

# INFO/CS 4300: Language and Information

## The Inverted Index Algorithm for Cosine Similarity

Name: \_\_\_\_\_

NetID: \_\_\_\_\_ Left colleague NetID: \_\_\_\_\_ Right colleague NetID: \_\_\_\_\_

### Algorithm sketch

**Iterate** through all query terms  $q_i$ :

for each term, **iterate** through its postings  $d_j$ :

**update** the respective **accumulators** by  $w_{iq} * w_{ij}$

**return** sorted documents by their final **accumulator** scores

(after normalizing each accumulator by the document norm)

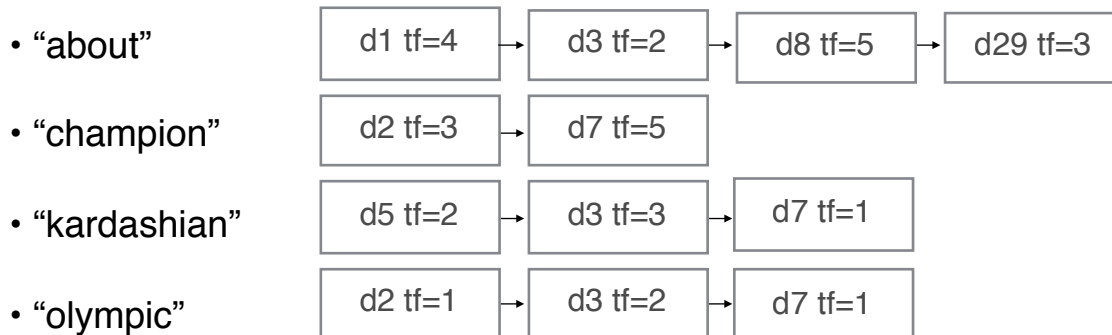
Instantiation:  $IDF_i = N / DF_i$ ,  $N = 300$ ,  $DF_i$  = number of docs that contain  $q_i$

$$w_{ij} = tf_{ij} * IDF_i$$

$w_{iq} = 1$  for all terms in the query,  $w_{iq} = 0$  otherwise

For this example, do not normalize by the document norm

### Inverted index



### IDF values

Key	Value
-----	-------

### Accumulators

Key	Value
-----	-------

“about”

“champion”

“kardashian”

“olympic”

### Final ranking