm2, 4) : test4 := \frac{qcoord\_2 \cdot doc4}{Norm(qcoord\_2, 2) \cdot Norm(doc4, 2)}$

$$test4 := 0.4304582469 \quad (40)$$

```
> doc5 := Column(Vtm2, 5):test5 := (qcoord_2.doc5)/(Norm(qcoord_2, 2)*Norm(doc5, 2))
```

$$test5 := -0.5662031223 \quad (41)$$

****

Answer: for Rank 2 analysis cosine similarity test
for query q="baking bread "shows :
Very good for

doc1 =0.9714

the other document shows less then 0.5 (less then relevant)

****

> $Sm3 := DiagonalMatrix\left(S_{1..3}, 3, 3\right);$

$$Sm3 := \begin{bmatrix} 3.01546471512117 & 0. & 0. \\ 0. & 1.41421356237309 & 0. \\ 0. & 0. & 1.20531141460241 \end{bmatrix} \quad (42)$$

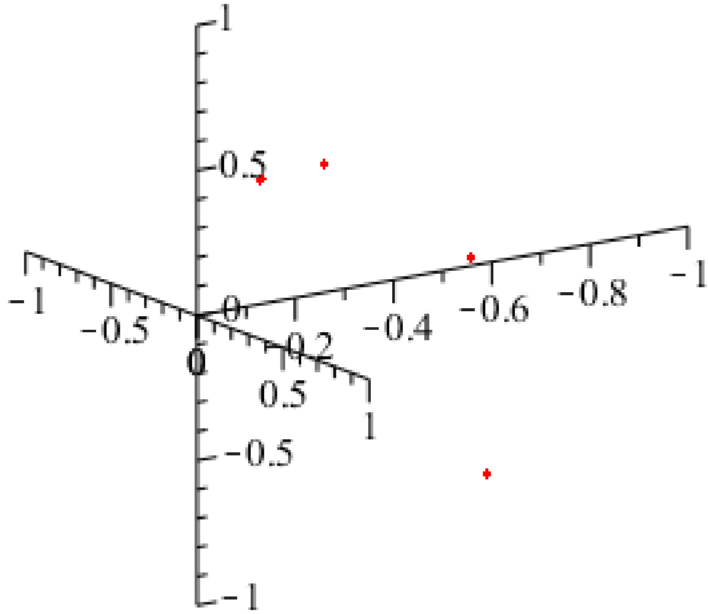
> $Um3 := SubMatrix(U, 1..6, 1..3);$

$$Um3 := \begin{bmatrix} -0.415139716033301 & 0.408248290463863 & 0.162729146205422 \\ -0.590723295429479 & 4.44089209850063\ 10^{-16} & -0.729386490672964 \\ -0.415139716033301 & 0.408248290463863 & 0.162729146205422 \\ -0.258865830839363 & 0. & 0.440768092169533 \\ -0.415139716033301 & -0.816496580927726 & 0.162729146205422 \\ -0.258865830839363 & 5.55111512312578\ 10^{-17} & 0.440768092169533 \end{bmatrix} \quad (43)$$

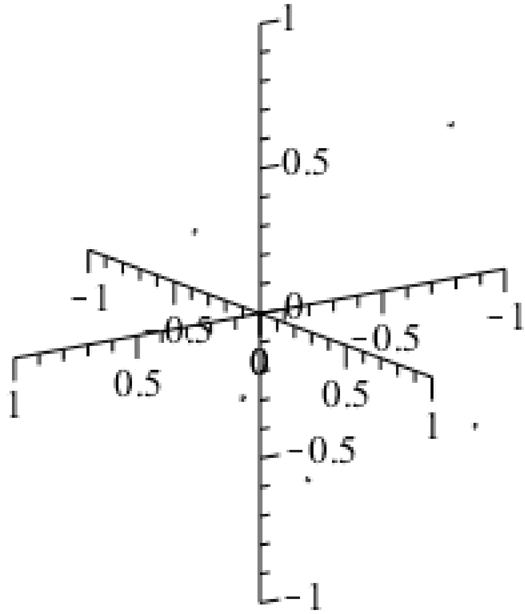
> $Vtm3 := SubMatrix(Vt, 1..3, 1..5)$

$$Vtm3 := [[-0.471238386697216, -0.137670228390193, -0.195897929916829, -0.780600778846626, -0.333568158307022], [0.577350269189626, -0.577350269189626, 9.71445146547012\ 10^{-17}, -1.11022302462516\ 10^{-16}, -0.577350269189626], [-0.335123515274566, 0.135010043241896, -0.605143601758358, 0.531262812684466, -0.470133558516463]] \quad (44)$$

> $plot1\_3d := pointplot3d(Um3, color = red, axes = normal, view = [-1..0, -1..1, -1..1]):$
> $display(plot1\_3d);$



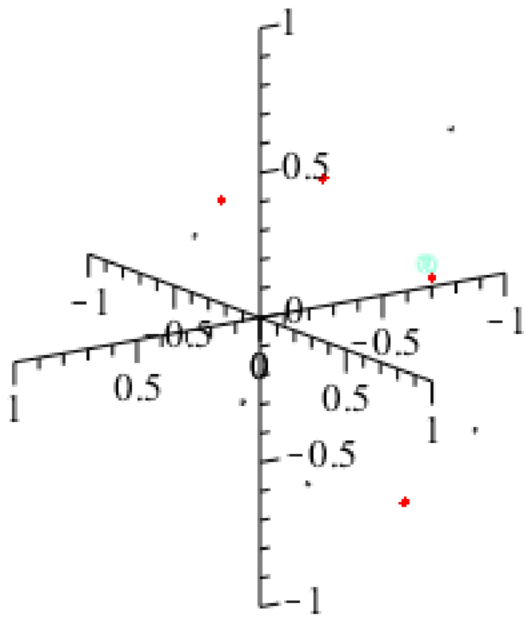
> $plot2\_3d := pointplot3d(Vtm3, color = black, symbol = asterisk, axes = normal, view = [-1 ..1, -1 ..1, -1 ..1]) : display(plot2\_3d);$



> $qcoord\_3 := Transpose(q).(Um3 \cdot MatrixInverse(Sm3));$

$$qcoord\_3 := \begin{bmatrix} -0.275340456780387 & 0.577350269189626 & 0.270020086483792 \end{bmatrix} \quad \textbf{(45)}$$

> $plot3\_3d := pointplot3d(qcoord\_3, symbol = box, axes = normal) : display(plot1\_3d, plot2\_3d, plot3\_3d);$



> $doc1 := Column(Vtm3, 1) : test1 := \dfrac{qcoord\_3 \cdot doc1}{Norm(qcoord\_3, 2) \cdot Norm(doc1, 2)}$

$$test1 := 0.6567412211 \tag{46}$$

> $doc2 := Column(Vtm3, 2) : test2 := \dfrac{qcoord\_3 \cdot doc2}{Norm(qcoord\_3, 2) \cdot Norm(doc2, 2)}$

$$test2 := -0.6127748991 \tag{47}$$

> $doc3 := Column(Vtm3, 3) : test3 := \dfrac{qcoord\_3 \cdot doc3}{Norm(qcoord\_3, 2) \cdot Norm(doc3, 2)}$

$$test3 := -0.2478655574 \tag{48}$$

> $doc4 := Column(Vtm3, 4) : test4 := \dfrac{qcoord\_3 \cdot doc4}{Norm(qcoord\_3, 2) \cdot Norm(doc4, 2)}$

$$test4 := 0.5466612853 \tag{49}$$

> $doc5 := Column(Vtm3, 4) : test5 := \dfrac{qcoord\_3 \cdot doc5}{Norm(qcoord\_3, 2) \cdot Norm(doc5, 2)}$

$$test5 := 0.5466612853 \tag{50}$$

****

Answer: for Rank 3 analysis cosine similarity test
for query q="baking bread "shows :
relevant for

best	**doc1 =0.6567**
less relavent	**doc4=doc5=0.5467**

****