COMP-202: Foundations of Programming

Lecture 25: Distributional Semantics

Jackie Cheung, Winter 2016

Announcements

Next class: Final review session

This week:

- Review tutorials
- A4 due tomorrow at 11:59pm
- Quiz 5 (due Friday)
  - Identify the base case and recursive step for two computations

Office Hours

Changed for this week and next week to:

Tuesday, 4pm-6pm, MC108N

Overview

Natural language processing

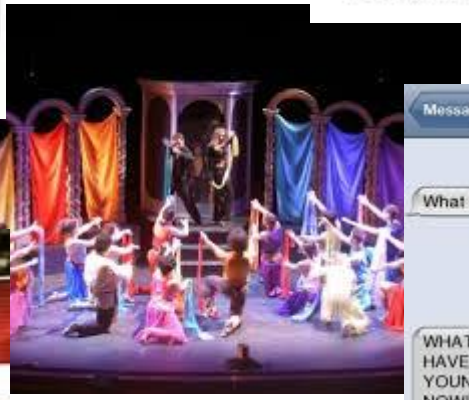
Distributional semantics

Lexical similarity

Language is Everywhere

18.

SHall I compare thee to a Summers day?
Thou art more louely and more temperate:
Rough windes do ſhake the darling buds of Maie,
And Sommers leaſe hath all too ſhort a date:
Sometime too hot the eye of heauen ſhines,
And often is his gold complexion dimm'd,
And euery faire from faire ſome-time declines,
By chance, or natures changing courſe vntrim'd:
But thy eternall Sommer ſhall not fade,
Nor looſe poſſeſſion of that faire thou ow'ſt,
Nor ſhall death brag thou wandr'ſt in his ſhade,
When in eternall lines to time thou grow'ſt,
So long as men can breath or eyes can ſee,
So long liues this, and this giues life to thee,



Languages Are Diverse

6000+ languages in the world

language
langue
ਭਾਸ਼ਾ
語言
idioma
Sprache
lingua

→The Great Language Game
http://greatlanguagegame.com/

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Domains of natural language

- Acoustic signals, phonemes, words, syntax, semantics, …
- Speech vs. text
- **Natural language understanding** (or **comprehension**) vs. **natural language generation** (or **production**)

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Goals

- Language technology applications
- Scientific understanding of how language works

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Methodology and techniques

- Gathering data: language resources
- Evaluation
- Statistical methods and machine learning
- Rule-based methods

Understanding and Generation

Natural language understanding (NLU)

- Language to form usable by machines or humans

Natural language generation (NLG)

- Traditionally, semantic formalism to text
- More recently, also text to text

Most work in NLP is in NLU

- c.f. linguistics, where most theories deal primarily with production

Personal Assistant App

Understanding

Call a taxi to take me to the airport in 30 minutes.

What is the weather forecast for tomorrow?

Generation

Machine Translation

I like natural language processing.

Automatische Sprachverarbeitung gefällt mir.

Understanding

Generation

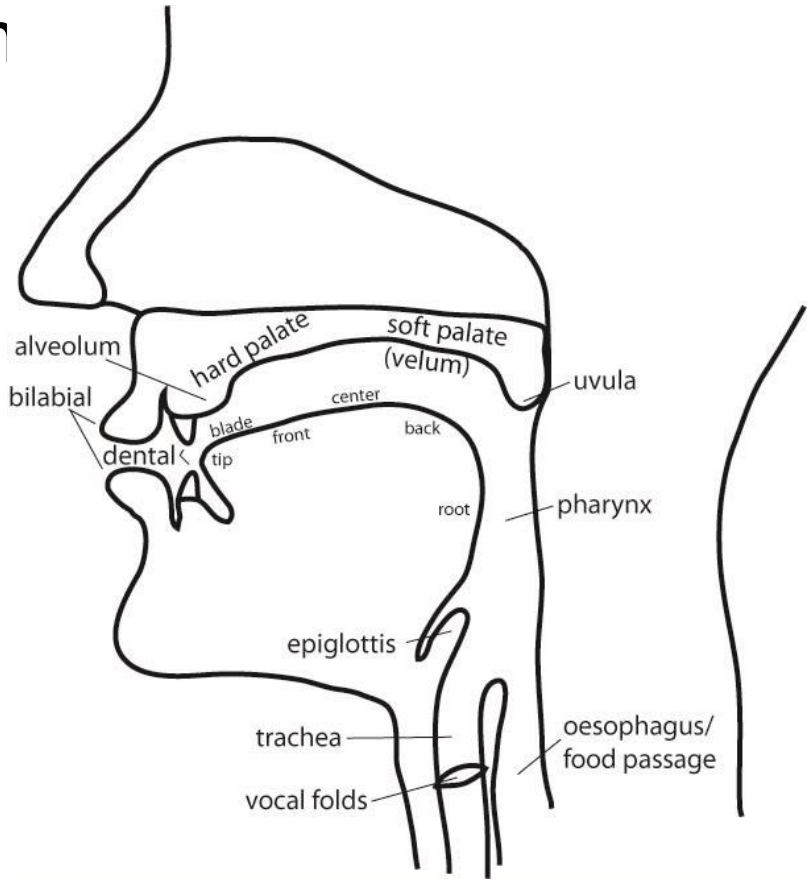
Domains of Language

The grammar of a language has traditionally been divided into multiple levels.

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Phonetics

Study of the speech sounds that make up lan



peach [phi:tsh]

perception

Involves closing of the lips, building up of pressure in the oral cavity, release with aspiration, …

Vowel can be described by its formants, …

Phonology

Study of the rules that govern sound patterns and how they are organized

peach	[phi:tsh]
speech	[spi:tsh]
beach	[bi:tsh]

The p in peach and speech are the same phoneme, but they actually are phonetically distinct!

Morphology

Word formation and meaning

antidisestablishmentarianism

anti- dis- establish -ment -arian -ism

establish

establish**ment**

establishment**arian**

establishmentarian**ism**

disestablishmentarianism

antidisestablishmentarianism

Syntax

Study of the structure of language

*I a woman saw park in the.

I saw a woman in the park.

There are two meanings for the sentence above! What are they? This is called **ambiguity**.

Semantics

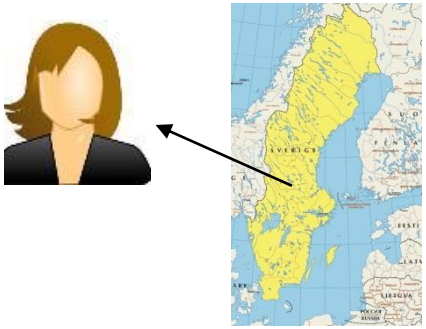
Study of the meaning of language

bank

Ambiguity in the **sense** of the word

Semantics

Ross wants to marry a Swedish woman.

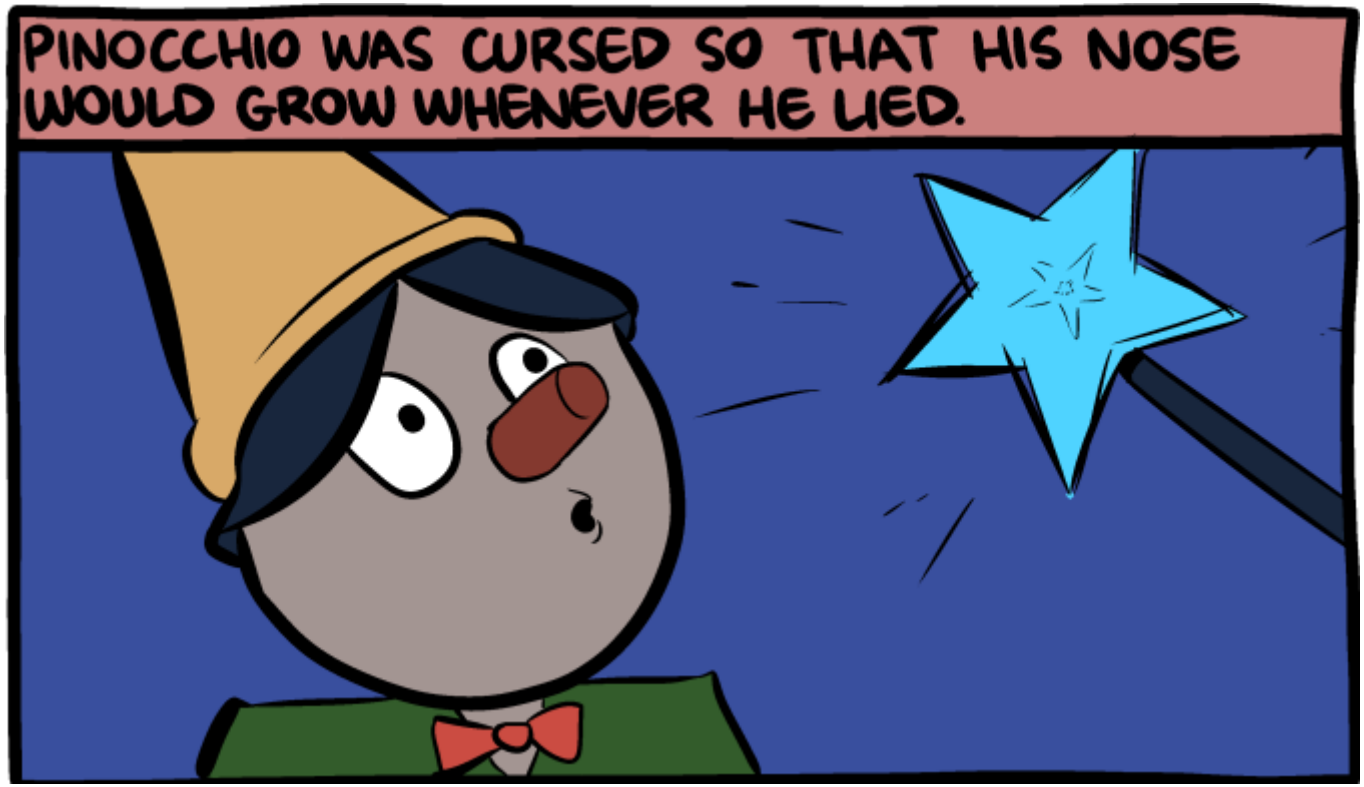


Pragmatics

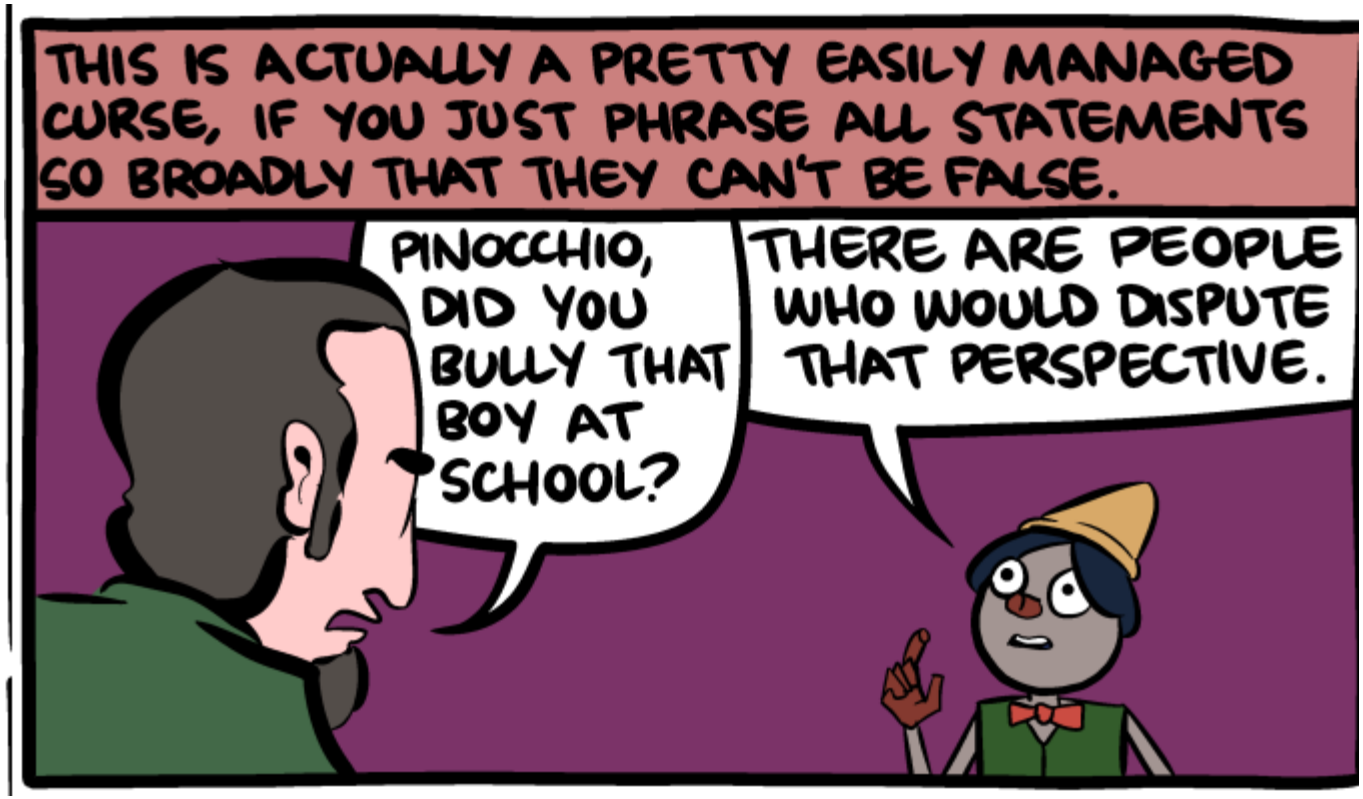
Study of the meaning of language in context.

→ Literal meaning (semantics) vs. meaning in context:

http://www.smbc-comics.com/index.php?id=3730



Pragmatics


THIS IS ACTUALLY A PRETTY EASILY MANAGED CURSE, IF YOU JUST PHRASE ALL STATEMENTS SO BROADLY THAT THEY CAN'T BE FALSE.
PINOCCHIO, DID YOU BULLY THAT BOY AT SCHOOL?
THERE ARE PEOPLE WHO WOULD DISPUTE THAT PERSPECTIVE.

Pragmatics


YOU CAN ALSO ANSWER TOUGH QUESTIONS BY JUST DESCRIBING THE CURRENT SITUATION.
PINOCCHIO! DID YOU EGG MY DOOR?
IF YOU'RE ACCUSING ME OF SUCH A DEED, I HAVE NOTHING MORE TO SAY.

Pragmatics


CURSED TO NEVER LIE, PINOCCHIO CEASED TO INTERACT WITH TRUTH, EVEN IN SAFE SITUATIONS.
SIR, WOULD YOU LIKE ANY DESSERT?
DESSERT WOULD BE DELICIOUS.

Discourse

Study of the structure of larger spans of language (i.e., beyond individual clauses or sentences)

I am angry at her.
She lost my cell phone.

I am angry at her.
The rabbit jumped and ate two carrots.

You are your friends

DISTRIBUTIONAL SEMANTICS

Word Relatedness

One task in **lexical semantics** is the determine how related two words are:

house and *home* are highly related

cat and *dog* are very related

music and *math* are somewhat related

democracy and *metal* are not very related

How can we come up with an automatic method to detect this?

Distributional Semantics

You shall know a word by the company it keeps.

Firth, 1957

Understand a term by the distribution of words that appear near the term

Basic Idea

Go through a corpus of text. For each word, keep a count of all of the words that appear in its context within a window of, say, 5 words.

John Firth was an English linguist and a leading figure in British linguistics during the 1950s.

Term-Context Matrix

Each row is a **vector representation** of a word

	the	*was*	*and*	*British*	*linguist*
Firth	5	7	12	6	9
figure	276	87	342	56	2
linguist	153	1	42	5	34
1950s	12	32	1	34	0
English	15	34	9	5	21

Context words

Co-occurrence counts

Target words

Vector

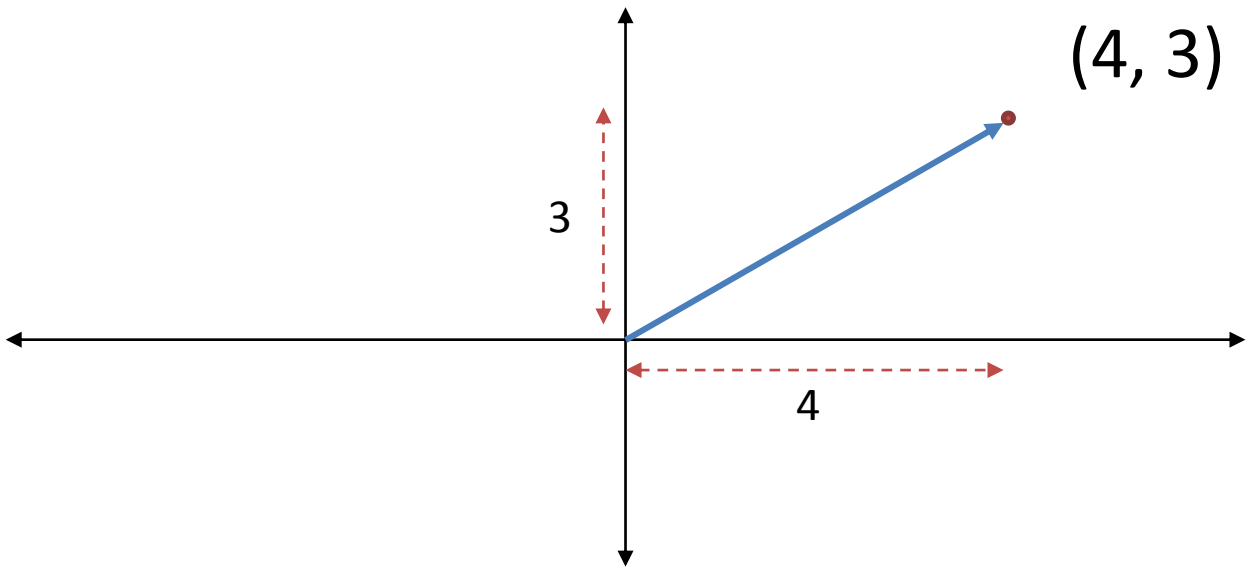
A **vector** as an ordered list of numbers

- The rows of the term-context matrix are each a vector, where each dimension of the vector is associated with a context word

	the	*was*	*and*	*British*	*linguist*
linguist	153	1	42	5	34

Visualizing Vectors

A 2D vector can be plotted onto an X-Y plane



Higher-dimensional vectors can be thought of similarly, but it's harder to visualize them.

Storing Vectors Using HashMap

We can store a vector as a HashMap<String, Integer>

	the	was	and	British	linguist
linguist	153	1	42	5	34

HashMap<String,Integer> with the following (key, value) pairs:

the = 153

was = 1

and = 42

British = 5

linguist = 34

Cosine Similarity

Compare word vectors A and B by

$$\mathrm{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

Dot product

$$A \cdot B = \sum_i A_i \times B_i$$

Vector norm

$$\|A\| = \sqrt{\sum_i A_i^2}$$

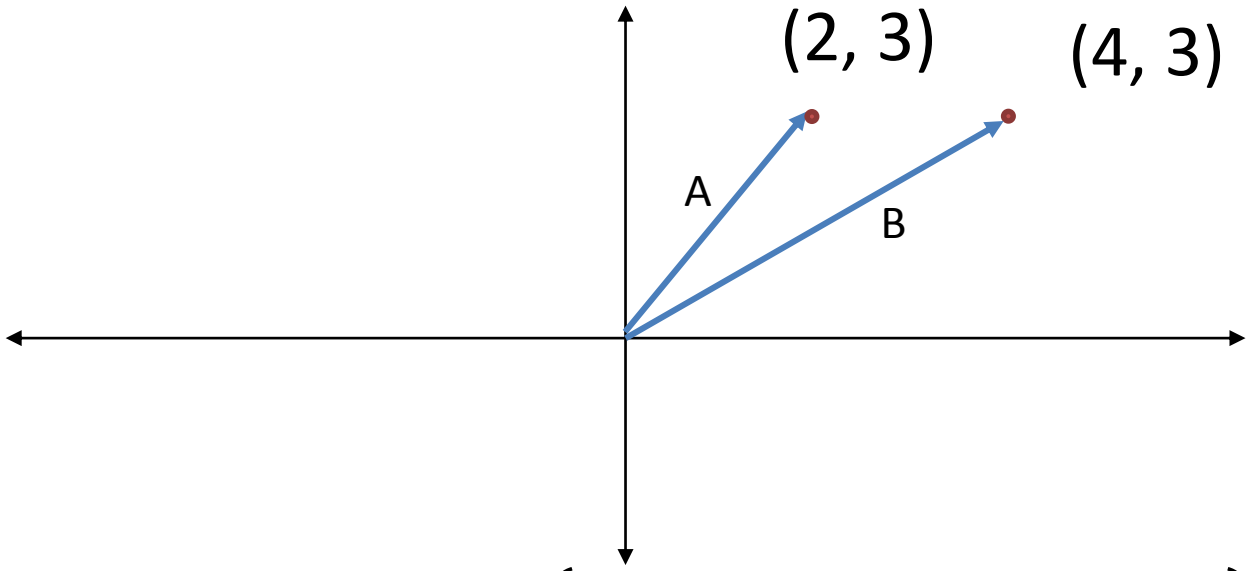
Meaning of Cosine Similarity

This corresponds to the cosine of the angle between the two vectors.

Range of values:

- 0 Vectors are orthogonal (not related)
- 1 Vectors point in the same direction

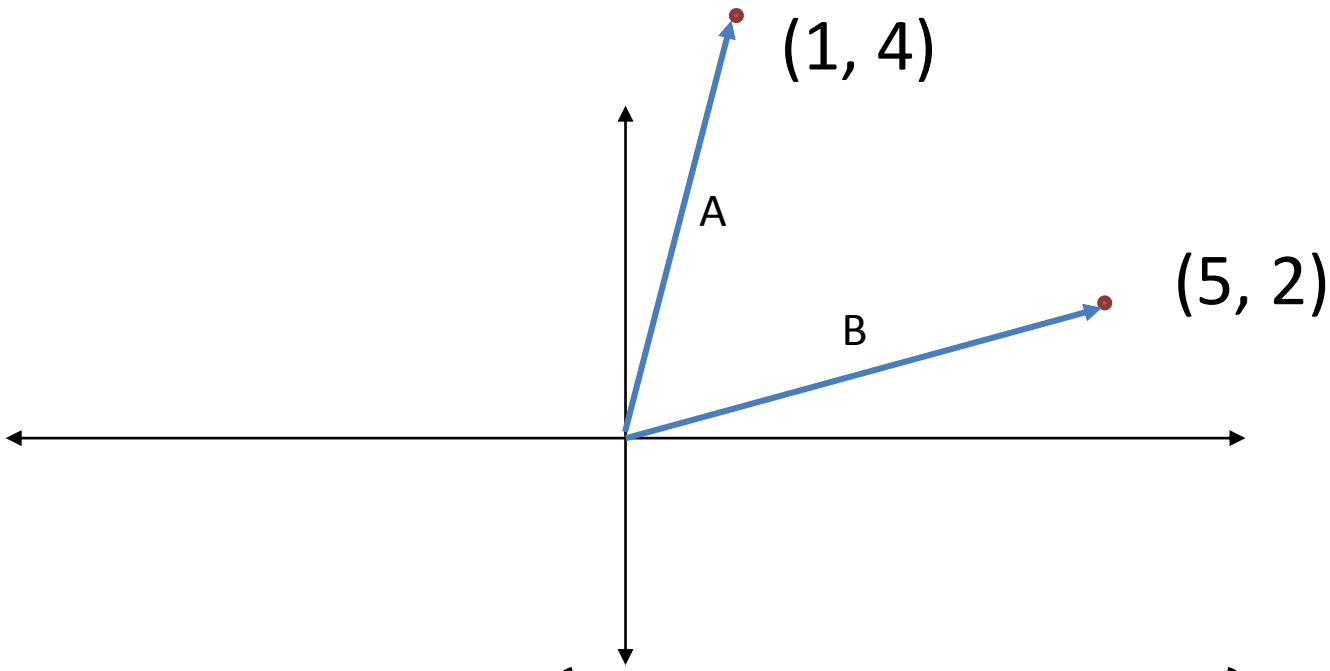
Example: Very Similar



$$\cos(A, B) = \frac{\left((2 \times 4) + (3 \times 3)\right)}{\sqrt{(2^2 + 3^2)}\sqrt{(4^2 + 3^2)}}$$

$$= 0.943$$

Example 2: Less Similar



$$\cos(A, B) = \frac{\left((1 \times 5) + (4 \times 2)\right)}{\sqrt{(1^2 + 4^2)}\sqrt{(5^2 + 2^2)}}$$

$$= 0.585$$

Exercise

Implement cosine similarity

$$\mathrm{sim}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

At-Home Challenge

Implement your own distributional-semantic model that is trained on a **large amount** of text

1. Find a collection of several hundreds/thousands/hundred thousands/millions of documents
2. Decide on a vocabulary of context words and a context window size
3. Write code to create these vectors
4. Write code to test them for word pairs that you know