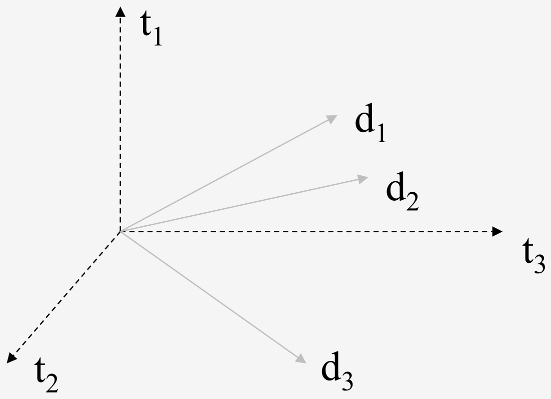
NLP

Text Similarity

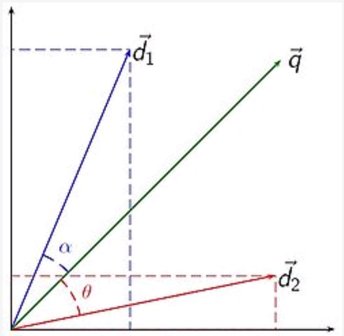
The Vector Space Model

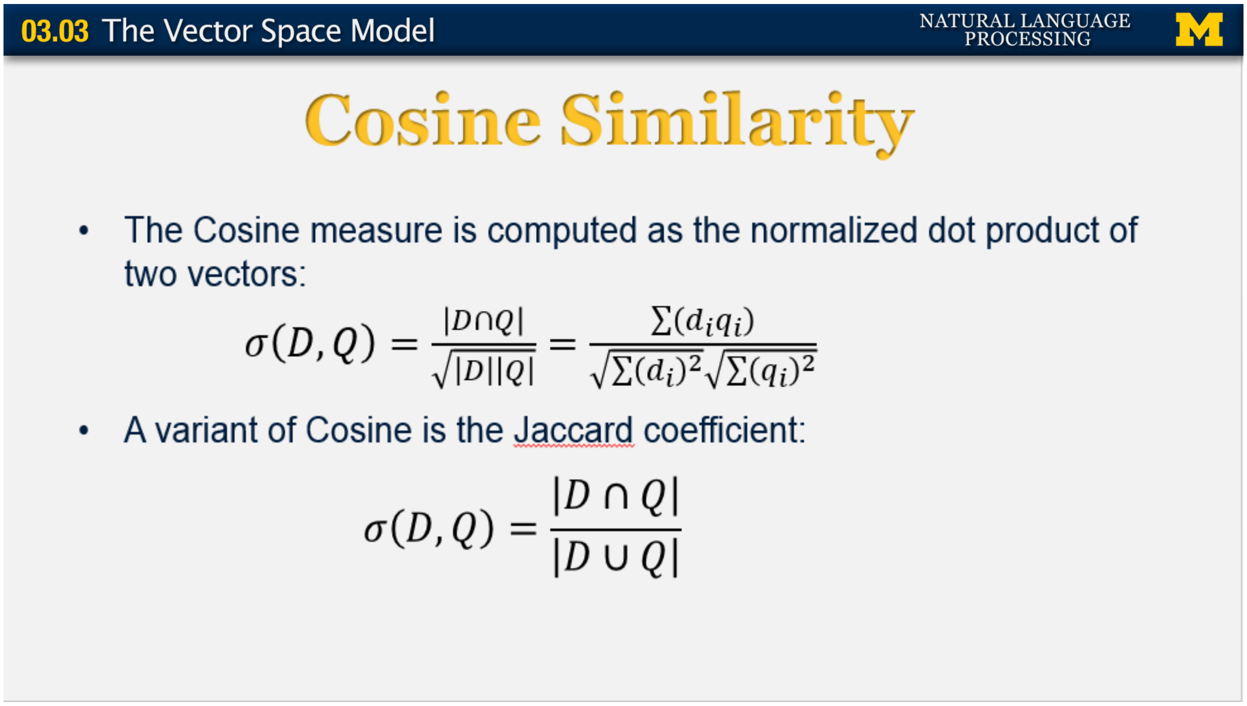
The Vector Space Model



Document Similarity

- Used in information retrieval to determine which document (d_1 or d_2) is more similar to a given query q.
- Note that documents and queries are represented in the same space.
- Often, the angle between two vectors (or, rather, the cosine of that angle) is used as a proxy for the similarity of the underlying documents.





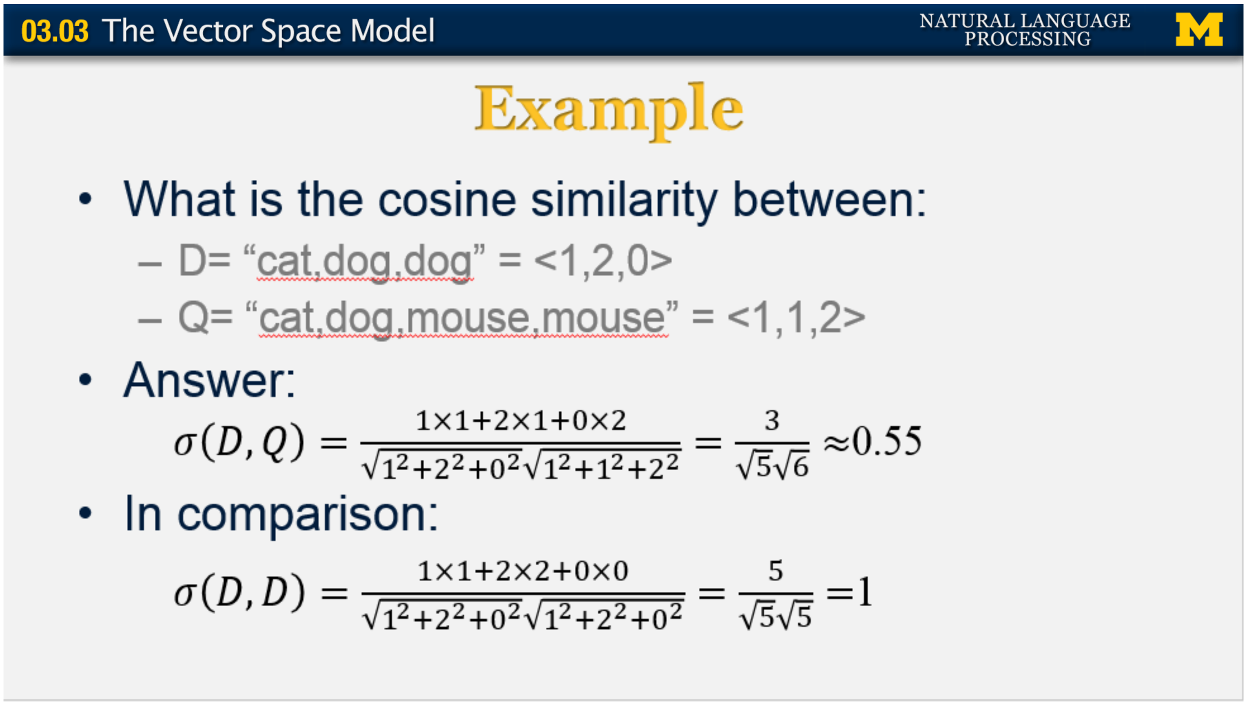
Cosine Similarity

- The Cosine measure is computed as the normalized dot product of two vectors:

$$\sigma(D,Q) = \frac{|D \cap Q|}{\sqrt{|D||Q|}} = \frac{\Sigma(d_i q_i)}{\sqrt{\Sigma(d_i)^2}\sqrt{\Sigma(q_i)^2}}$$

- A variant of Cosine is the Jaccard coefficient:

$$\sigma(D,Q) = \frac{|D \cap Q|}{|D \cup Q|}$$



Example

- What is the cosine similarity between:
  - D= “cat,dog,dog” = <1,2,0>
  - Q= “cat,dog,mouse,mouse” = <1,1,2>
- Answer:

$$\sigma(D,Q) = \frac{1\times1+2\times1+0\times2}{\sqrt{1^2+2^2+0^2}\sqrt{1^2+1^2+2^2}} = \frac{3}{\sqrt{5}\sqrt{6}} \approx 0.55$$

- In comparison:

$$\sigma(D,D) = \frac{1\times1+2\times2+0\times0}{\sqrt{1^2+2^2+0^2}\sqrt{1^2+2^2+0^2}} = \frac{5}{\sqrt{5}\sqrt{5}} = 1$$

Quiz

- Given the three documents

  D_1 = <1,3>

  D_2 = <10,30>

  D_3 = <3,1>

- Compute the cosine scores

  $\sigma(D_1,D_2)$

  $\sigma(D_1,D_3)$

- What do the numbers tell you?

Answers to the Quiz

$\sigma(D_1,D_2) = 1$

one of the two documents is a scaled version of the other

$\sigma(D_1,D_3) = 0.6$

swapping the two dimensions results in a lower similarity

Quiz

- What is the range of values that the cosine score can take?

Answer to the Quiz

- In general, the cosine function has a range of [–1,1]
- However, when the two vectors are both in the first quadrant (since all word counts are non–negative), the range is [0,1].

Text Similarity

The Vector Space Model Applied to Word Similarity

Distributional Similarity

- Two words that appear in similar contexts are likely to be semantically related, e.g.,
  - schedule a test ***drive*** and investigate **Honda**'s financing options
  - **Volkswagen** debuted a new version of its front-wheel-***drive*** Golf
  - the **Jeep** reminded me of a recent ***drive***
  - Our test ***drive*** took place at the wheel of loaded **Ford** EL model
- "You will know a word by the company that it keeps." (J.R. Firth 1957)

Distributional Similarity

- The context can be any of the following:
  - The word before the target word
  - The word after the target word
  - Any word within *n* words of the target word
  - Any word within a specific syntactic relationship with the target word (e.g., the head of the dependency or the subject of the sentence)
  - Any word within the same sentence
  - Any word within the same document

Association Strength

- Frequency matters: we want to ignore spurious word pairings.
- However, frequency alone is not sufficient.
- A common technique is to use pointwise mutual information (PMI).
- Here *w* is a word and *c* is a feature from the context $\mathrm{PMI}(w,c)=\log P(w,c)/P(w)P(c)$

NLP