Comparing Distributional Profiles:
Cosine Similarity

Remember Distributional Semantics?

- John Rupert Firth and Zellig Harris
  - the most precise way of determining a word’s meaning is by investigating the meanings of the words that occur along with that word.[2]
- Distributional Semantics
  - “linguistic items with similar distributions have similar meanings”[3]

Distributional Profiles

- LL calculations → distributional profiles for every word in your document
- We could analyze these by hand (like the HW)
- But if you have 5000 types in your document...
- you will have a 5000x5000 matrix to analyze
- This is difficult to do by hand

Measuring Similarity

- If we want to find which word meanings are most similar
- We can use cosine similarity
- We have used it already
- But here is how it works

Cosine Similarity

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

- In Python, that looks like this...

Cosine Similarity Python Code

```
c1 = {ἐγώ: 0.00015636973, αὐτός: -0.00656411755…}
c2 = {ἐγώ: 0.58764128248, αὐτός: 0.00000217846…}
terms = set(c1).union(c2)
dotprod = sum(c1.get(k, 0) * c2.get(k, 0) for k in terms)
magA = math.sqrt(sum(c1.get(k, 0)**2 for k in terms))
magB = math.sqrt(sum(c2.get(k, 0)**2 for k in terms))
return dotprod / (magA * magB)
```

- But that is extremely slow, so...

We use the SKLearn function!

```
from sklearn.metrics.pairwise import pairwise_distances
CS_Dists = pairwise_distances(LL, metric = 'cosine')
```

- This does two things:
  - First, it vectorizes the whole calculation
  - Second, it does the whole thing in C
- And now you are finished!

The results!

Greek Word	English Translation	Cosine Distance
θεός	God	1.78745906965E-014
ὁ	the	0.1901116586
σύ	you	0.338482433
ἐξομολογέω	confess	0.3459202137
οὐρανός	heaven	0.3493830955
ἅγιος	holy	0.3554291573
δοῦλος	servant	0.355503958
λαός	people	0.3631526666
χριστός	anointed	0.3776761066
οἶκος	house	0.3779294495

What do they mean?

- They are not synonyms
- Instead, they suggest topics that God is associated with
- The interpretation comes in when looking at what these topics are
- So what do these results tell us about God in the Old Testament?

Homework

- Calculate cosine similarity for the LL lists you produced for last week's homework
- And then interpret:
  - Choose 5 important words
  - Compare the 10 most similar words for each word
  - Do this for all 4 window sizes (4, 8, 12, 16)
  - Write one page:
  - Does this suggest the same window size as LL? What makes you say this?