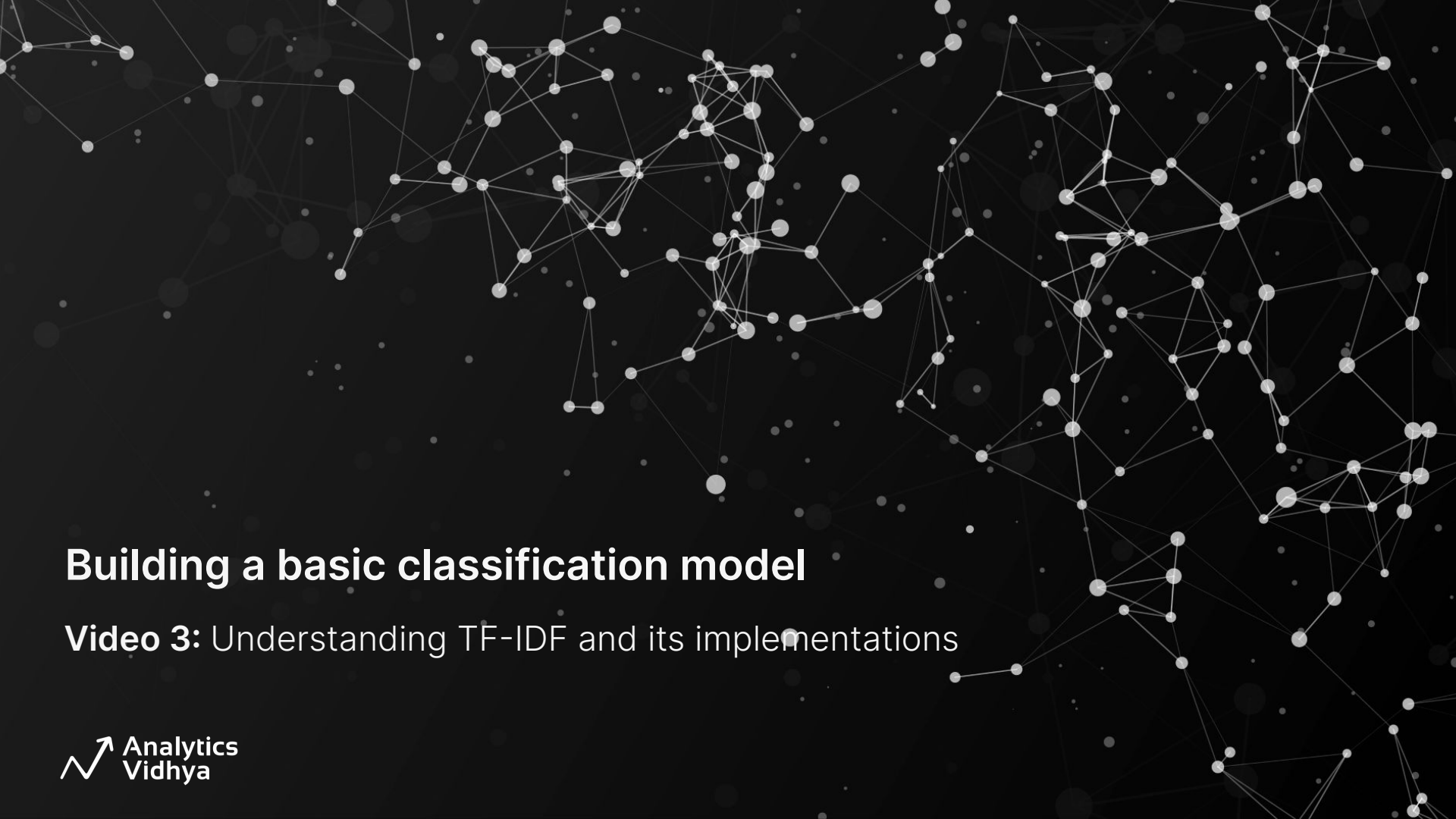

Building a basic classification model
Video 3: Understanding TF-IDF and its implementations
Analytics Vidhya

Vectorization

One Hot Encoding

Count Vectorizer

TF-IDF

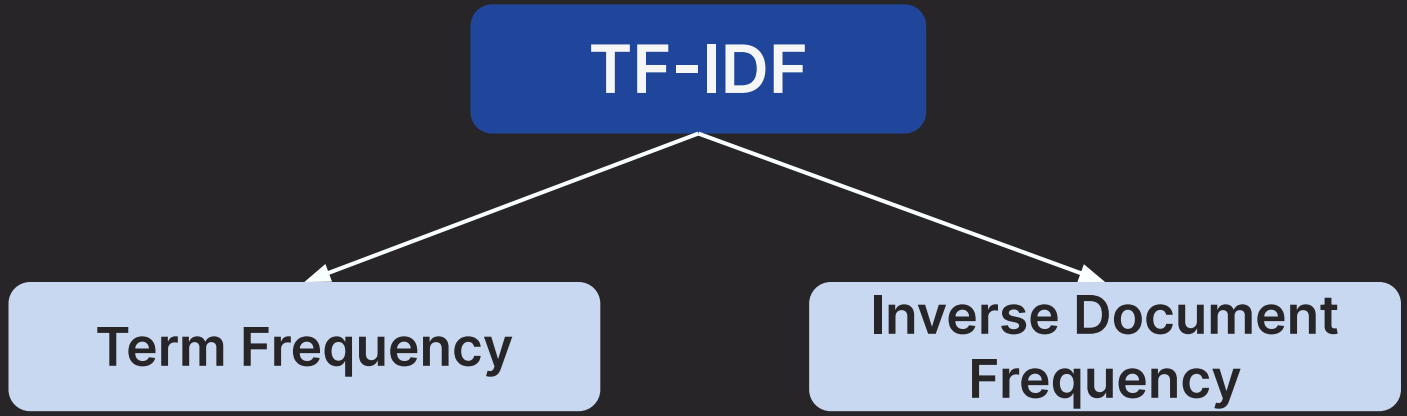
Word Embedding

Term Frequency - Inverse Document Frequency

Evaluate how relevant a word is to a document

Balance the term frequency with the uniqueness of the term.

Components of TF-IDF



Term Frequency (TF)

Measures the frequency of a word in a document

$$\text{TF} = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

Analytics Vidhya

Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

Analytics Vidhya

Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

life	4

Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

life	4
beautiful	1

Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

life	4
beautiful	1
Total	10

Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

life	4
beautiful	1
Total	10

$$TF = \frac{4}{10}$$

$$TF = \frac{1}{10}$$

Analytics Vidhya

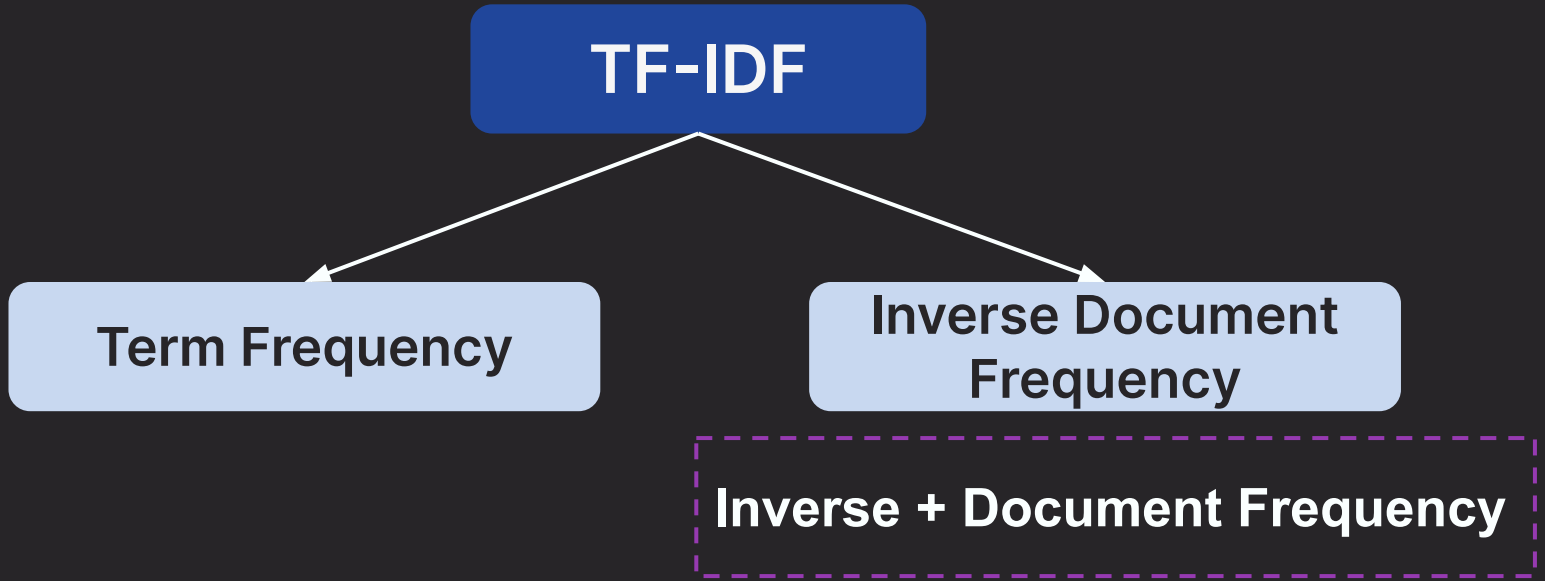
Term Frequency (TF)

Document 1 : “Life is life, life is beautiful and full of life.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

Components of TF-IDF



Document Frequency (DF)

Number of documents in which the word appears

$$DF = \frac{\text{Number of document with term t}}{\text{Total number of document D}}$$

Document Frequency (DF)

Document 1 : “Life is life, life is beautiful and full of life.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

life

$$DF = \frac{3}{3} = 1$$

beautiful

$$DF = \frac{1}{3}$$

Inverse Document Frequency (IDF)

Assess the importance of the term across the corpus.

$$\text{IDF} = \log \frac{\text{Total number of document D}}{\text{Number of document with term t}}$$

IDF

Document 1 : “Life is life, life is beautiful and full of life.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

life

$$\text{IDF} = \log \frac{\text{Total number of document D}}{\text{Number of document with term t}} = \log \frac{3}{3} = 0$$

IDF

Document 1 : “Life is life, life is beautiful and full of life.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

beautiful

$$\text{IDF} = \log \frac{\text{Total number of document D}}{\text{Number of document with term t}} = \log \frac{3}{1} = 0.47$$

Term Frequency - Inverse Document Frequency

TF = TF X IDF

life

$$TF = \frac{4}{10}$$

X

$$IDF = 0$$

TF-IDF = 0

Term Frequency - Inverse Document Frequency

TF = TF X IDF

life

$$TF = \frac{4}{10} \quad X \quad IDF = 0$$

TF-IDF = 0

beautiful

$$TF = \frac{1}{10} \quad X \quad IDF = 1.099$$

TF-IDF = 0.047

Term Frequency - Inverse Document Frequency

TF =TF X IDF

- TF-IDF increases with the number of occurrences of a word in a specific document but offset by frequency of word in corpus.

Analytics Vidhya

Term Frequency - Inverse Document Frequency

	life	beautiful	surprising	struggle	turn	teaches	resilience	growth
Documents	0	0.1099	-	-	-	-	-	-

Analytics Vidhya

Importance of TF-IDF

- Information retrieval
- Text Mining
- Search Engine Optimization
- Document Summarization
- Topic Modeling

Analytics Vidhya

Jupyter Notebook