Jaccard Similarity

Given a string, we can break it up into a set of words (i.e., *tokenize* it). E.g., dogs chase cats $\rightsquigarrow$ [dogs, chase, cats].

Jaccard Similarity

Given a string, we can break it up into a set of words (i.e., *tokenize* it). E.g., dogs chase cats ⇝ [dogs, chase, cats].

Given two sets of words/tokens A and B, the Jaccard similarity of A and B is defined by

$$\mathrm{Jacc}(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Jaccard Similarity

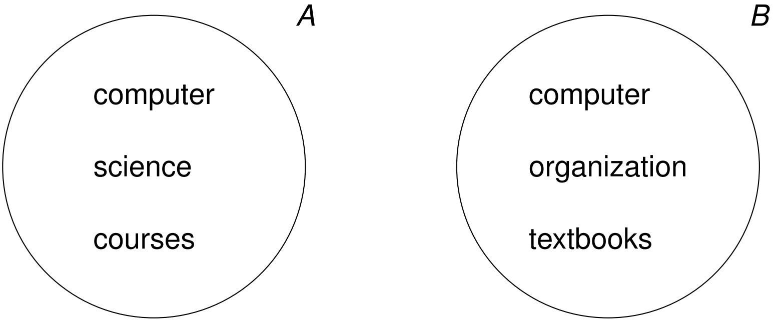
Given a string, we can break it up into a set of words (i.e., *tokenize* it). E.g., dogs chase cats $\rightsquigarrow$ [dogs, chase, cats].

Given two sets of words/tokens A and B, the Jaccard similarity of A and B is defined by

$$\text{Jacc}(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Example

Consider the strings *computer science courses* and *computer organization computer textbooks*.



A

computer

science

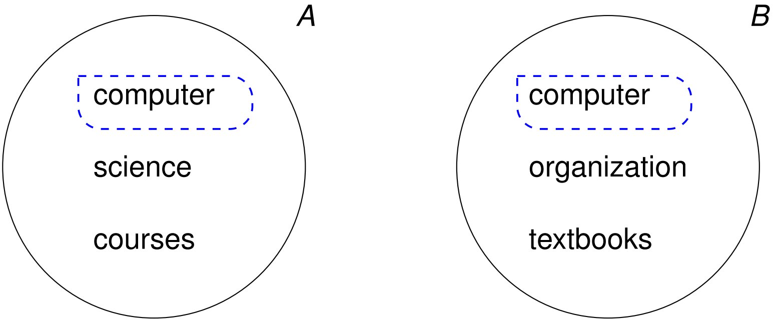
courses

B

computer

organization

textbooks



A

computer

science

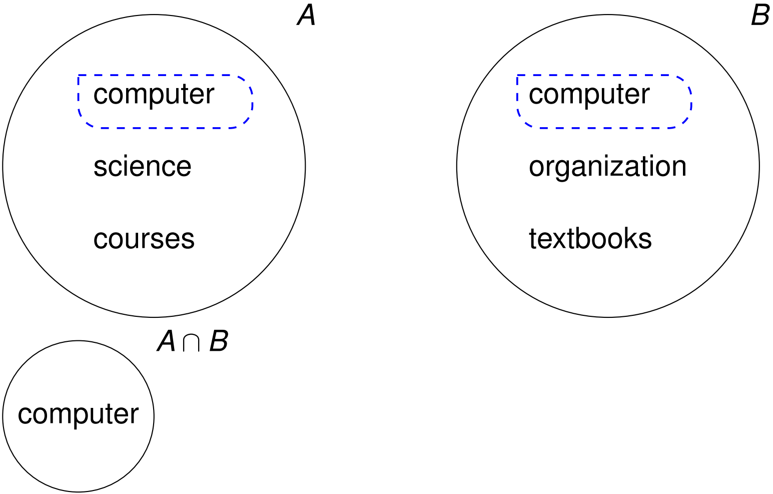
courses

B

computer

organization

textbooks



A

computer

science

courses

B

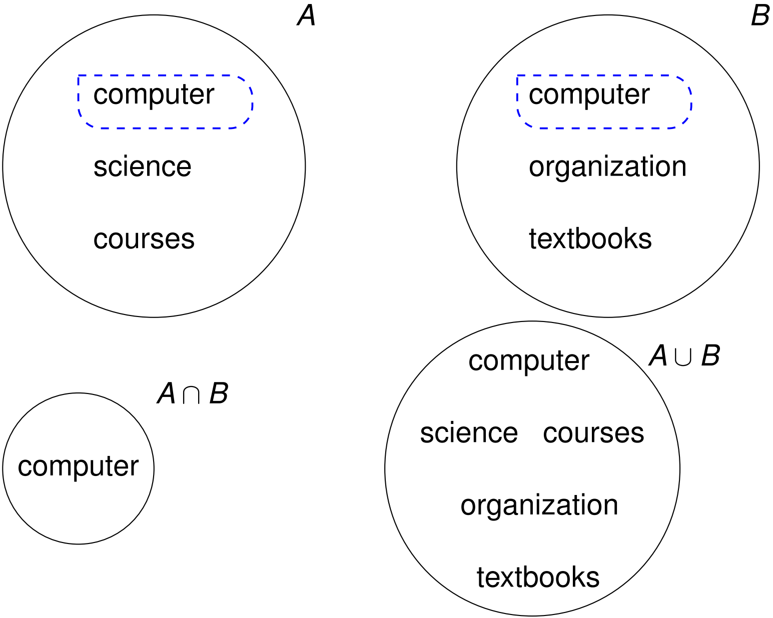
computer

organization

textbooks

$A \cap B$

computer



A

computer

science

courses

B

computer

organization

textbooks

$A \cap B$

computer

$A \cup B$

computer

science courses

organization

textbooks

Thus, $\text{Jacc}(A, B) = \frac{1}{5}$.

In general, Jaccard is actually a similarity in the sense that it has the following properties:

- $0 \leq \text{Jacc}(A, B) \leq 1$

In general, Jaccard is actually a similarity in the sense that it has the following properties:

- $0 \leq \text{Jacc}(A, B) \leq 1$
- $\text{Jacc}(A, B) = 0$ means the strings have no words in common

In general, Jaccard is actually a similarity in the sense that it has the following properties:

- $0 \leq \text{Jacc}(A, B) \leq 1$
- $\text{Jacc}(A, B) = 0$ means the strings have no words in common
- $\text{Jacc}(A, B) = 1$ means the strings are the same, up to re-ordering and duplicate words

In general, Jaccard is actually a similarity in the sense that it has the following properties:

- $0 \leq \text{Jacc}(A, B) \leq 1$
- $\text{Jacc}(A, B) = 0$ means the strings have no words in common
- $\text{Jacc}(A, B) = 1$ means the strings are the same, up to re-ordering and duplicate words
- Higher value generally means “more similar”

What about the strings *dogs chase cats* and *cats chase dogs*?

What about the strings *dogs chase cats* and *cats chase dogs*?

Jacc$(A, B) = 1$, even though the meanings are quite different.

What about the strings *dogs chase cats* and *cats chase dogs*?

$\text{Jacc}(A, B) = 1$, even though the meanings are quite different.

This is an example of what's called the *bag-of-words model*, where we don't care about word order, subject vs. object, etc.