HSLU Lucerne University of Applied Sciences and Arts

Vector Space Model

NLP
Andreas Marfurt

Information Technology
27.02.2025

Overview

- Representing text
  - Words
  - Sentences/Documents
- Bag of words
  - One-hot encoding
  - TF-IDF
- Cosine similarity

(You have seen this in ADML. Check if my explanation matches your memory.)

Representing Text

- Represent text with numbers, so we can compute something with it
- How should we do that?

Representing Text

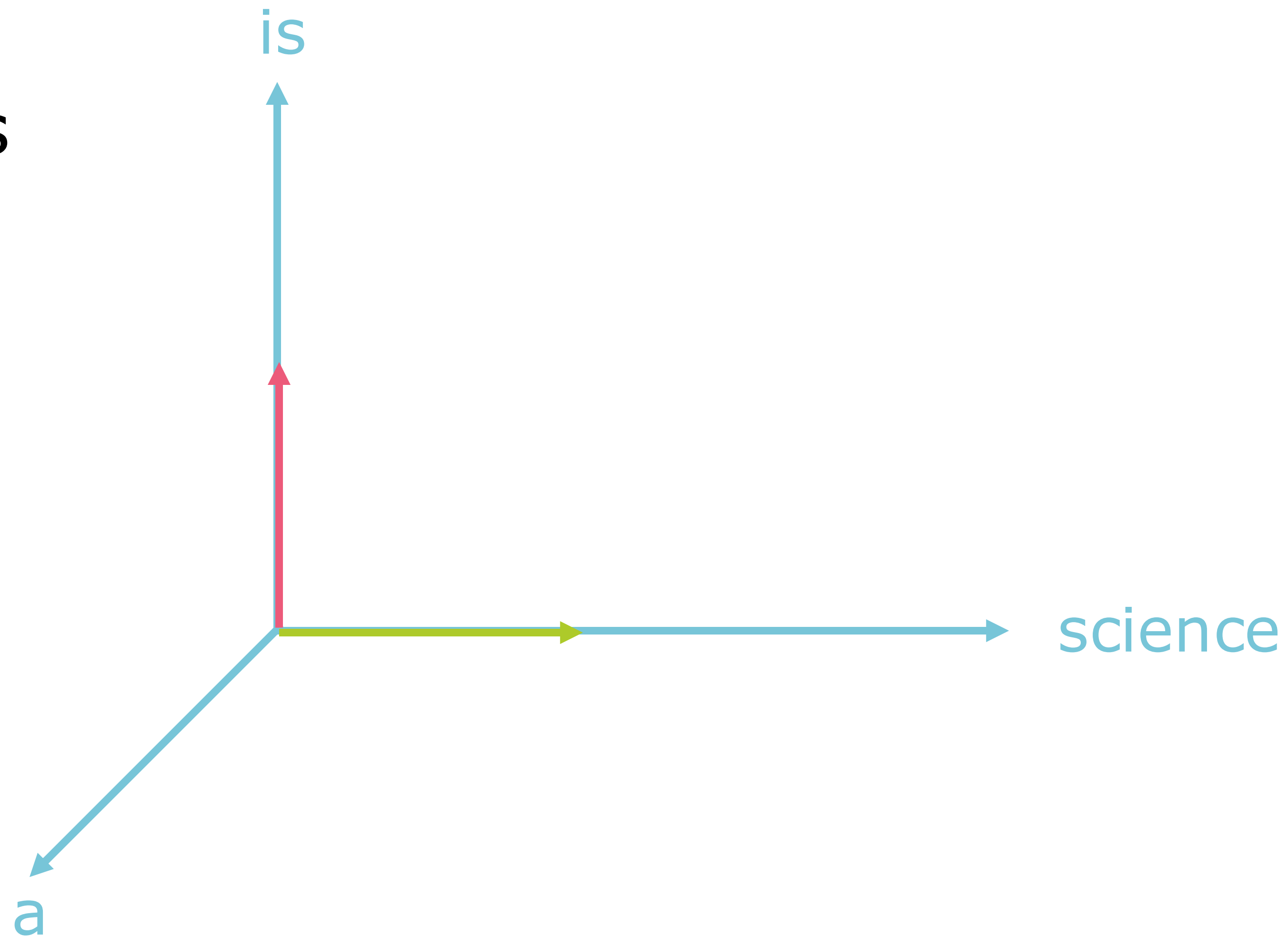
- Document 1: "Science is a rigorous, systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about everything." (Wikipedia)
- Document 2: "The last question was asked for the first time, half in jest, on May 21, 2061, at a time when humanity first stepped into the light." (The Last Question)

Representing Text

- Build a vocabulary: [science, is, a, rigorous, …]
- Use one-hot encoding
  - Vector with length of vocabulary
  - Each dimension is a word
  - Word vector is all 0, except 1 for the position of the word in the vocabulary
  - Example "is": [0, 1, 0, 0, …]

Vector Space Model

- Word vectors



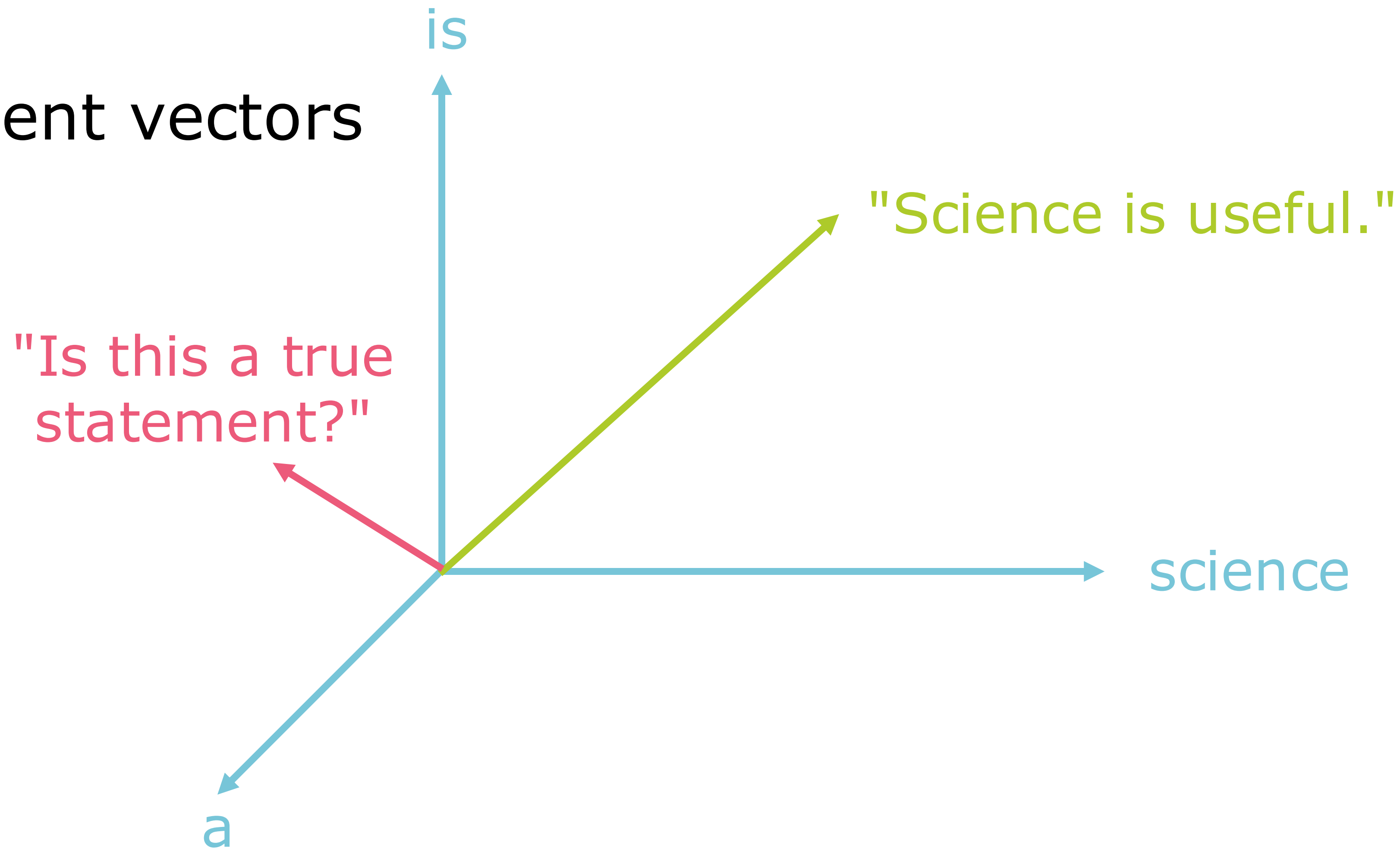
Bag of Words Model

- Document vectors: Combination of all their words
  - 1 if present, 0 if absent (= max of word vectors)
  - Use word count (= sum of word vectors)

- Word ordering doesn't play a role in this model, just counts
  - This is called the "bag of words" (BoW) model

Bag of Words Model

- Document vectors



Bag of Words Model

- Build a word-document matrix (large and sparse)

	Document 1	Document 2	Document 3	Document 4
Word 1	1	0	3	0
Word 2	0	0	1	0
Word 3	0	0	0	1
Word 4	2	0	1	0
Word 5	0	0	0	0
…				

Bag of Words Model

- Build a word-document matrix (large and sparse)

Document representation

Word representation

	Document 1	Document 2	Document 3	Document 4
Word 1	1	0	3	0
Word 2	0	0	1	0
Word 3	0	0	0	1
Word 4	2	0	1	0
Word 5	0	0	0	0
...				

Bag of Words Model

- Pros: Simple to understand, efficient and effective for easy tasks

- Cons: Word order matters, does not relate words (all are equally far apart)
  - cat and feline
    vs.
  - cat and beach

Bag of Words Model

- Intuition: If a term appears 50 times in a document, it is more important for that document than if it appears only once
- Term frequency (TF): TF(term t, doc d) = count(t, d)

- … but 50 times as important?
  - Can use sublinear function, e.g. logarithmic:
    TF(t, d) = ln(1 + count(t, d))

Other variants: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Bag of Words Model

- Some words (stopwords) appear often in almost all documents (the, a, an, is, I, am, …)
  - They are not indicative of the content of the document
- Idea: Down-weight the terms that appear in many documents
- Document frequency (DF_t): In how many documents does term *t* appear?
- Inverse document frequency (IDF):
  $IDF(t) = \log(|D| / DF_t)$

 Other variants: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

TF-IDF Model

- Combines frequency (TF) with how much information the term provides (IDF)
- TF-IDF(t, d) = TF(t, d) x IDF(t)

TF-IDF Model

- Weights of frequent words get reduced

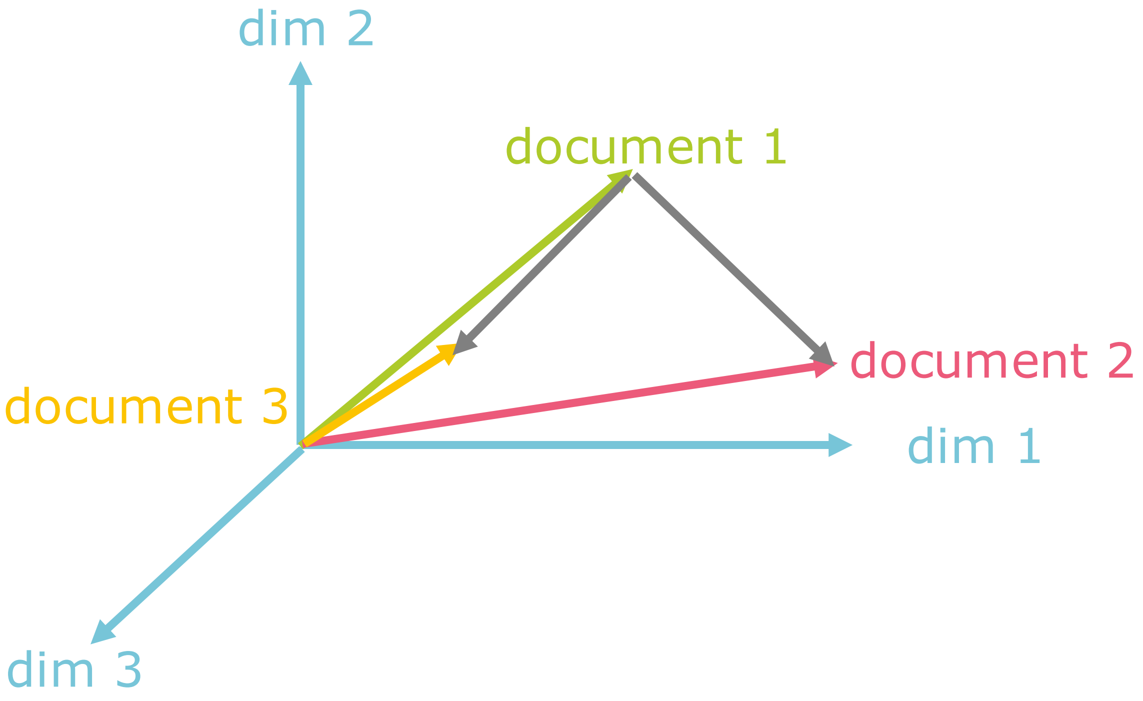
	Document 1	Document 2	Document 3	Document 4
Word 1	0.05	0	0.33	0
Word 2	0	0	0.04	0
Word 3	0	0	0	0.2
Word 4	0.001	0	0.1	0
Word 5	0	0	0	0
...				

Measuring Similarity

- We would like to relate documents, determine their similarity
- Find similar documents
  - Retrieval: Find similar documents to a query "document"
  - Recommendation: Similar articles to the ones you liked

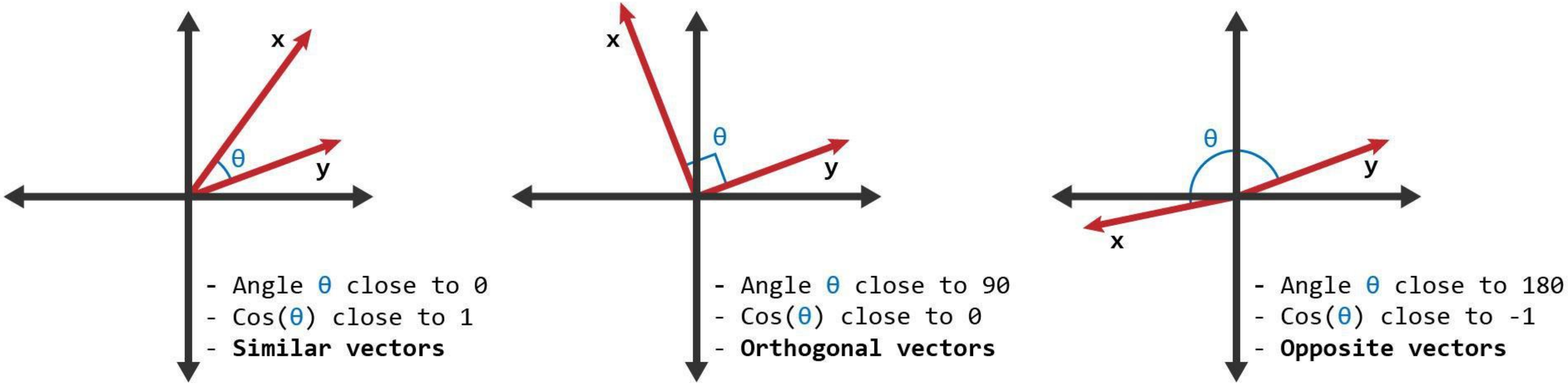
Measuring Similarity

- Euclidean distance:
  Distance between points
  $euclid(d_1, d_2) = |d_1 - d_2|$
- Should d_2 and d_3 be equally far from d_1?
  - Not if our dimensions are terms/semantic concepts
  - Direction is more important than length
- Curse of dimensionality:
  Everything is equally far apart in high dimensions



Measuring Similarity

- Cosine similarity: Angle between vectors
  $\cos(d_1, d_2) = d_1 d_2 /(|d_1||d_2|)$



Vector Space Model

- Pros
  - simple, well-founded approach
  - continuous degree of similarity between queries and documents
  - ranks documents according to relevance
  - allows for partial matching
- Cons
  - documents/queries with similar content but different term vocabularies (e.g., synonyms or plurals) will not be associated
  - word order in documents is ignored (“parking fine” vs. “fine parking”)