Cosine Similarity

Natalie Parde
UIC CS 421

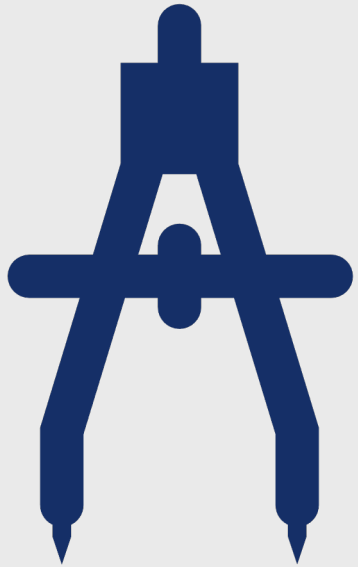
Now that we know how to create a vector space model, how can we use it to compute similarity between words?

- **Cosine similarity**
  - Based on the **dot product** (also called **inner product**) from linear algebra
    - $\text{dot product}(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \cdots + v_N w_N$
- Similar vectors (those with large values in the same dimensions) will have high values; dissimilar vectors (those with zeros in different dimensions) will have low values

Why don't we just use the dot product?

- More frequent words tend to co-occur with more words and have higher co-occurrence values with each of them
- Thus, the **raw dot product will be higher for frequent words**
- This isn't good! ☹
  - We want our similarity metric to tell us how similar two words are regardless of frequency
- The simplest way to fix this problem is to **normalize for the vector length** (divide the dot product by the lengths of the two vectors)

Normalized Dot Product = Cosine of the angle between two vectors



- The cosine similarity metrics between two vectors **v** and **w** can thus be computed as:
  - $\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$
- This value ranges between 0 (dissimilar) and 1 (similar) for frequency or TF-IDF vectors

Example: Computing Cosine Similarity

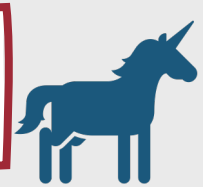
	glitter	data	computer
unicorn	442	8	2
digital	5	1683	1670
information	5	3982	3325

cos(unicorn, information) = ?

Example: Computing Cosine Similarity

	glitter	data	computer
unicorn	442	8	2
digital	5	1683	1670
information	5	3982	3325

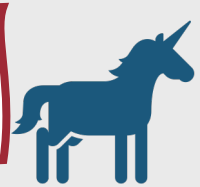
$$\cos(\text{unicorn, information}) = \frac{[442,8,2]\cdot[5,3982,3325]}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}}$$



Example: Computing Cosine Similarity

	glitter	data	computer
unicorn	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{unicorn, information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}}$$



Example: Computing Cosine Similarity

	glitter	data	computer
unicorn	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{unicorn, information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = 0.017$$

Example: Computing Cosine Similarity

	glitter	data	computer
unicorn	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{unicorn, information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = 0.017$$

$$\cos(\text{digital, information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = 0.996$$

Example: Computing Cosine Similarity

	glitter	data	computer
unicorn	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{unicorn, information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = 0.017$$

$$\cos(\text{digital, information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = 0.996$$

Result: *information* is way closer to *digital* than it is to *unicorn*!