Scoring with the Jaccard Coefficient

Take 1 : Jaccard Coefficient

: 집합 A, B가 얼마나 비슷한지 나타내는 계수

- A commonly used measure of **overlap of two sets A and B** is the Jaccard Coefficient
  - Jaccard(A,B) = $|A \cap B|$ / $|A \cup B|$
  - Jaccard(A,A) = 1
  - 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

- Query : ides of march
- Document 1: caesar died in march
- Document 2 : the long march
- Jaccard(q,d1) = $\frac{1}{6}$, Jaccard(q,d2) = $\frac{1}{5}$
- 즉, 문장의 길이가 짧은 Document 2가 Document 1보다 높은 점수를 받음 → **문장의 길이**의 영향을 받음

Issues with Jaccard for Scoring

- It doesn't consider **term frequency**
  - **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
  - use $|A \cap B|$ / $|\sqrt{A \cup B}|$, instead of $|A \cap B|$ / $|A \cup B|$ for length normalization

출처 : stanford IR 강의 (https://www.youtube.com/watch?v=MiX8_JVP6PE&list=PLaZQkZp6WhWwoDuD6pQCmgVyDbUWl_ZUi&index=8)