

Lecture 3:
Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Lecture 3: Similarity

COMP90049
Knowledge Technologies

<https://powcoder.com>

Sarah Erfani and Karin Verspoor and Jeremy Nicholson, CIS

Add WeChat powcoder

Semester 2, 2017



THE UNIVERSITY OF
MELBOURNE

Lecture 3: Similarity

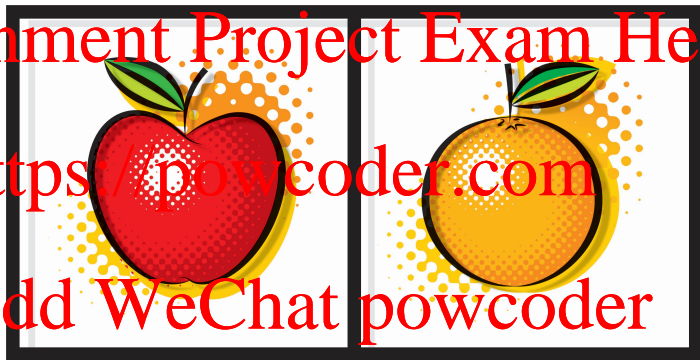
COMP90049
Knowledge
Technologies

Comparing things

Sets of descriptors
Features, Vectors

Comparing Documents

Distance
Measures



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things

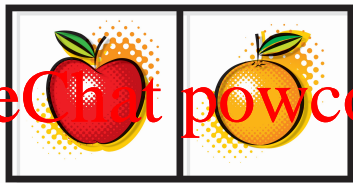
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures



<https://powcoder.com>



Add WeChat powcoder

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

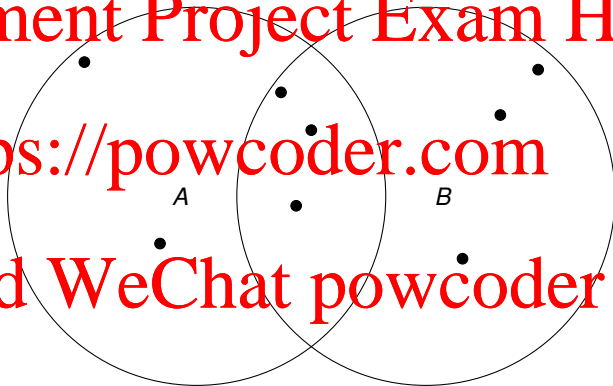
Comparing
Documents

Distance
Measures

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things

Sets of descriptors

Features, Vectors

Comparing
Documents

Distance
Measures

Many similarity assessments can be framed as set intersection.

- Amazon: Book purchases
- Netflix: Movies that you have watched

Recommendations

- Rating sets (stars)
 - thresholding using ratings
 - different subsets for different ratings
- Categories of items
 - generalisation
 - book or movie genres

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Assignment Project Exam Help

We have discussed similarity at an intuitive level.
<https://powcoder.com>
How do we measure similarity quantitatively?

Add WeChat powcoder

Lecture 3: Similarity

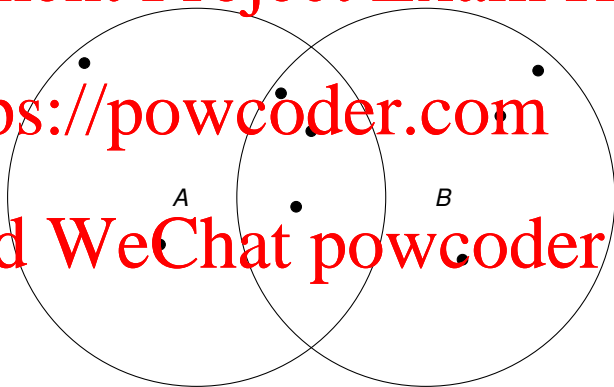
COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

$$\frac{|A \cap B|}{|A \cup B|}$$



$$\text{sim}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{8}$$

Lecture 3: Similarity

COMP90049

Knowledge

Technologies

Comparing things

Sets of descriptors

Features, Vectors

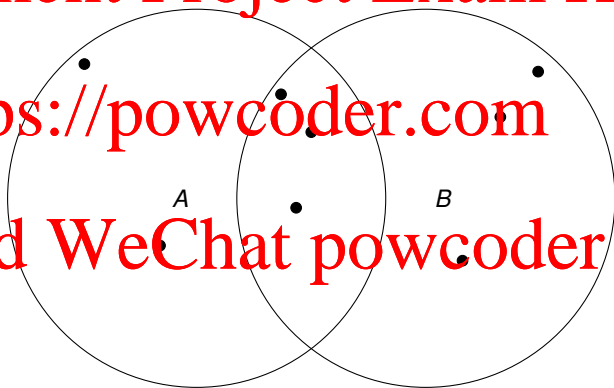
Comparing

Documents

Distance

Measures

$$\frac{2|A \cap B|}{|A| + |B|}$$



$$sim(A, B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2 * 3}{5 + 6} = \frac{6}{11}$$

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Assignment Project Exam Help

A model is our attempt to understand and represent the nature of reality through a particular lens, be it architectural, biological or mathematical.

<https://powcoder.com>

An model is an **abstraction** of the entity that we are trying to model, c.f. fruit above.

Add WeChat powcoder

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things

Sets of descriptors

Features, Vectors

Comparing
Documents

Distance
Measures

A *feature vector* is an n -dimensional vector of *features* that represent some object.

A *feature* or *attribute* is any distinct aspect, quality, or characteristic of that object

<https://powcoder.com>

- Features may be symbolic/categorical/discrete (e.g. colour, gender)
- Features may be ordinal (e.g. cool < mild < hot [temperature])
- Features may be numeric/continuous (e.g., height, age)

Add WeChat powcoder

A vector locates an object (document, person, ...) as a point in n -space. The angle of the vector in that space is determined by the relative weight of each term.

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things

Sets of descriptors

Features, Vectors

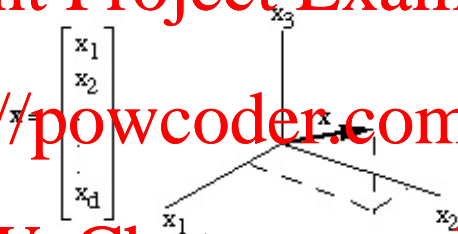
Comparing
Documents

Distance
Measures

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things

Sets of descriptors

Features, Vectors

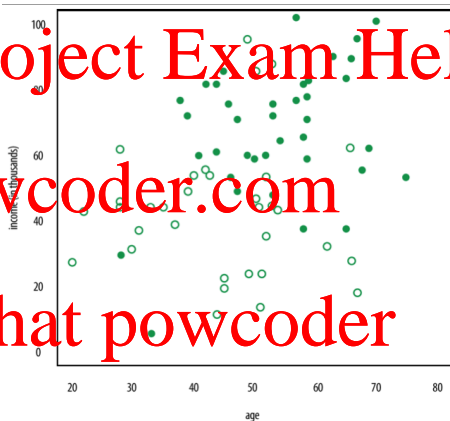
Comparing

Documents

Distance

Measures

age	income	credit
33	8	low
58	42	low
49	79	low
49	17	low
58	26	high
44	71	high
...		



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

How should we compare documents to assess their similarity?

- String-level similarity (e.g., edit distance)
- Sets of common substrings (sentences, phrases, words, n-grams)
- “bag of words”

How similar are these sentences?

- 1 Mary is quicker than John.
- 2 John is quicker than Mary.
- 3 Mary is slower than John.
- 4 Jane is quicker than Mary.

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Assignment Project Exam Help

1 Mary is quicker than John.

2 John is quicker than Mary.

3 Mary is slower than John.

4 Jane is quicker than Mary.

<https://powcoder.com>

Add WeChat powcoder

Sentence	"Mary"	"John"	"Jane"	"quicker"	"slower"
1	1	1	0	1	0
2	1	1	0	1	0
3	1	1	0	0	1
4	1	0	1	1	0

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

One of the earliest models proposed for retrieval of documents (information retrieval in 1962) was the vector-space model.

Suppose there are n distinct indexed terms in the collection. Then each document d can be thought of as a vector

$$\langle w_{d,1}, w_{d,2}, \dots, w_{d,t}, \dots, w_{d,n} \rangle$$

where $w_{d,t}$ is a weight describing the importance of term t in d .

(Most $w_{d,t}$ values will be zero, because most documents only contain a tiny proportion of a collection's terms.)

Intuitively, if some other document d' has a vector

$$\langle w_{d',1}, w_{d',2}, \dots, w_{d',t}, \dots, w_{d',n} \rangle$$

where the weights are close to those of d – in particular, if the non-zero w values are for much the same set of terms – then d and d' are likely to be similar in topic.

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Assignment Project Exam Help

We have discussed similarity at an intuitive and quantitative level.

<https://powcoder.com>

$$sim_J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{8}$$

$$sim_D(A, B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2 * 3}{5 + 6} = \frac{6}{11}$$

Add WeChat powcoder

What is the relationship between similarity and distance?

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

A distance measure on a space is a function that takes two points in a space as arguments.

- 1 No negative distances.

$$d(x, y) \geq 0$$

- 2 Distances are positive, except for the distance from a point to itself.

$$d(x, y) = 0 \text{ if and only if } x = y$$

- 3 Distance is symmetric.

$$d(x, y) = d(y, x)$$

- 4 The *triangle inequality* typically holds.
(Distance measures the length of the *shortest path* between two points.)

$$d(x, y) \leq d(x, z) + d(z, y)$$

Euclidean Distance

Lecture 3: Similarity

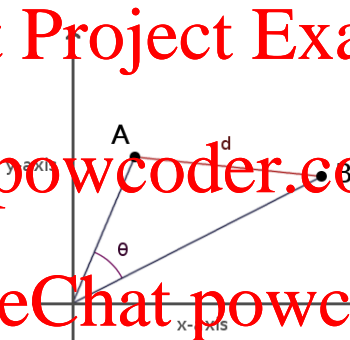
COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Given two items A and B , and their corresponding feature vectors \vec{a} and \vec{b} , respectively, we can calculate their similarity via their distance d in euclidean space:



Assignment Project Exam Help

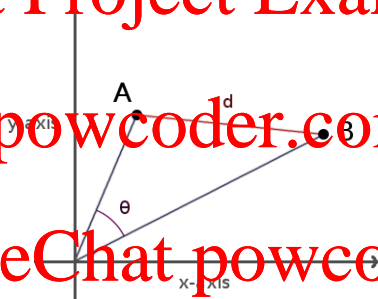
<https://powcoder.com>

Add WeChat powcoder

In n-dimensional space:

$$d(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

Given two items A and B , and their corresponding feature vectors \vec{a} and \vec{b} , respectively, we can calculate their similarity via their *vector cosine* (the cosine of the angle θ between the two vectors):



$$\text{sim}(A, B) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}}$$

Lecture 3: Similarity

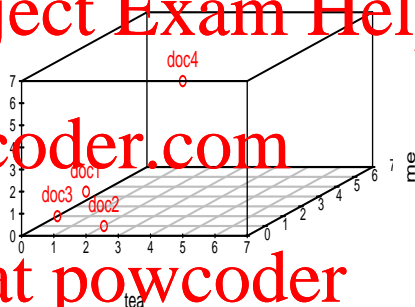
COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Point	tea	me	two
doc1	2	0	2
doc2	2	1	0
doc3	0	2	0
doc4	5	0	7

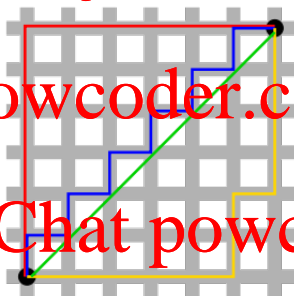


Add WeChat powcoder

- Doc4, like Doc1, is all about “tea” and “two”.
- But because it is longer, it is in a space by itself.

["City block" distance or "Taxicab geometry" or " L_1 distance"]

Given two items A and B , and their corresponding feature vectors \vec{a} and \vec{b} respectively, we can calculate their similarity via their distance d based on the absolute differences of their cartesian coordinates.



<https://powcoder.com>

Add WeChat powcoder

In n -dimensional space:

$$d(A, B) = \sum_{i=1}^n |a_i - b_i|$$

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

Relative entropy:

$$D(x \parallel y) = \sum_i x_i (\log_2 x_i - \log_2 y_i)$$

or alternatively *skew divergence*:

$$D_\alpha(x, y) = D(x \parallel \alpha y + (1 - \alpha)x)$$

or *Jensen-Shannon divergence*:

$$JSD(x, y) = \frac{1}{2} D(x \parallel m) + \frac{1}{2} D(y \parallel m)$$

where $m = \frac{1}{2}(x + y)$

NB: Probability will be reviewed next lecture!

Lecture 3: Similarity

COMP90049
Knowledge
Technologies

Comparing things
Sets of descriptors
Features, Vectors

Comparing
Documents

Distance
Measures

- How can we represent a set of objects?
- What are some methods for measuring similarity between objects?

Reading

- On distance measures:
Chapter 3, especially Section 3.5
Mining of Massive Datasets

<http://infolab.stanford.edu/~ullman/mmds.html>

- On document representation:
Chapter 6
Information Retrieval, Manning *et al.*

<http://nlp.stanford.edu/IR-book/html/htmledition/scoring-term-weighting-and-the-vector-space-model-1.html>