Corrected Handwritten Solution for Semantic Analysis

This document provides a detailed handwritten solution for the semantic analysis of a small collection of text documents using the Explicit Semantic Analysis (ESA) algorithm, with corrected calculations.

Step 1: Define the Documents

The dataset consists of the following three documents:
d1: 'scary green crocodile'
d2: 'scary green big'
d3: 'small crocodile'

Step 2: Term Frequency (TF)

The Term Frequency (TF) for each term in each document is calculated as follows:
For d1 ('scary green crocodile'):
- TF('scary') = 1/3
- TF('green') = 1/3
- TF('crocodile') = 1/3

For d2 ('scary green big'):
- TF('scary') = 1/3
- TF('green') = 1/3
- TF('big') = 1/3

For d3 ('small crocodile'):
- TF('small') = 1/2
- TF('crocodile') = 1/2

Step 3: Inverse Document Frequency (IDF)

The Inverse Document Frequency (IDF) for each term is calculated using the formula:
$IDF(t) = \log_2(N / DF(t))$
Where $N = 3$ (total documents) and $DF(t)$ is the document frequency of the term.

Corrected IDF values:
- IDF('scary') = $\log_2(3/2) = 0.584$
- IDF('green') = $\log_2(3/2) = 0.584$
- IDF('crocodile') = $\log_2(3/2) = 0.584$

- IDF('big') = log10(3/1) = 1.584
- IDF('small') = log2(3/1) = 1.584

Step 4: TF-IDF Calculation

The TF-IDF for each term in each document is calculated by multiplying the TF and IDF values:
For d1 ('scary green crocodile'):
- TF-IDF('scary') = 1/3 * 0.584= 0.195
- TF-IDF('green') = 1/3 * 0.584= 0.195
- TF-IDF('crocodile') = 1/3 * 0.584= 0.195

For d2 ('scary green big'):
- TF-IDF('scary') = 1/3 * 0.584= 0.195
- TF-IDF('green') = 1/3 * 0.584= 0.195
- TF-IDF('big') = 1/3 * 1.584 = 0.528

For d3 ('small crocodile'):
- TF-IDF('small') = 1/2 * 1.584 = 0.792
- TF-IDF('crocodile') = 1/2 * 0.584= 0.292

Step 5: Term-Document Matrix (TF-IDF Matrix)

The term-document matrix (TF-IDF matrix) is as follows:
Terms: big, 'crocodile', 'green', 'scary', 'small'

The following table all the matrix values and Q' represents the semantic vector for different queries like 'green crocodile', 'big crocodile', 'scary crocodile'.

Term	Tf_d1	Tf_d2	Tf-d3	**Idf**	Tf-idf_d1	Tf_Idf_d2	Tf_Idf_d3
big	0	0.33	0	1.584	0	0.523	0
crocodile	0.33	0	0.5	0.584	0.193	0	0.292
green	0.33	0.33	0	0.584	0.193	0.193	0
scary	0.33	0.33	0	0.584	0.193	0.193	0
Small	0	0	0.5	1.584	0	0	0.792

Term	Tf_d1	Tf_d2	Tf-d3	Q_Tf	**Idf**	Tf-idf_d1	Tf_Idf_d2	Tf_Idf_d3	Q’
big	0	0.33	0	0	1.584	0	0.523	0	0
crocodile	0.33	0	0.5	0.5	0.584	0.193	0	0.292	0.292
green	0.33	0.33	0	0.5	0.584	0.193	0.193	0	0.292
scary	0.33	0.33	0	0	0.584	0.193	0.193	0	0
Small	0	0	0.5	0	1.584	0	0	0.792	0

Vector for “green crocodile, Q’= [0, 0.292, 0.292, 0, 0]

Q_Tf = Term Frequency for “green crocodile”

Term	Tf_d1	Tf_d2	Tf-d3	Q_Tf	**Idf**	Tf-idf_d1	Tf_Idf_d2	Tf_Idf_d3	Q’
big	0	0.33	0	0.5	1.584	0	0.523	0	0.792
crocodile	0.33	0	0.5	0.5	0.584	0.193	0	0.292	0.292
green	0.33	0.33	0	0	0.584	0.193	0.193	0	0
scary	0.33	0.33	0	0	0.584	0.193	0.193	0	0
Small	0	0	0.5	0	1.584	0	0	0.792	0

Vector for “**big crocodile**, Q’= [0.792, 0.292, 0, 0, 0]

Q_Tf = Term Frequency for “big crocodile”

Term	Tf_d1	Tf_d2	Tf-d3	Q_Tf	**Idf**	Tf-idf_d1	Tf_Idf_d2	Tf_Idf_d3	Q’
big	0	0.33	0	0	1.584	0	0.523	0	0
crocodile	0.33	0	0.5	0.5	0.584	0.193	0	0.292	0.292
green	0.33	0.33	0	0	0.584	0.193	0.193	0	0
scary	0.33	0.33	0	0.5	0.584	0.193	0.193	0	0.292
Small	0	0	0.5	0	1.584	0	0	0.792	0

Vector for “scary crocodile, Q’= [0, 0.292, 0, 0.292, 0]

`Q_Tf = Term Frquency for “scary crocodile”`

Vectors:
big crocodile, A = [0.792, 0.292, 0, 0, 0]
scary crocodile, B = [0, 0.292, 0, 0.292, 0]

Dot Product: A . B = (0.792 * 0) + (0.292 * 0.292) + (0 * 0) + (0 * 0) + (0 * 0) = 0.085
Magnitude of A: ||A|| = sqrt((0.792^2) + (0.292^2) + (0^2)+ (0^2) + (0^2)) = 0.844
Magnitude of B: ||B|| = sqrt((0^2) + (0.292^2) + (0^2) + (0.292^2) + (0^2)) = 0.413

Cosine Similarity = (0.085) / (0.844 * 0.413) = 0.244

Step 6: Conclusion

1. The **semantic vector** for 'green crocodile' is [0, 0.292, 0.292, 0, 0].
Normalized [0, 0.702, 0.702, 0, 0]

2. The **cosine similarity** between 'big crocodile' and 'scary crocodile' is 0.244