The Vector Space Model

Debapriyo Majumdar
Indian Statistical Institute
debapriyo@isical.ac.in

Basics of Ranking

- Boolean retrieval models simply return documents satisfying the Boolean condition(s)
  - Among those, are all documents equally “good”?
  - No
- Case: single term query
  - Not all documents containing the term are equally associated with that term
- From Boolean model to term weighting
  - Weight of a term in a document is 1 or 0 in Boolean model
  - Use more granular term weighting
- Weight of a term in a document: represent how much the term is important in the document and vice versa

Term weighting – TF.iDF

- How important is a term t in a document d
- Intuition 1: More times a term is present in a document, more important it is
  - Term frequency (TF)
- Intuition 2: If a term is present in many documents, it is less important particularly to any one of them
  - Document frequency (DF)
- Combining the two: TF.iDF (term frequency × inverse document frequency)
  - Many variants exist for both TF and DF

Formulation of term frequency (TF)

1. Simplest term frequency: Number of times a term t occurs in a document d: $\mathrm{freq}(t, d)$
   - If a term a is present 10 times, b is present 2 times, is a 5 times more important than b in that document?
   - No, importance does not grow linearly with frequency
2. Logarithmically scaled frequency: $1 + \log(\mathrm{freq}(t, d))$, or even $1 + \log(1 + \log(\mathrm{freq}(t, d)))$
   - Still, long documents on same topic would have the more frequency for the same terms
3. Augmented frequency: avoid bias towards longer documents

Half the score for just being present

Rest is a function of frequency

$$TF(t,d) = 0.5 + \frac{0.5 \times \mathrm{freq}(t,d)}{\max\{\mathrm{freq}(w,d) \mid w \in d\}}$$

for all t in d; 0 otherwise

(Inverse) document frequency (iDF)

- Inverse document frequency of a term t

$$\mathrm{iDF}(t) = \log\left[\frac{N}{\mathrm{DF(t)}}\right],$$

where N = total number of documents

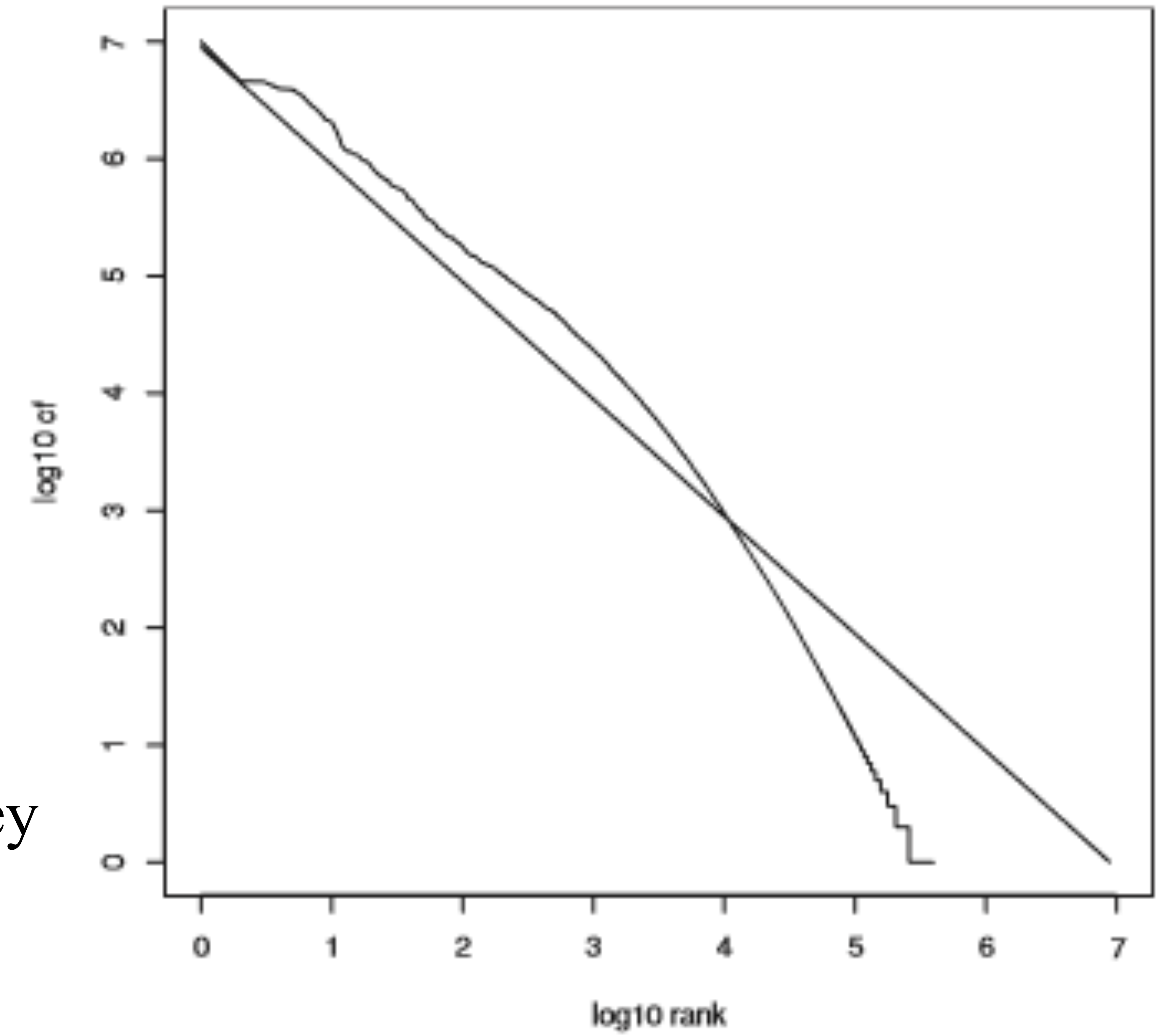
$\mathrm{DF}(t)$ = number of documents in which t occurs

<u>Distribution of terms</u>

- Zipf's law: Let $T = \{t_1, \ldots, t_m\}$ be the terms, sorted by decreasing order of the number of documents in which they occur. Then

$$\mathrm{DF}(t_i) \propto \frac{1}{i}$$

- In other words, $\log \mathrm{DF}(t_i) = \log c + (-1) \cdot \log i$ for some constant c



Zipf's law fitting for Reuter's RCV1 collection

Vector space model

	d_1	d_2	d_3	d_4	d_5	q
diwali	0.5	0	0	0	0	
india	0.2	0.2	0	0.2	0.1	1
flying	0	0.4	0	0	0	
population	0	0	0	0.5	0	
autumn	0	0	1	0	0	
statistical	0	0.3	0	0	0.2	1

Each term represents a dimension

Documents are vectors in the term-space

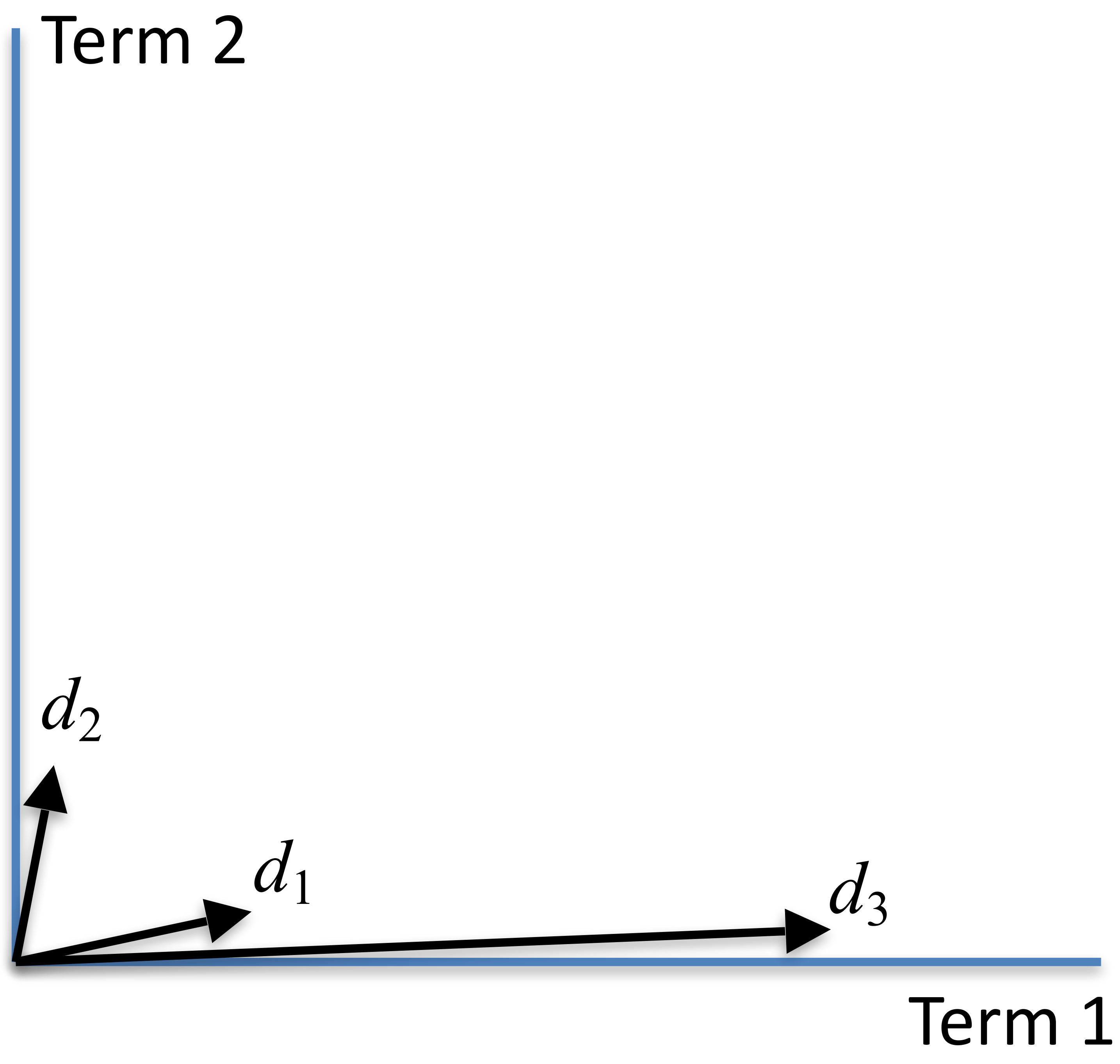
Term-document matrix: a very sparse matrix

Entries are scores of the terms in the documents (Boolean → Count → Weight)

Query is also a vector in the term-space

- Vector similarity: inverse of “distance”
- Euclidean distance?

Problem with Euclidean distance



Problem

- Topic-wise d_3 is closer to d_1 than d_2 is
- Euclidean distance wise d_2 is closer to d_1

Dot product seems to solves this problem

Vector space model

	d_1	d_2	d_3	d_4	d_5	q
diwali	0.5	0	0	0	0	
india	0.2	0.2	0	0.2	0.1	1
flying	0	0.4	0	0	0	
population	0	0	0	0.5	0	
autumn	0	0	1	0	0	
statistical	0	0.3	0	0	0.2	1
$q^T d$	0.2	0.5	0	0.2	0.3	

Vector similarity: dot product

$$sim(q,d) = q^T d$$

Each term represents a dimension

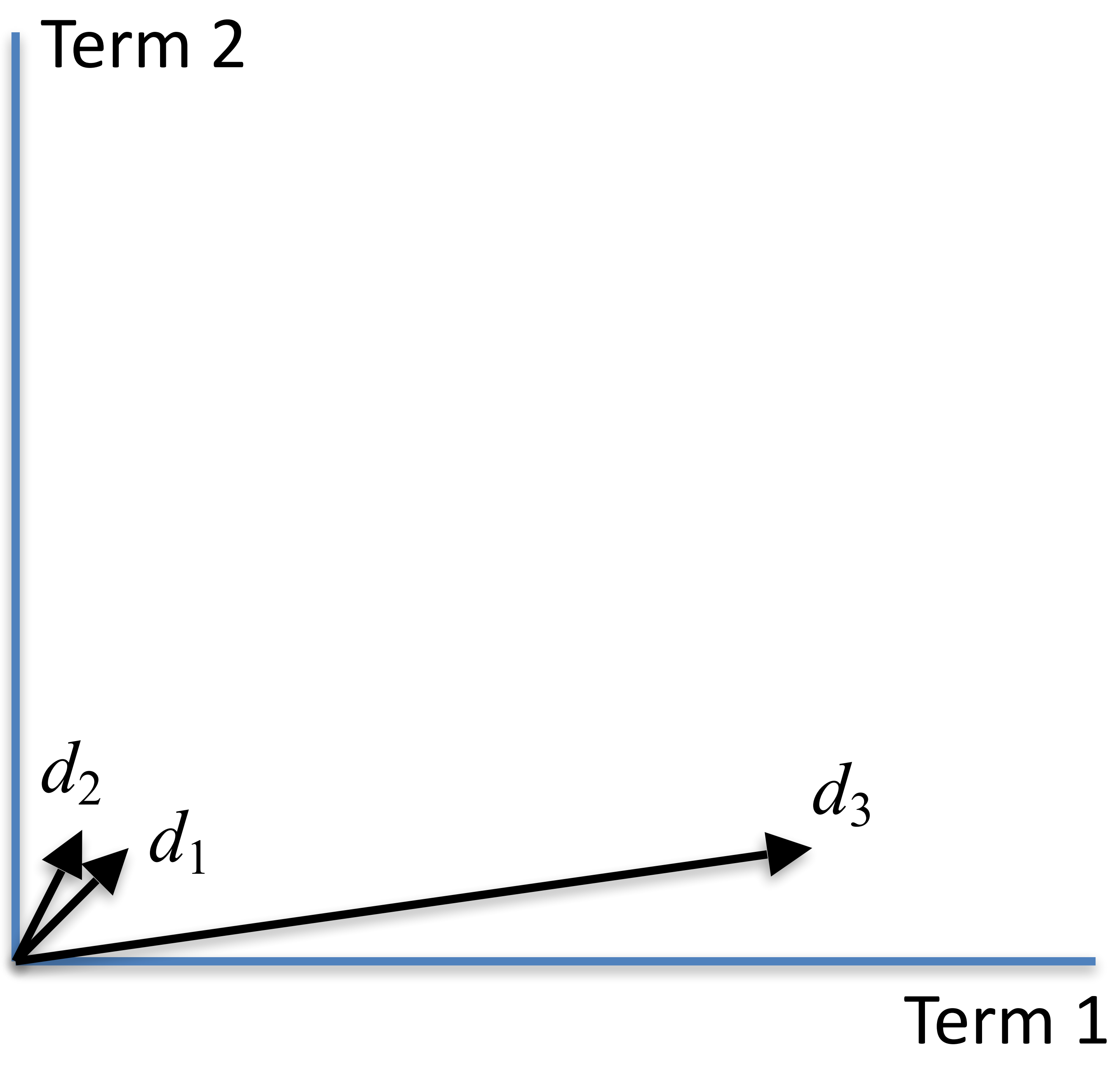
Documents are vectors in the term-space

Term-document matrix: a very sparse matrix

Entries are scores of the terms in the documents (Boolean → Count → Weight)

Query is also a vector in the term-space

Problem with dot product



Problem

- Topic-wise d_2 is closer to d_1 than d_3 is
- Dot product of d_3 and d_1 is greater because of the length of d_3

- Consider angle
  - Cosine of the angle between two vectors
  - Same direction: 1 (similar)
  - Orthogonal: 0 (unrelated)
  - Opposite direction: -1 (opposite)

Vector space model

	d_1	d_2	d_3	d_4	d_5	q
diwali	0.5	0	0	0	0	
india	0.2	0.2	0	0.2	0.1	1
flying	0	0.4	0	0	0	
population	0	0	0	0.5	0	
autumn	0	0	1	0	0	
statistical	0	0.3	0	0	0.2	1

Each term represents a dimension

Documents are vectors in the term-space

Term-document matrix: a very sparse matrix

Entries are scores of the terms in the documents (Boolean → Count → Weight)

Query is also a vector in the term-space

cosine of the angle between the vectors

$$\mathrm{sim}_{\cos}(q, d) = \cos(\theta_{q,d}) = \frac{q^T d}{\|q\| \|d\|}$$

References

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. "Introduction to Information Retrieval", Cambridge University Press. 2008.