Vector space model

Term Frequency - Inverse Document Frequency (tf-idf)

- Product of Term Frequency (tf) and Inverse Document Frequency (idf)

$$W_{t,d} = [1 + \log_{10}(tf_{t,d})] \; X \; [\log_{10}(N/df_t)]$$

Analytics Vidhya

Term Frequency - Inverse Document Frequency (tf-idf)

Query (q): book of Analytics

Doc 1 (d1): This book is of Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	d1(tf-idf)	d2(tf-idf)
Analytics	0	0
Big-Data	0	0.3
book	0.3	0
data	0	0.3
examining	0	0.3
is	0	0
large	0	0.3
of	0	0
process	0	0.3
the	0	0.3
this	0.3	0
volume	0	0.3

Term Frequency - Inverse Document Frequency (tf-idf)

Query (q): book of Analytics

Doc 1 (d1): This book is of Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	d1(tf-idf)	d2(tf-idf)
Analytics	0	0
Big-Data	0	0.3
book	0.3	0
data	0	0.3
examining	0	0.3
is	0	0
large	0	0.3
of	0	0
process	0	0.3
the	0	0.3
this	0.3	0
volume	0	0.3

Vector space model

- Each document is represented as a vector

Term	d1(tf-idf)	d2(tf-idf)
Analytics	0	0
Big-Data	0	0.3
book	0.3	0
data	0	0.3
examining	0	0.3
is	0	0
large	0	0.3
of	0	0
process	0	0.3
the	0	0.3
this	0.3	0
volume	0	0.3

Vector space model

- Each document is represented as a vector
- Dimension of vector space: |V|

Term	d1(tf-idf)	d2(tf-idf)
Analytics	0	0
Big-Data	0	0.3
book	0.3	0
data	0	0.3
examining	0	0.3
is	0	0
large	0	0.3
of	0	0
process	0	0.3
the	0	0.3
this	0.3	0
volume	0	0.3

Vector space model

- Each document is represented as a vector
- Dimension of vector space: |V|
- Terms are the axes

Term	d1(tf-idf)	d2(tf-idf)
Analytics	0	0
Big-Data	0	0.3
book	0.3	0
data	0	0.3
examining	0	0.3
is	0	0
large	0	0.3
of	0	0
process	0	0.3
the	0	0.3
this	0.3	0
volume	0	0.3

Vector space model for ranked retrieval

- Each document is represented as a vector
- Dimension of vector space: |V|
- Terms are the axes
- Convert the queries into vectors

Vector space model for ranked retrieval

- Each document is represented as a vector
- Dimension of vector space: |V|
- Terms are the axes
- Convert the queries into vectors
- Rank the documents based on their proximity to the query in vector space

Vector space model for ranked retrieval

- Each document is represented as a vector
- Dimension of vector space: |V|
- Terms are the axes
- Convert the queries into vectors
- Rank the documents based on their proximity to the query in vector space

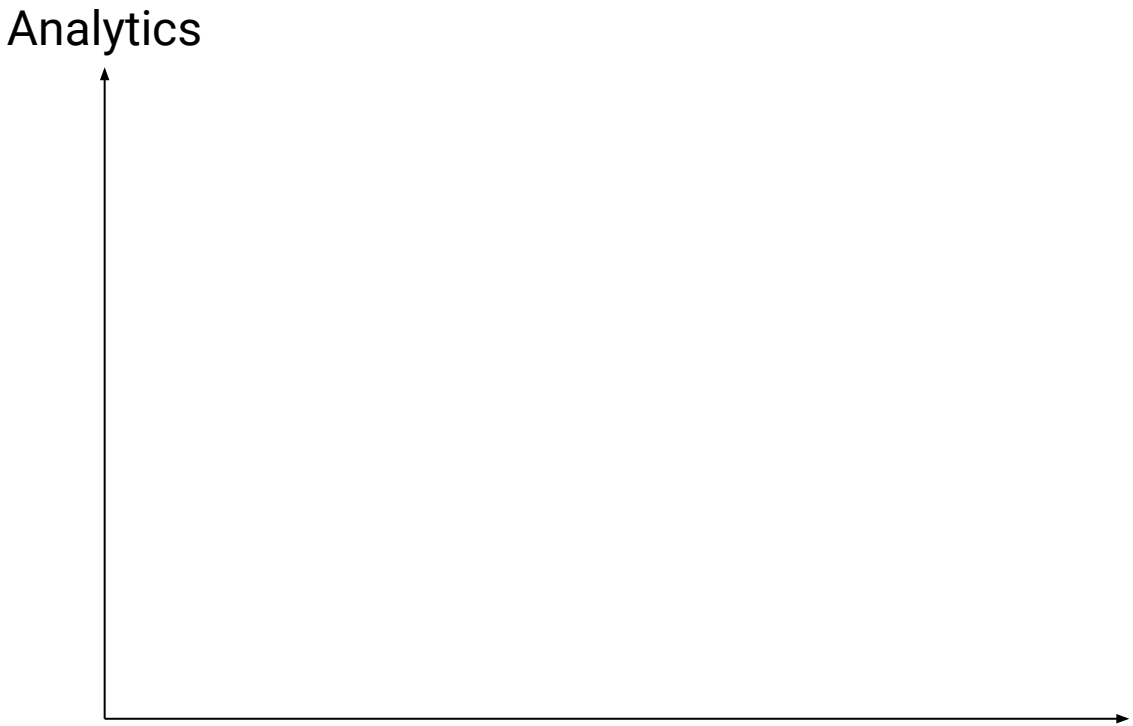
Proximity = Similarity

Vector space model for ranked retrieval

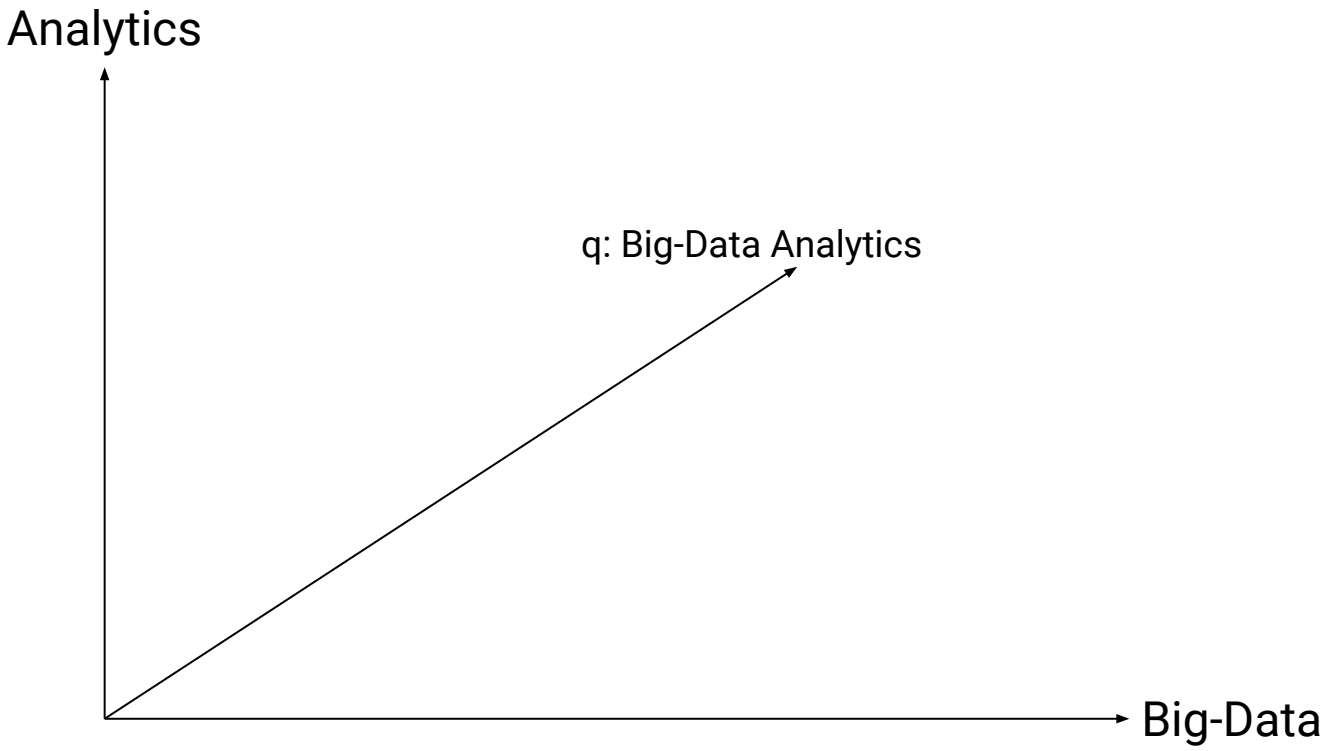
- Each document is represented as a vector
- Dimension of vector space: |V|
- Terms are the axes
- Convert the queries into vectors
- Rank the documents based on their proximity to the query in vector space

Proximity = Similarity = Inverse of the distance

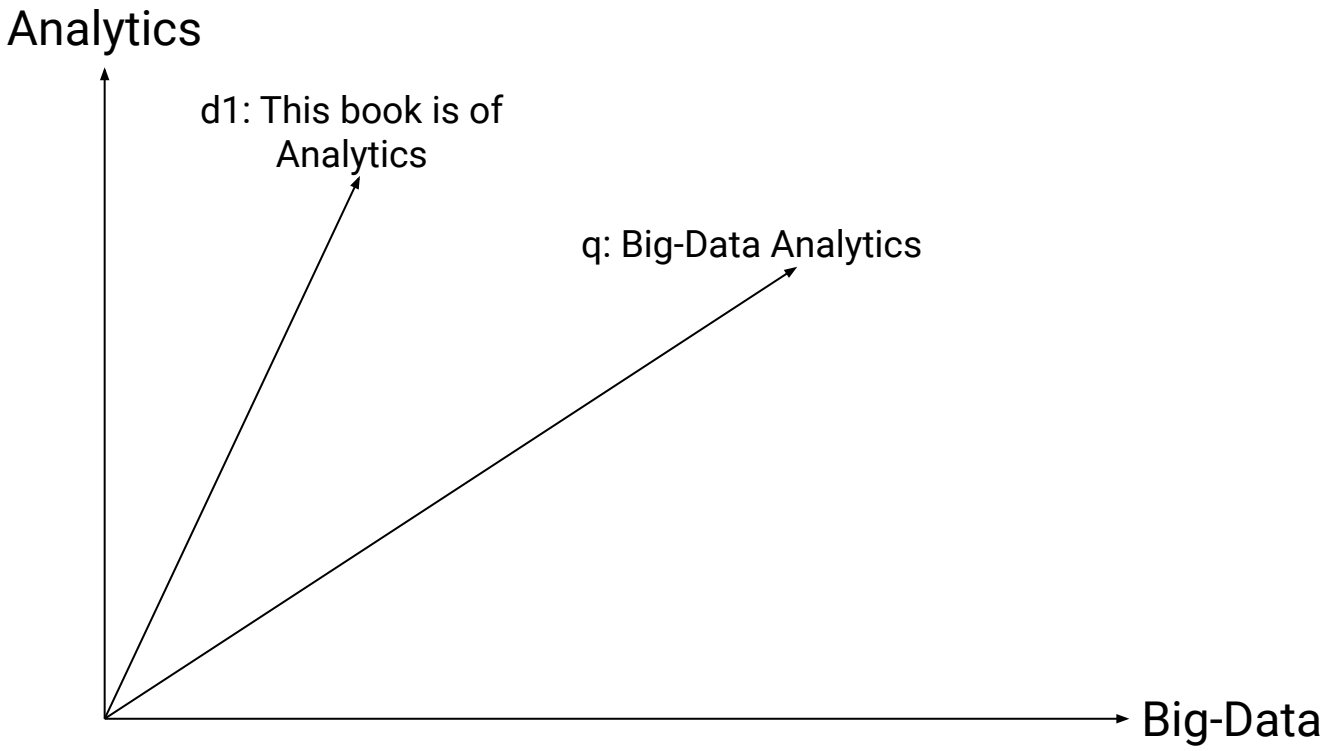
Vector space model for ranked retrieval



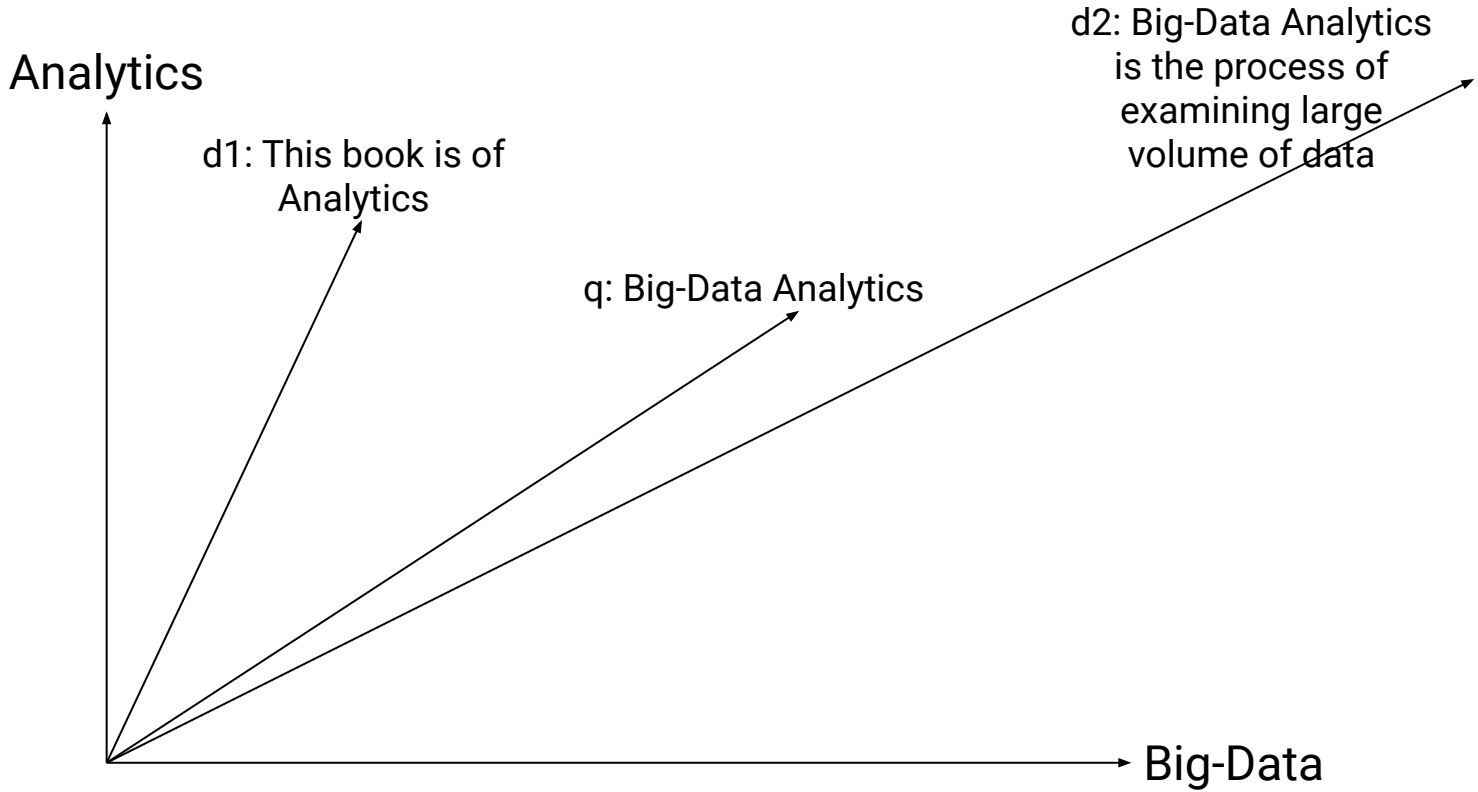
Vector space model for ranked retrieval



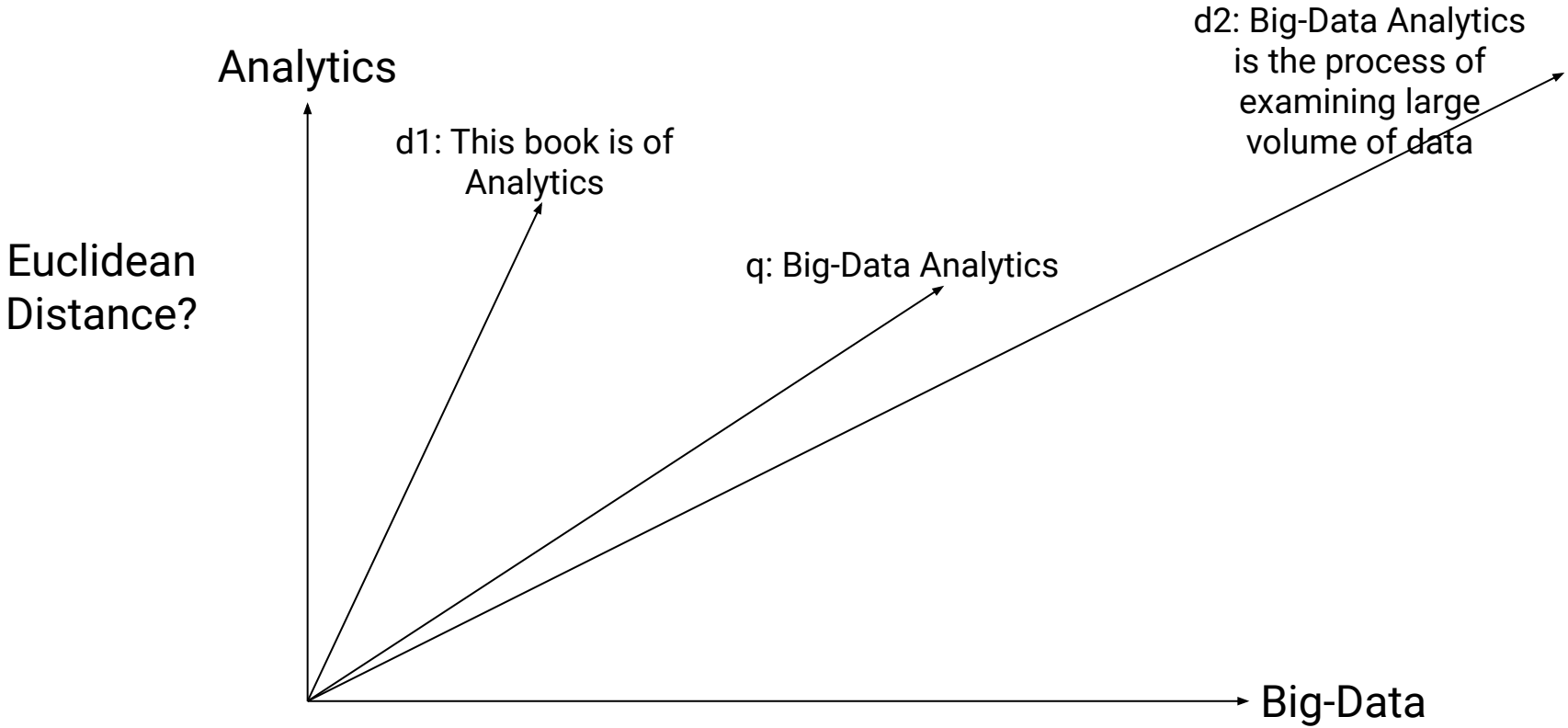
Vector space model for ranked retrieval



Vector space model for ranked retrieval

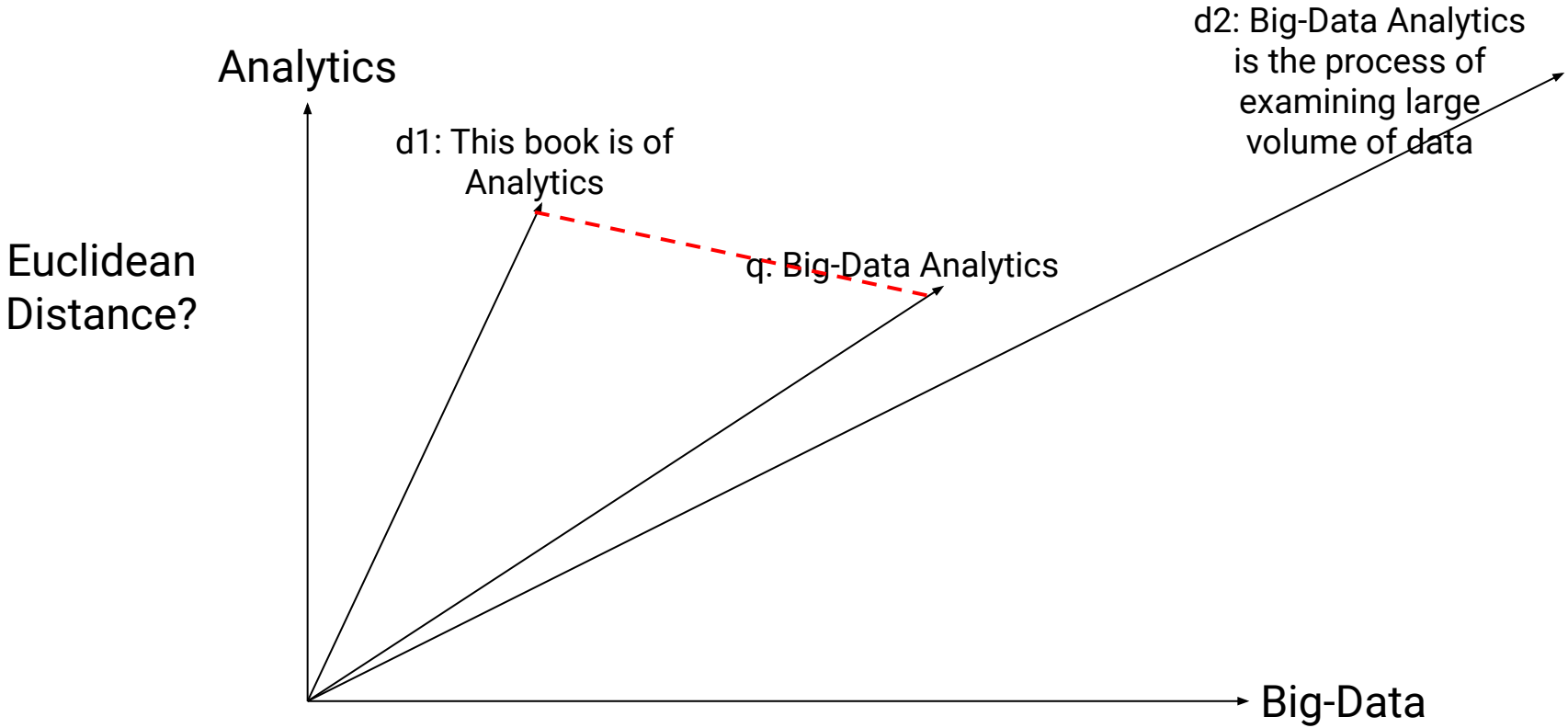


Vector space model for ranked retrieval



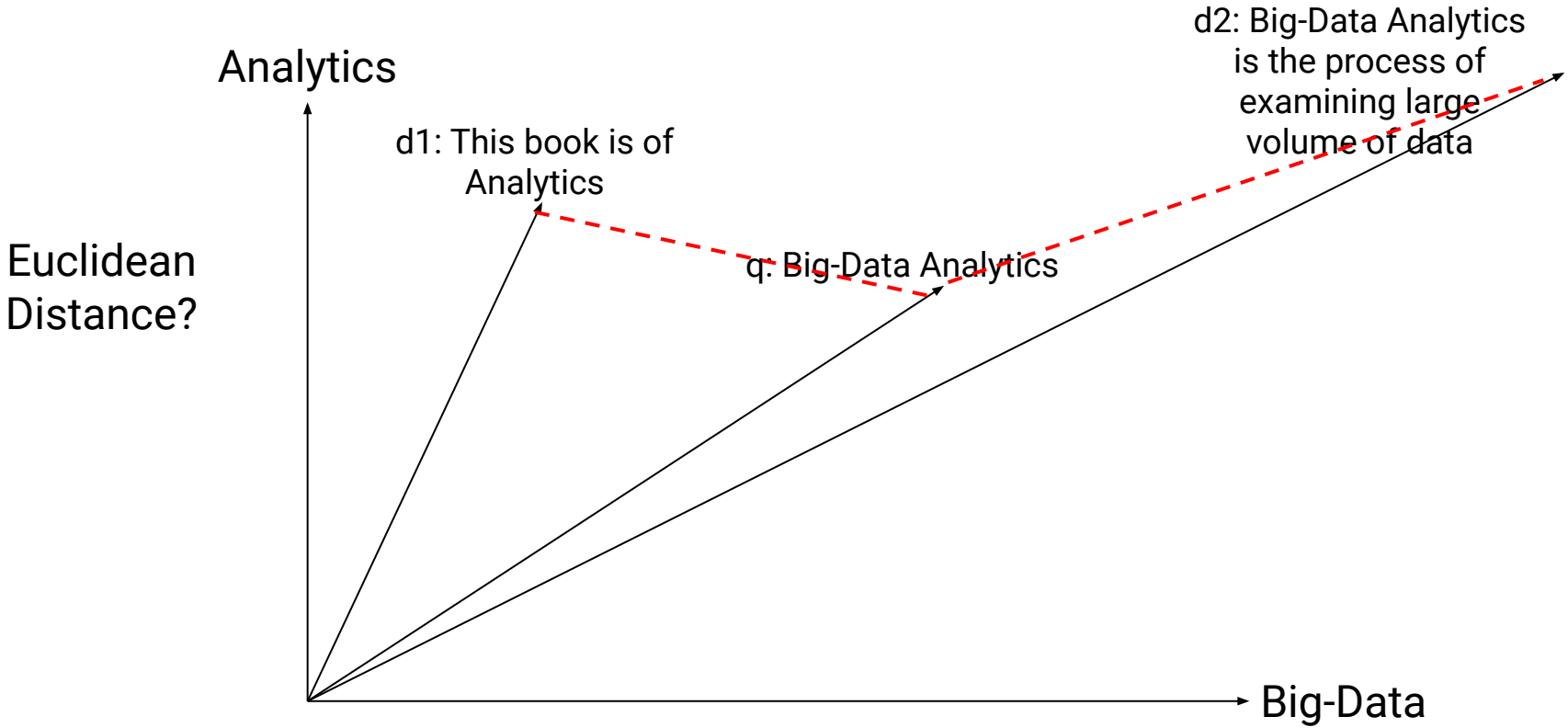
Analytics Vidhya

Vector space model for ranked retrieval



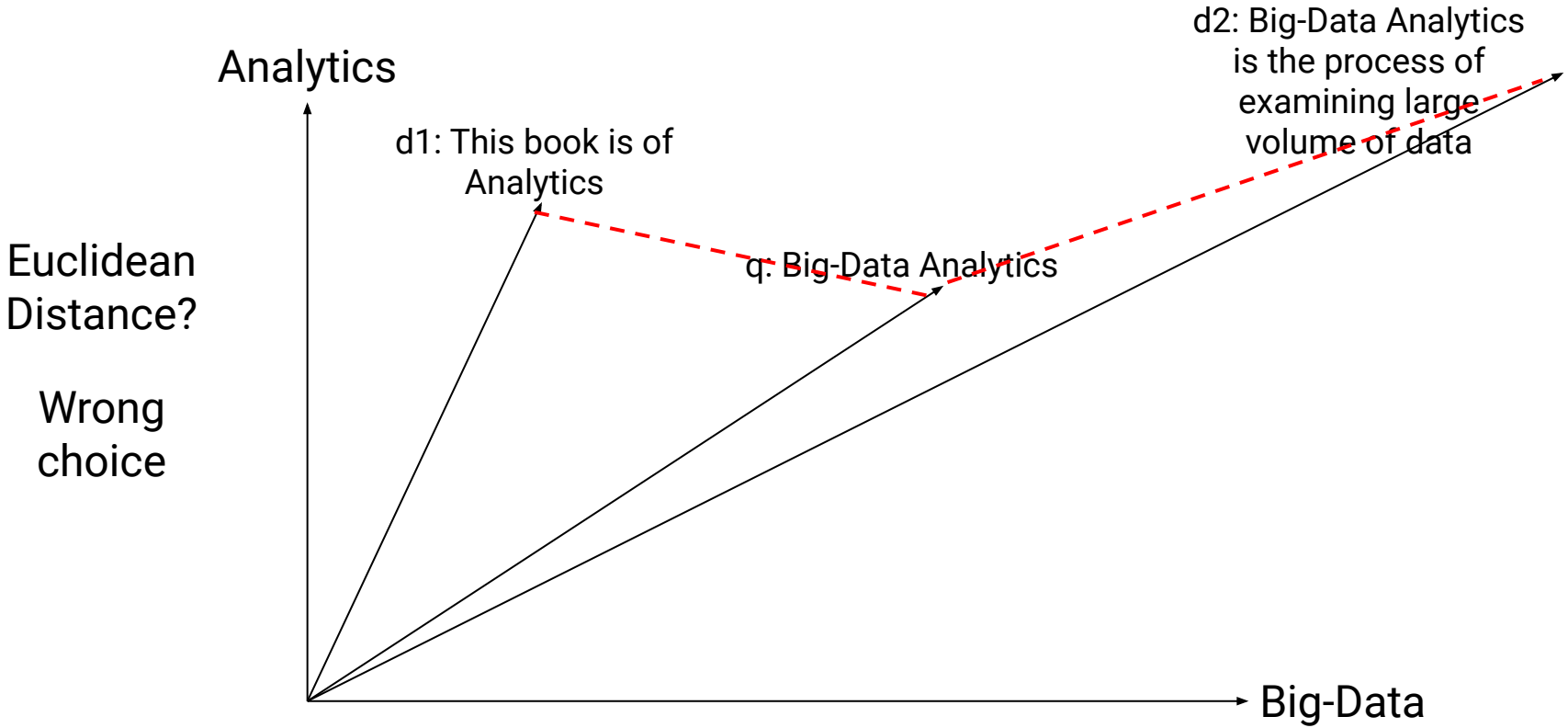
Analytics Vidhya

Vector space model for ranked retrieval



Analytics Vidhya

Vector space model for ranked retrieval



Analytics Vidhya

Vector space model for ranked retrieval

Rank based
on the angle

Vector space model for ranked retrieval

Rank based on the angle

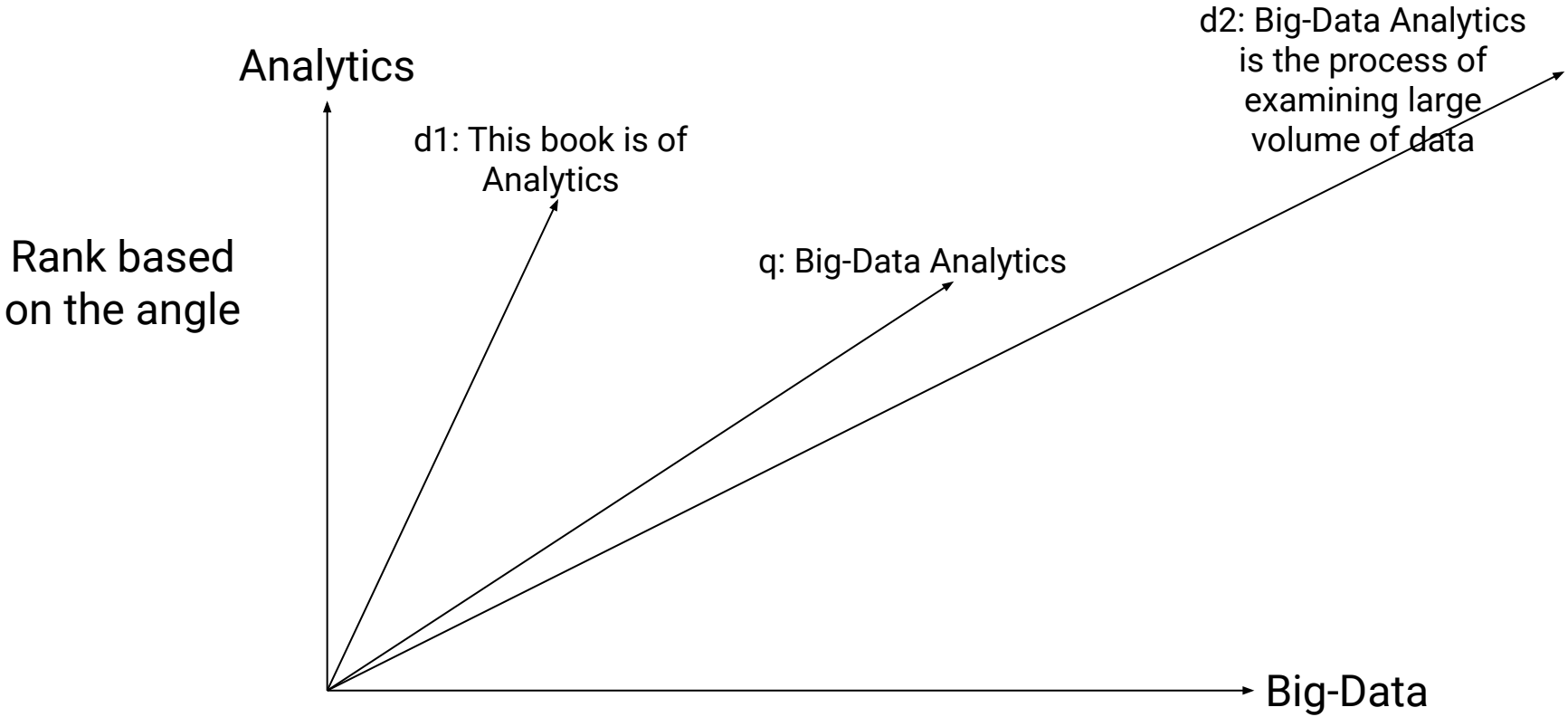
Vector space model for ranked retrieval

Rank based on the angle

Vector space model for ranked retrieval

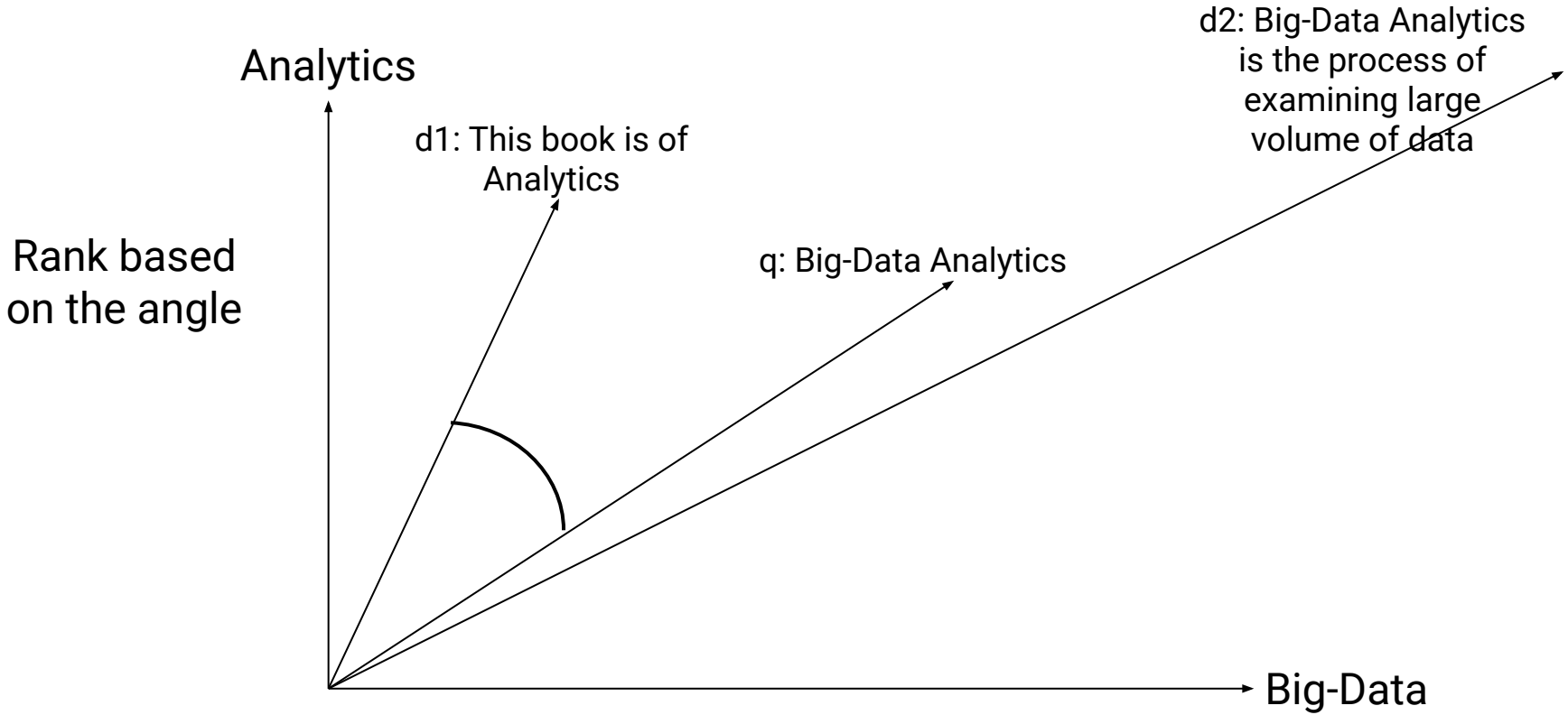
Rank based on the angle

Vector space model for ranked retrieval



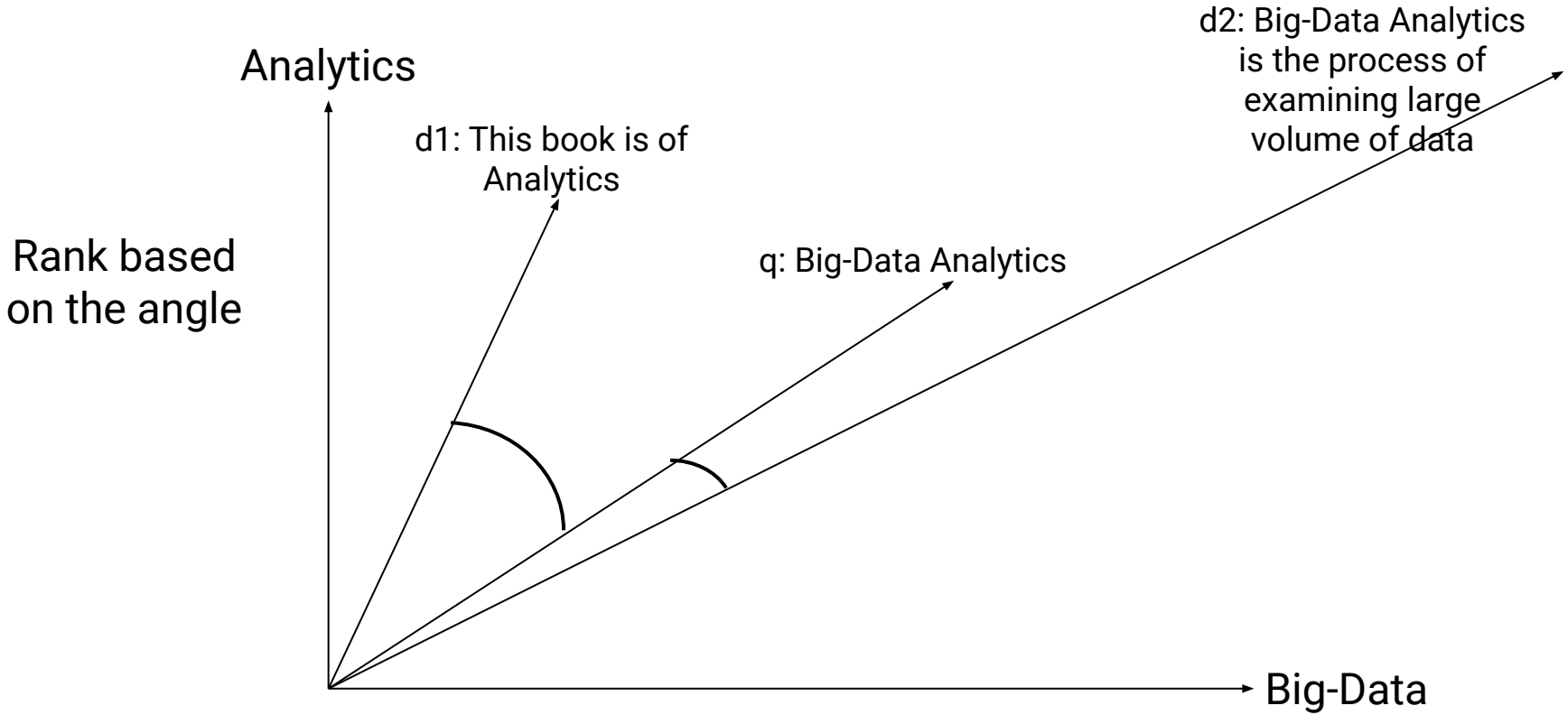
Analytics Vidhya

Vector space model for ranked retrieval



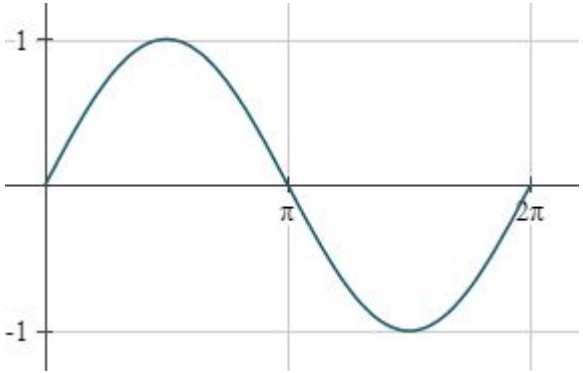
Analytics Vidhya

Vector space model for ranked retrieval

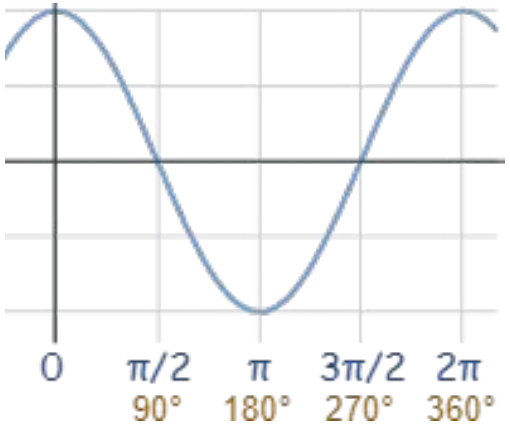


Analytics Vidhya

Vector space model for ranked retrieval



Sine



Cosine

Vector space model for ranked retrieval

Cosine Similarity

Vector space model for ranked retrieval

Cosine Similarity

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

Analytics Vidhya

Vector space model for ranked retrieval

Cosine Similarity

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

Vector space model for ranked retrieval

Cosine Similarity for vector space model

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|}$$

Vector space model for ranked retrieval

Cosine Similarity for vector space model

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Analytics Vidhya

Vector space model for ranked retrieval

Cosine Similarity for vector space model

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i is the weight of ith term in the query

Analytics Vidhya

Vector space model for ranked retrieval

Cosine Similarity for vector space model

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i is the weight of ith term in the query
- d_i is the weight of ith term in the document

Analytics Vidhya

Vector space model for ranked retrieval

Cosine Similarity for vector space model

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i is the weight of ith term in the query
- d_i is the weight of ith term in the document
- |V| is the vocabulary size

Analytics Vidhya

Vector space model for ranked retrieval

- Represent query and each document as a weighted vector

Vector space model for ranked retrieval

- Represent query and each document as a weighted vector
- Calculate the cosine similarity for query and each document

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Analytics Vidhya

Vector space model for ranked retrieval

- Represent query and each document as a weighted vector
- Calculate the cosine similarity for query and each document
- Rank documents with respect to query (higher the cosine similarity score, lesser the angle and more the similarity)

Vector space model for ranked retrieval

- Represent query and each document as a weighted vector
- Calculate the cosine similarity for query and each document
- Rank documents with respect to query (higher the cosine similarity score, lesser the angle and more the similarity)
- Return top K documents

Thank You

Analytics Vidhya