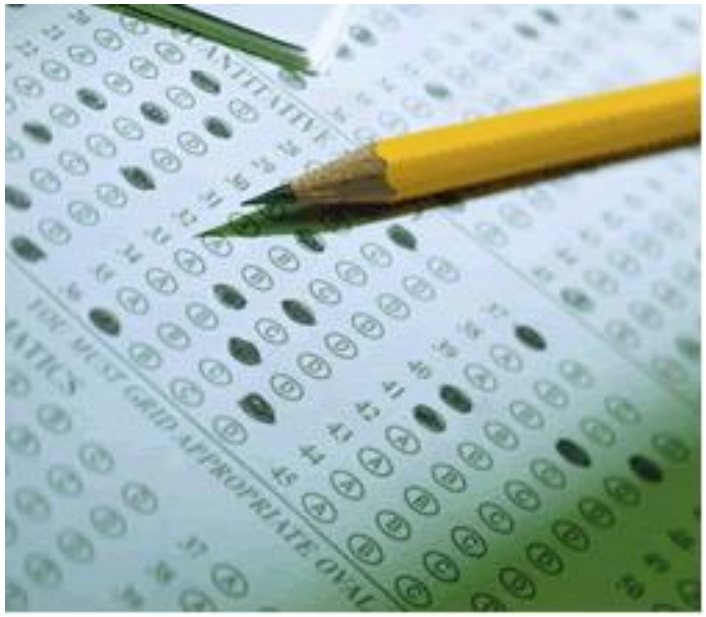
Word Embeddings



SML310: Research Projects in Data Science, Fall 2019

Michael Guerzhoy

The SAT/TOEFL Synonym Task

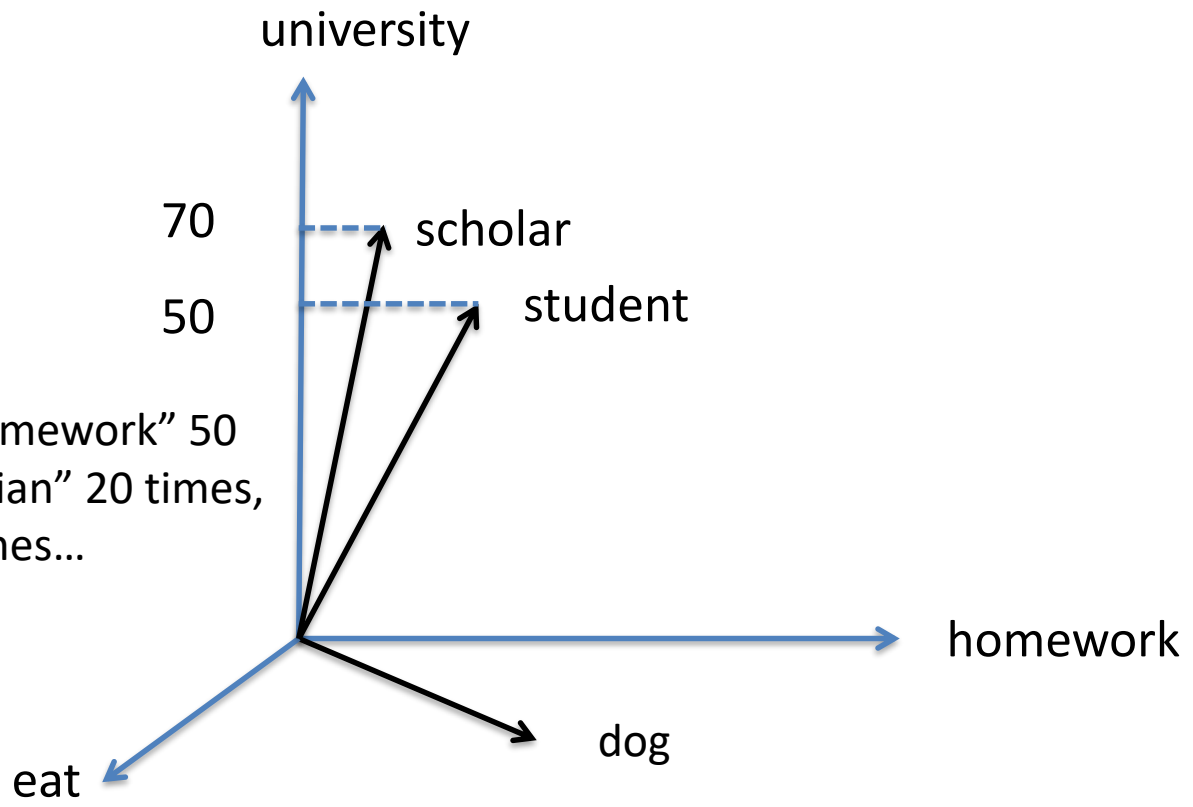
- "Choose the word which best expresses the meaning of the given word"
  - Nifty
    a) Thrifty
    b) Caffeinated
    c) Groovy
    d) Shifty

How Related Are Two Words?

- The Distributional Hypothesis in Linguistics: *words are related if they appear in similar contexts*
- Idea: compute *descriptors* of words by processing novels from Project Gutenberg, and recording how often the word *w* occurs in the same sentence as every other word
  - "dog" occurs together with "homework" 50 times, together with "veterinarian" 20 times, together with "school" 5 times…
  - A good (and common) exercise for using dictionaries/associative arrays!
  - Words are similar if their descriptors are similar

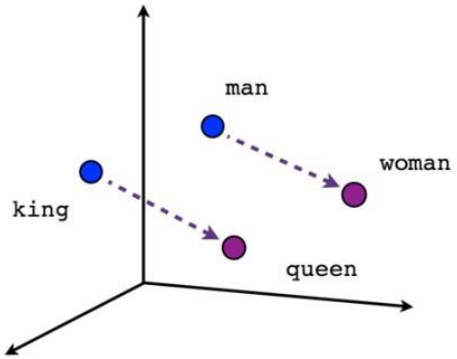
Word Embeddings

- We can represent words as vectors in n-dimensional space!

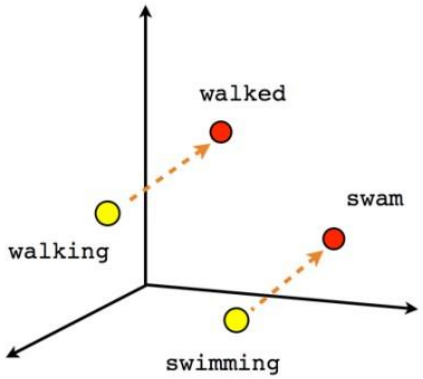


"dog" occurs together with "homework" 50 times, together with "veterinarian" 20 times, together with "university" 0 times…

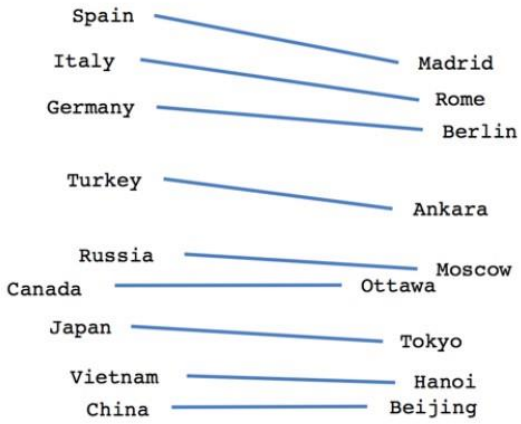
Word Embeddings: Word2Vec



Male-Female



Verb tense



Country-Capital

- Word2Vec is a more sophisticated word embedding scheme than what we're doing here
- With Word2Vec, we can get things like $v_{king} = v_{queen} + (v_{man} - v_{woman})$
  - No explicit encoding: we just use the Distributional Hypothesis again!