Applications

Jordan Boyd-Graber

University of Maryland

Information Retrieval

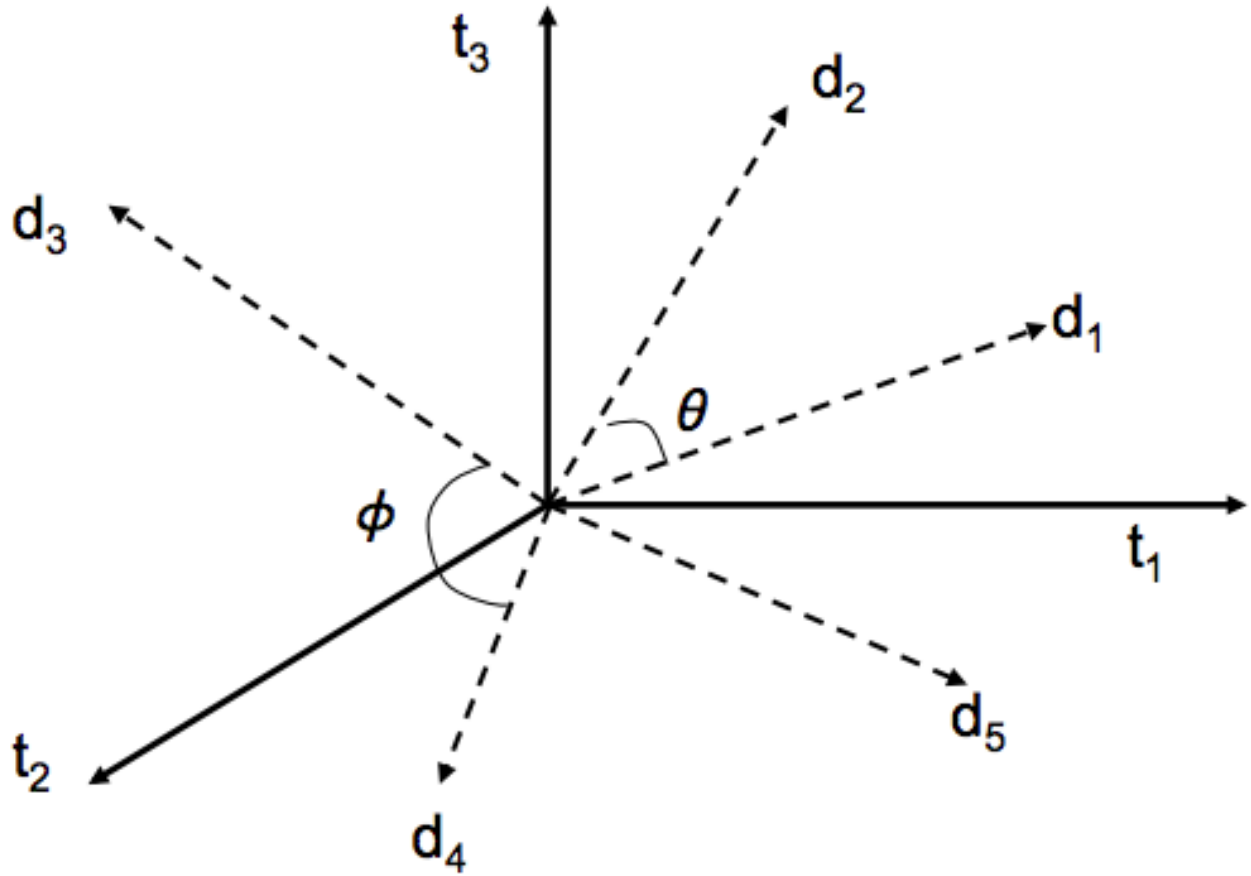
Slides adapted from Jimmy Lin

IR is worth a lot of money . . .

Prerequisites

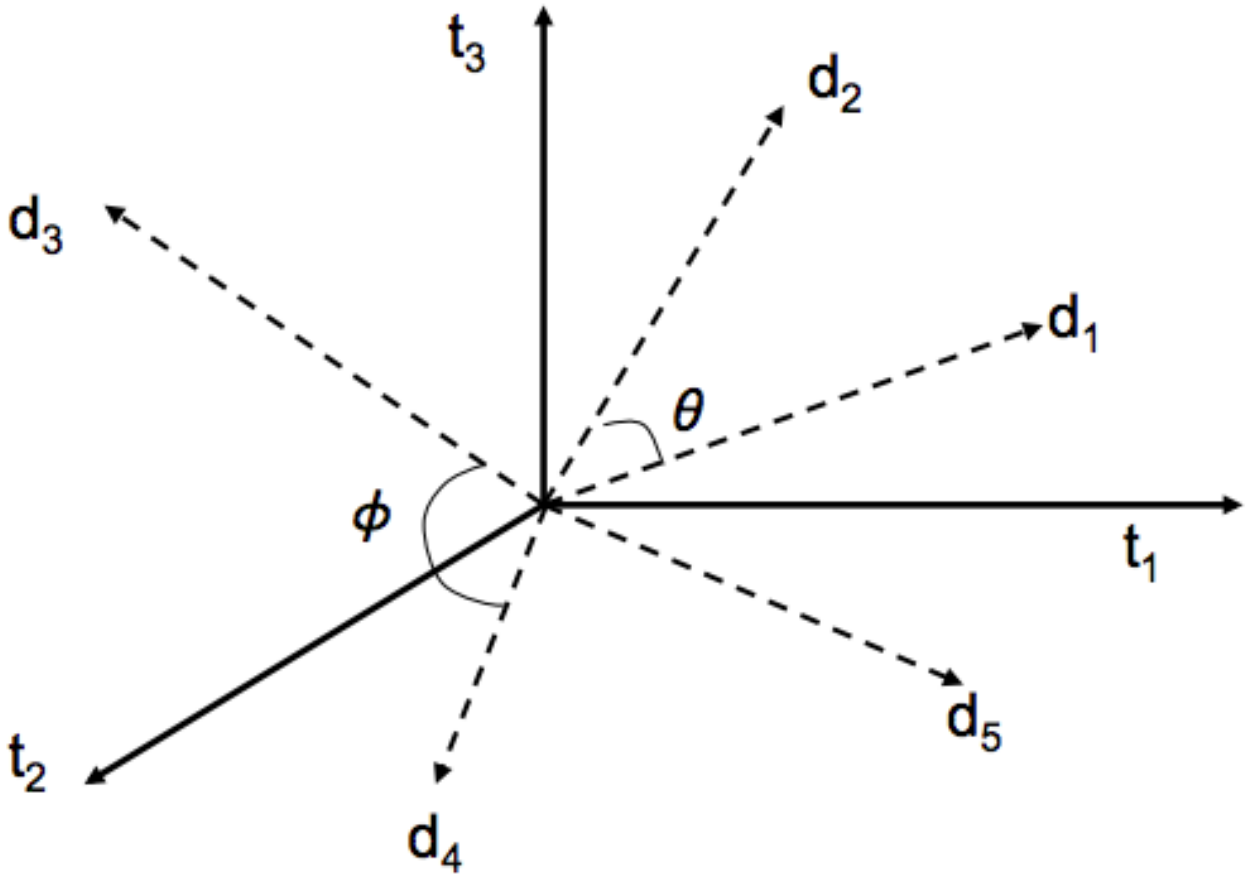
- Search a “collection” of **documents**
- Each document contains **terms** (words)
- Users create queries

Representing documents



Each document is vector $d_i = \langle w_{i,1}, \ldots w_{i,V} \rangle$ (each word is dimension)

Representing documents



Each document is vector $d_i = \langle w_{i,1}, \ldots w_{i,V} \rangle$ (each word is dimension)
Most of these are zero!

Intuitions

- Term weights consist of two components
  - Local: how important is the term in this document?
  - Global: how important is the term in the collection?
- Here's the intuition:
  - Terms that appear often in a document should get high weights
  - Terms that appear in many documents should get low weights
- How do we capture this mathematically?
  - Term frequency (local)
  - Inverse document frequency (global)

tf-idf Term Weighting

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \tag{1}$$

- Word i's weight in document j
- Frequency of word i in document j
- Help with interpretation

tf-idf Term Weighting

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \tag{1}$$

- Word i's weight in document j
- Frequency of word i in document j
- Help with interpretation
- Tension: prediction or interpretation

tf-idf Term Weighting

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \tag{1}$$

- Word i's weight in document j
- Frequency of word i in document j
- Help with interpretation
- Tension: prediction or interpretation