Text Mining and Language Analytics

Lecture 4

Word embeddings I

Dr Stamos Katsigiannis

2023-24

Durham University

What do words mean?

- We can check a dictionary!

- Sense/concept
  - The meaning component of a word

- There are relations between senses across different words

Lemma

Definition

For the word “**bat**”

Senses

Definitions of bat

Noun

1. an implement with a handle and a solid surface, usually of wood, used for hitting the ball in games such as baseball, cricket, and table tennis.
2. a mainly nocturnal mammal capable of sustained flight, with membranous wings that extend between the fingers and connecting the forelimbs to the body and the hindlimbs to the tail.

Verb

1. (of a team or a player in sports such as baseball) take in turns the role of hitting rather than fielding.
   “Ruth came to bat in the fifth inning”
2. hit at (someone or something) with the palm of one's hand.
   “he batted the flies away”
3. flutter (one's eyelashes or eyelids), typically in a flirtatious manner.
   “she batted her long dark eyelashes at him”

* Acquired from Google Translate on 25/11/2020: https://translate.google.com

Durham University

How are these related?

How are these related?

cow

horse

How are these related?

Durham University

How are these related?

doctor

hospital

How are these related?

white

black

Durham University

Word relations in terms of sense

- Synonymy → Same meaning in some or all contexts
  - couch/sofa, automobile/car, bike/bicycle
- Antonymy → Opposites with respect to one feature of meaning
  - Define binary opposition or be at opposite ends of a scale
  - Be reversives
  - long/short, fast/slow, dark/light, rise/fall, up/down
- Similarity → Similar meanings. Not synonyms but sharing some element of meaning
  - car/bicycle, cow/horse, pencil/pen
- Relatedness → Related in any way, perhaps via a semantic frame or field
  - car/gasoline, hospital/doctor
- Connotation → Words have affective meanings
  - Positive connotations (e.g. happy), Negative connotations (e.g. sad)

Durham University

Word relations: Semantic field

语意领域

- Words belong to the same semantic field when they:
  - Cover a particular semantic domain
  - Bear structured relations with each other

- Examples:
  - **Universities**
    - Lecturer, professor, student, college, university, faculty
  - **Restaurants**
    - Waiter, menu, food, dish, chef, plate, table, order
  - **Houses**
    - Bed, door, window, kitchen, family, garden, roof
  - **Hospitals**
    - Nurse, doctor, bed, surgeon, anaesthetic, hospital

Word relations: Superordinate/Subordinate (I)

上级/下级

- A sense is subordinate of another if the first sense is more specific, denoting a subclass of the other
  - horse is a subordinate of animal
  - train is a subordinate of vehicle

- A sense is superordinate of another if the first is more broad, denoting a superclass of the other
  - animal is superordinate of horse
  - vehicle is superordinate of train

Example:

Superordinate	animal	vehicle	furniture	fruit	tool	material	colour
Subordinate	horse	train	table	apple	wrench	metal	red

Durham University

Word relations: Superordinate/Subordinate (II)

- **Levels are not symmetric!**

- One level of category is distinguished from the others → **Basic level**

Example:

Superordinate	Basic	Subordinate
vehicle	car	sedan
		coupe
		convertible
	bicycle	road bike
		mountain bike
	aeroplanes	jet aeroplanes
		propeller aeroplanes

Word relations: Superordinate/Subordinate (III)

- Basic level things are “human-sized”
  - Distinctive actions
  - Learned earliest in childhood
  - Names are shorter
  - Names are most frequent

- For example, consider chairs:
  - We know how to interact with a chair (sit/stand up)
  - Not so clear for superordinate categories like furniture
    - Try to imagine a furniture without thinking of a basic level category (e.g. bed, table, chair, sofa, etc.)

Durham University

Senses/Concepts

- **Concepts or Word Senses**
  - Have a complex many-to-many association with words
    - Multiple senses per word
    - Multiple words with the same sense
  - Have relations with each other
    - Synonymy
    - Antonymy
    - Similarity
    - Relatedness
    - Connotation
    - Superordinate/subordinate

Durham University

How to define a sense/concept?

- **Word meaning** → A concept defined by **necessary** and **sufficient** conditions

  必要条件

- Necessary condition for being X → Condition C that X must satisfy in order to be an X
  - **If not C then not X**

充分条件

- Sufficient condition for being X → Condition C that if something satisfies it then it must be an X
  - **If and only if C, then X**

正方形

- Think about conditions for X being a square:
  - "*X has four sides*" is **necessary** to be square
  - The following necessary conditions jointly are **sufficient** to be square
    - *"X has exactly four sides"*
    - *"Each of X's sides is straight"*
    - *"X is a closed figure"*
    - *"X lies in a plane"*
    - *"Each of X's sides is equal in length to each of the others"*
    - *"Each of X's interior angles is equal to the others"*
    - *"The sides of X are joined only at their ends"*

Durham University

Features

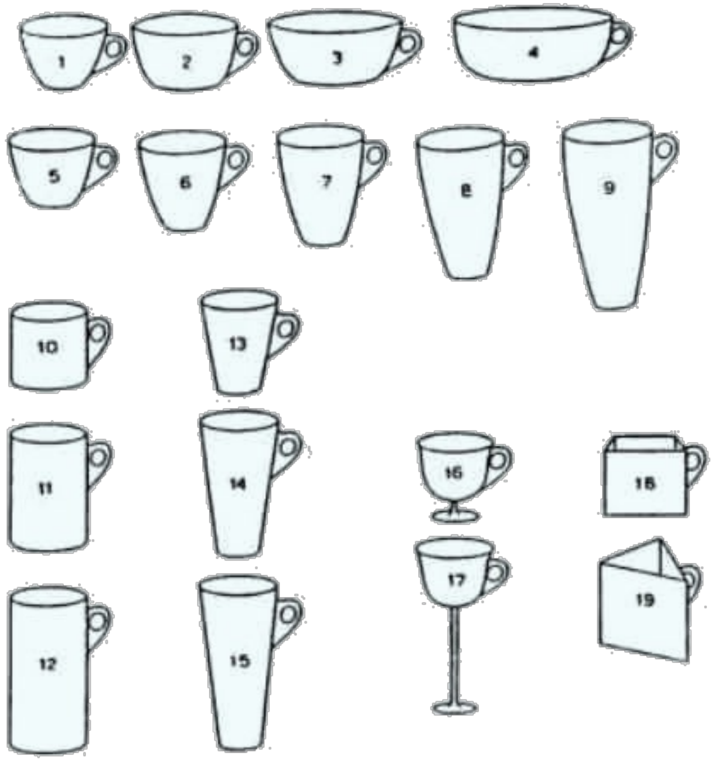
- Consider conditions as features that describe a word

- **Problem:**
  - Features are complex
  - Features may be context-dependent

Is it a cup or a mug? →

A cup is used for tea and is smaller. A Mug is used for coffee or chocolate and is thicker.

* http://www.mugs.coffee/coffee-mug-knowledge/difference-between-cup-and-mug/

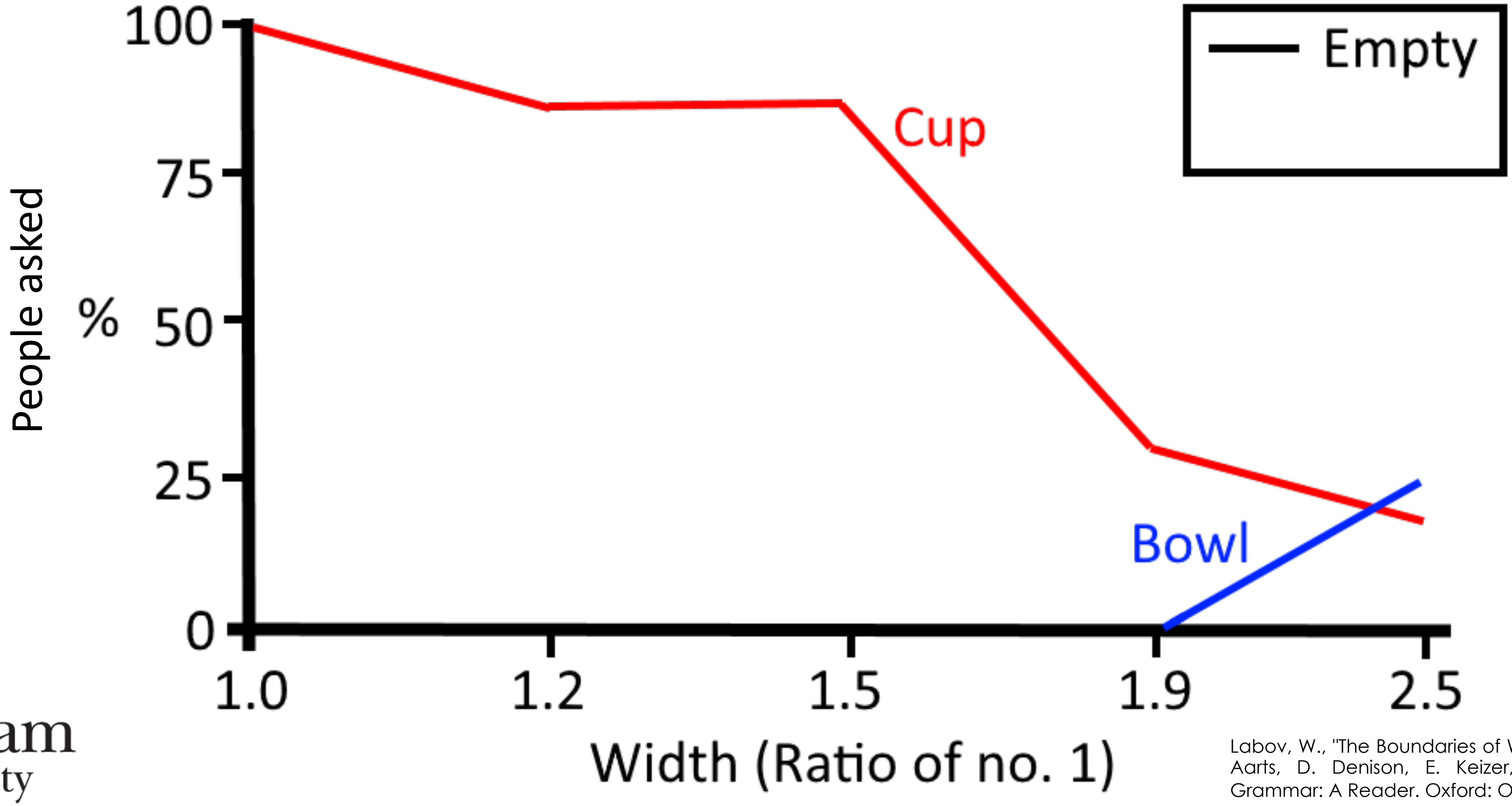


(William Labov, 2004)

Durham University

Category depends on complex features

When does a cup start being a bowl?

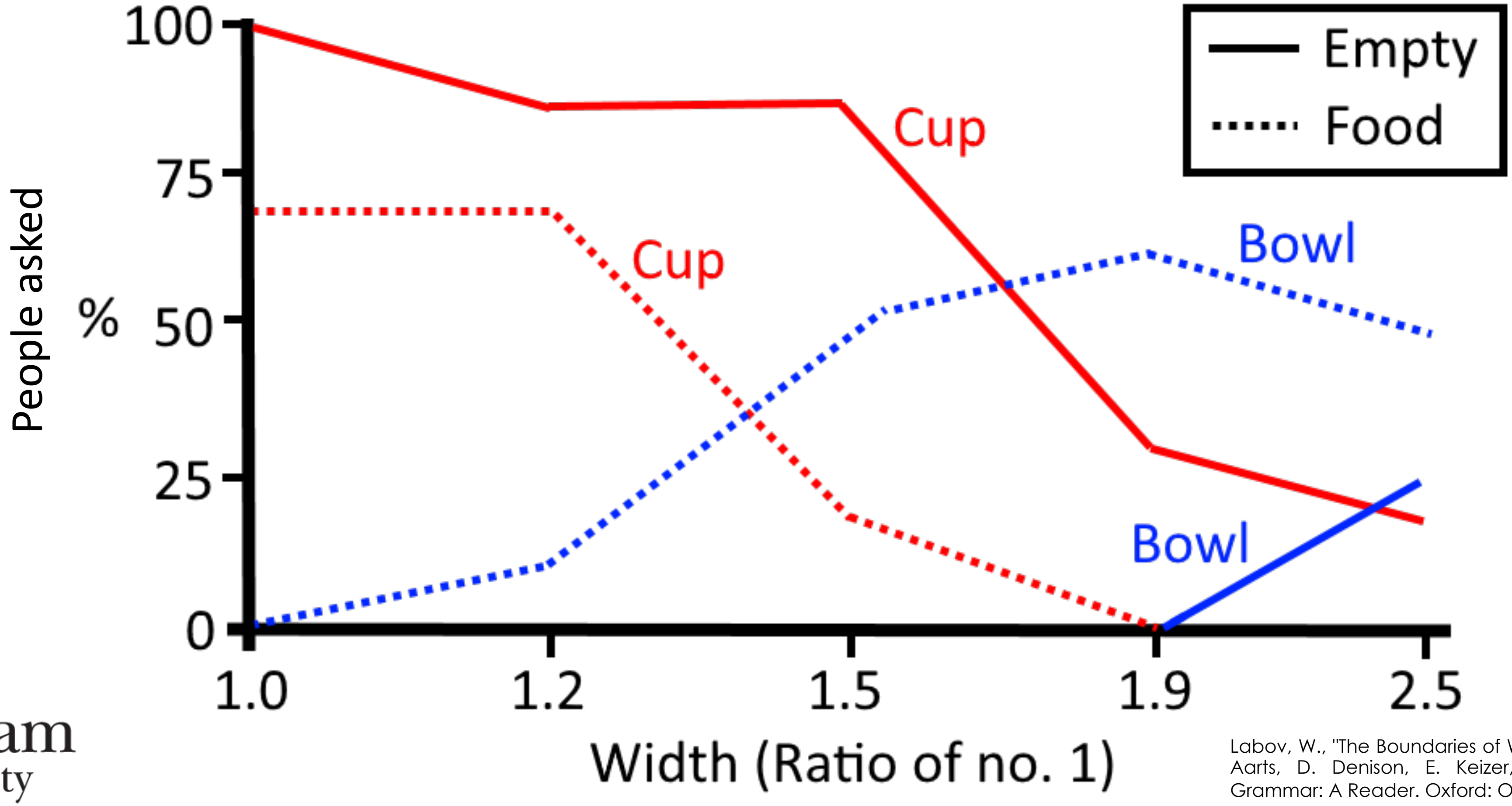


Labov, W., "The Boundaries of Words and their Meanings". In B. Aarts, D. Denison, E. Keizer, & G. Popova (Eds.), Fuzzy Grammar: A Reader. Oxford: Oxford University Press, 2004.

Durham University

Category depends on the context

When does a cup start being a bowl?



Labov, W., "The Boundaries of Words and their Meanings". In B. Aarts, D. Denison, E. Keizer, & G. Popova (Eds.), Fuzzy Grammar: A Reader. Oxford: Oxford University Press, 2004.

Word representation

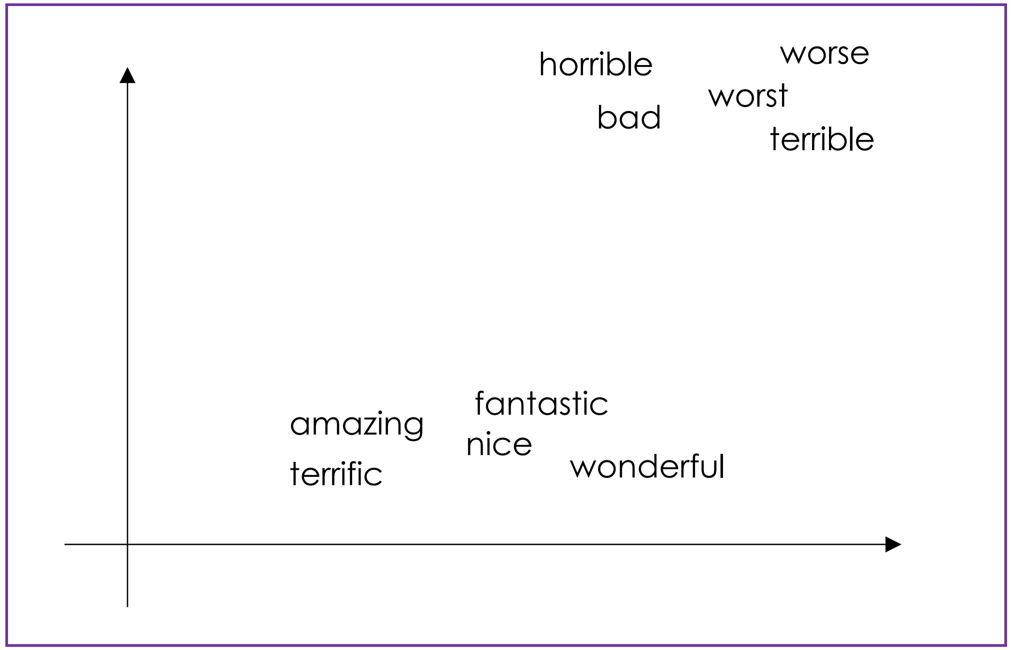
- We can represent words as numerical vectors
- Consider a vocabulary $V=\{w_1, w_2, w_3, \ldots\}$
- w_i can be represented by a vector of $|V|$ elements, with all elements being 0 and the i^{th} element being 1
- Words are now represented by a vector in a $|V|$-dimensional space
- Such vector is called an embedding because it's embedded into a space

嵌入

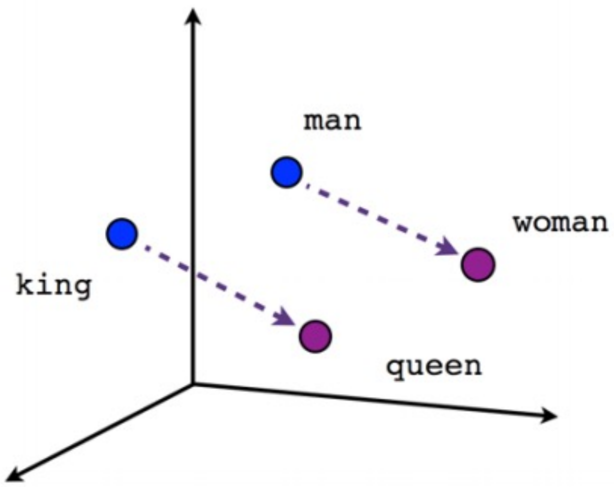
Embeddings and word relations (I)

- **Do word embeddings encode relations between words?**

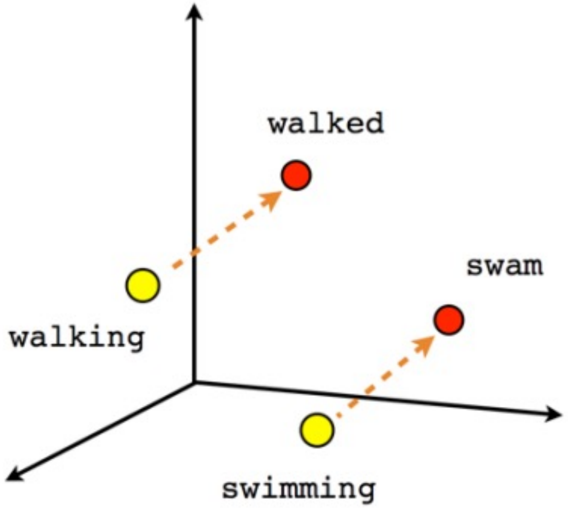
- Ideally, similar and related words should be closer in space than unrelated words



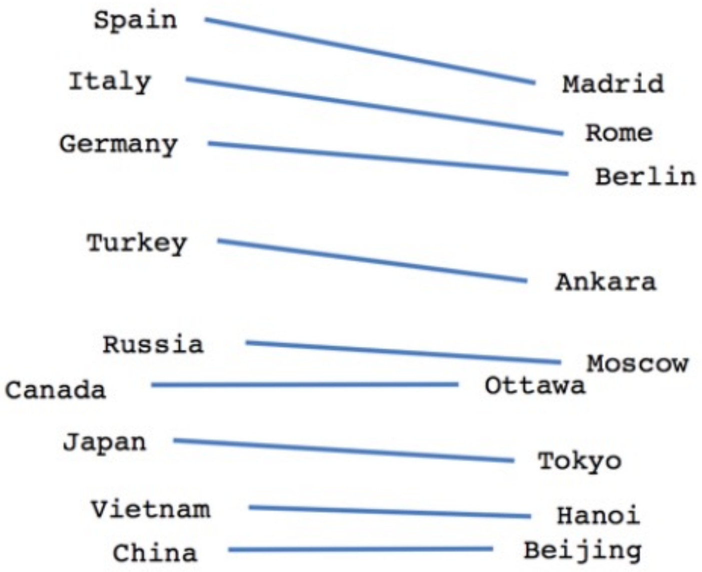
Embeddings and word relations (II)



Male-Female



Verb tense



Country-Capital

Ideally:

vector[king] - vector[queen] ≈ vector(man) - vector(woman)

vector[walking] - vector[walked] ≈ vector[swimming] - vector[swam]

vector[UK] - vector[London] ≈ vector[Japan] - vector[Tokyo]

Durham University

Similarity between embeddings

- **How to measure similarity between embeddings?**

- Distance between vectors can be used!

余弦距离

- **Cosine distance** typically used for word similarity
  - Calculates the cosine of the angle between two vectors

$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\cdot\sqrt{\sum_{i=1}^{N} b_i^2}}$$

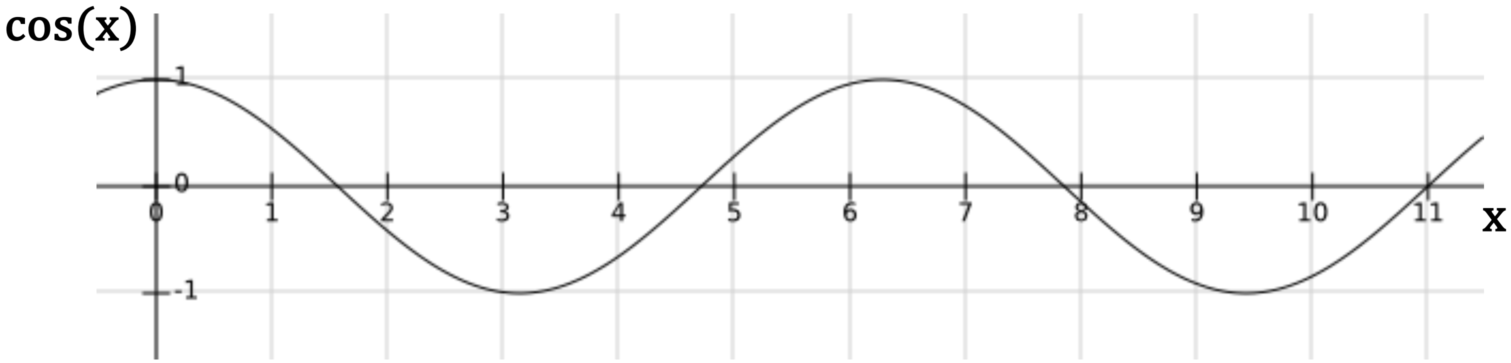
a_i and b_i are the i^th^ elements of vectors a and b

Durham University

Cosine similarity metric

- Remember from linear algebra
  - Dot-product of $(\vec{a}, \vec{b}) \rightarrow \vec{a} \cdot \vec{b} = \sum_{i=1}^{N} a_i b_i = a_1 b_1 + a_2 b_2 + \ldots + a_N b_N$
  - Vector length of $\vec{a} \rightarrow |\vec{a}| = \sqrt{\sum_{i=1}^{N} a_i^2}$
- Cosine value:
  - -1 → Vectors point in opposite directions
  - 1 → Vectors point in same direction
  - 0 → Vectors are orthogonal

For word vectors:
The closer to +1 the more similar the words



Durham University

Embeddings: One-Hot encoding

- **Problem with One-Hot encoding:** Distance between words always the same

- Vocabulary: {car, dog, aeroplane, test}:
  - car → (1, 0, 0, 0)
  - dog → (0, 1, 0, 0)
  - aeroplane → (0, 0, 1, 0)
  - test → (0, 0, 0, 1)

Example

$$\cos(\overrightarrow{car}, \overrightarrow{dog}) = \frac{1 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 + 0 \cdot 0}{(1^2 + 0^2 + 0^2 + 0^2) \cdot (0^2 + 1^2 + 0^2 + 0^2)} = \frac{0}{1} = 0$$

- **All vectors are orthogonal** → $\cos(\vec{a}, \vec{b}) = 0$
  - cos(car, dog) = 0, cos(dog, aeroplane) = 0, cos(dog, test) = 0, …

- What if we use a different distance metric?
  - Euclidean distance → Always equal to $\sqrt{2}$
  - Manhattan distance → Always equal to 2

Euclidean distance

$$Euclidean(\vec{a}, \vec{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_N - b_N)^2}$$

Manhattan distance

$$Manhattan(\vec{a}, \vec{b}) = |a_1 - b_1| + |a_2 - b_2| + \ldots + |a_N - b_N|$$

Durham University

Embeddings: Word context

单词上下文

- **How to capture the context of a word?** → Maybe look at its neighbours!

- "*You shall know a word by the company it keeps*" (J.R. Firth, 1957)

John R. Firth
1890-1960

- Words typically exist within a context
  - e.g. The probability of the word dog appearing within a text about animals is much higher than the probability of the word transistor appearing in the same text
  - e.g. The probability that the word bed would be close to the word sleep is higher than being close to the word stadium

- Context can be:
  - A text
  - A sentence
  - Neighbouring words

Firth, John R., "A synopsis of linguistic theory 1930–1955", in Studies in linguistic analysis, 1–32. Oxford: Blackwell, 1957

Embeddings: Word-Word matrix

- Consider the following text: I like playing football. I enjoy sports. Do I enjoy football?
- Consider the context of a word as the 1 previous and the 1 following word within a sentence
- Word-Word matrix → Frequency (TF) of each pair of words within the context
  - Usually TF-IDF used to avoid bias of very frequent words (e.g. I, the, it, …)

Also known as: Co-occurrence matrix

Context (columns) / Word (rows)

	I	like	playing	football	enjoy	sports	Do
I	0	1	0	0	2	0	1
like	1	0	1	0	0	0	0
playing	0	1	0	1	0	0	0
football	0	0	1	0	1	0	0
enjoy	2	0	0	1	0	1	0
sports	0	0	0	0	1	0	0
Do	1	0	0	0	0	0	0

Word vector

I → (0, 1, 0, 0, 2, 0, 1)
like → (1, 0, 1, 0, 0, 0, 0)
playing → (0, 1, 0, 1, 0, 0, 0)
football → (0, 0, 1, 0, 1, 0, 0)
enjoy → (2, 0, 0, 1, 0, 1, 0)
sports → (0, 0, 0, 0, 1, 0, 0)
Do → (1, 0, 0, 0, 0, 0, 0)

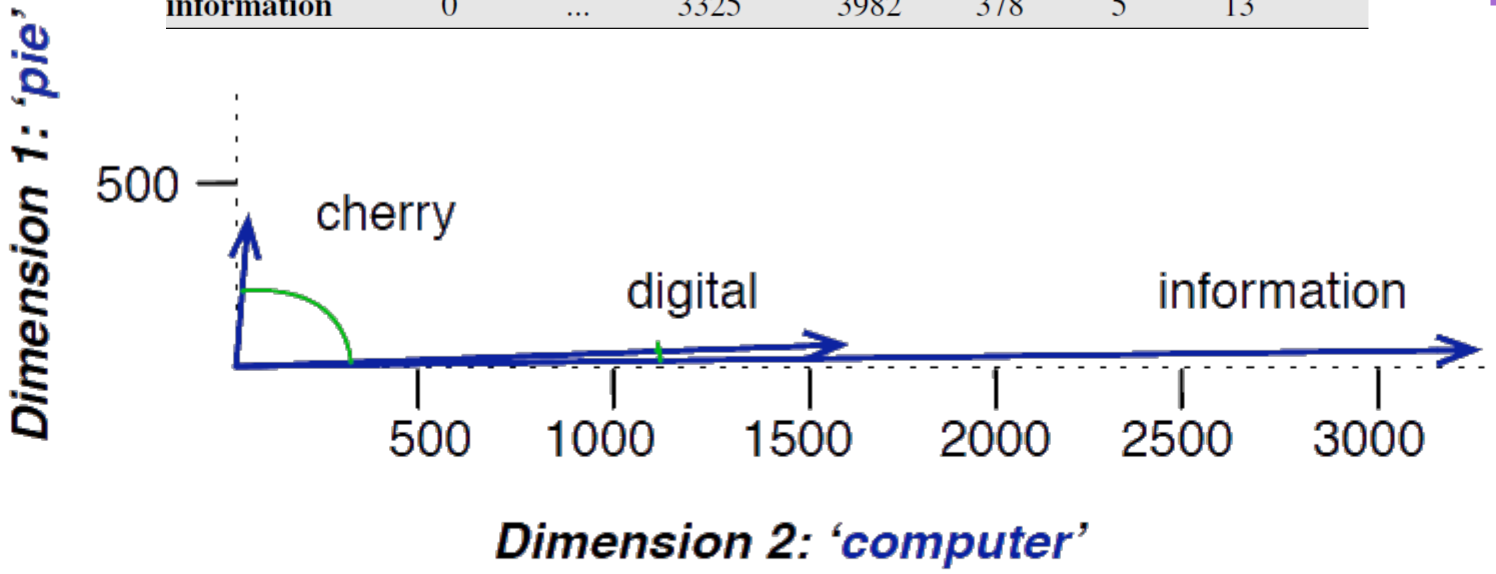
Embeddings: TF

Words cherry, information and digital projected on the **computer** and **pie** dimensions:

$$\textit{cosine}(\text{information}, \text{digital}) < \textit{cosine}(\text{information}, \text{cherry})$$

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

Co-occurrence matrix for four words in the Wikipedia corpus



Durham University

Daniel Jurafsky, James H. Martin, "Speech and Language Processing", 3rd edition

Questions?

Durham University