Introduction to

Information Retrieval

Scoring with the Jaccard coefficient

Take 1: Jaccard coefficient

- A commonly used measure of overlap of two sets *A* and *B* is the Jaccard coefficient
- $\text{jaccard}(A,B) = |A \cap B| / |A \cup B|$
- $\text{jaccard}(A,A) = 1$
- $\text{jaccard}(A,B) = 0$ if $A \cap B = 0$
- *A* and *B* don’t have to be the same size.
- Always assigns a number between 0 and 1.

Jaccard coefficient: Scoring example

- What is the query-document match score that the Jaccard coefficient computes for each of the two documents below?
- Query: *ides of march*
- Document 1: *caesar died in march*
- Document 2: *the long march*

Issues with Jaccard for scoring

- It doesn't consider *term frequency* (how many times a term occurs in a document)
  - Rare terms in a collection are more informative than frequent terms
  - Jaccard doesn't consider this information
- We need a more sophisticated way of normalizing for length
  - Later in this lecture, we'll use $|A \cap B| / \sqrt{|A \cup B|}$

    . . . instead of |A ∩ B|/|A ∪ B| (Jaccard) for length normalization.

Introduction to

Information Retrieval

Scoring with the Jaccard coefficient