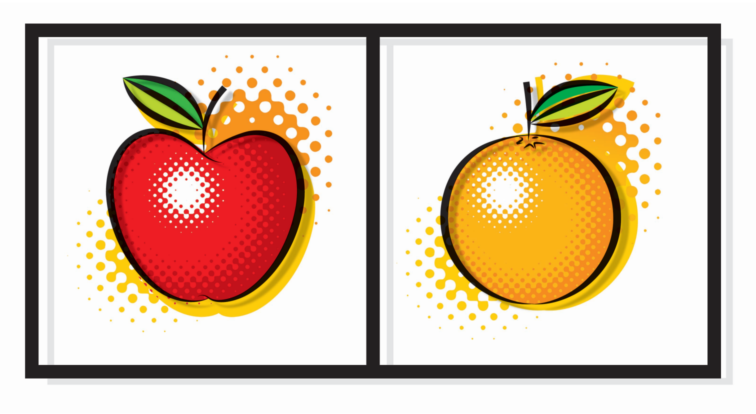
Lecture 3: Similarity

COMP90049
Knowledge Technologies

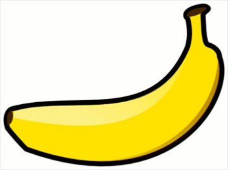
Sarah Erfani and Karin Verspoor, CIS

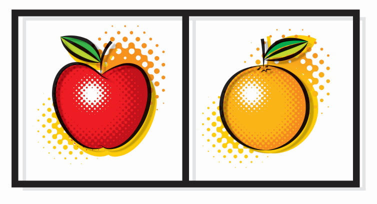
Semester 2, 2018

Compare and Contrast

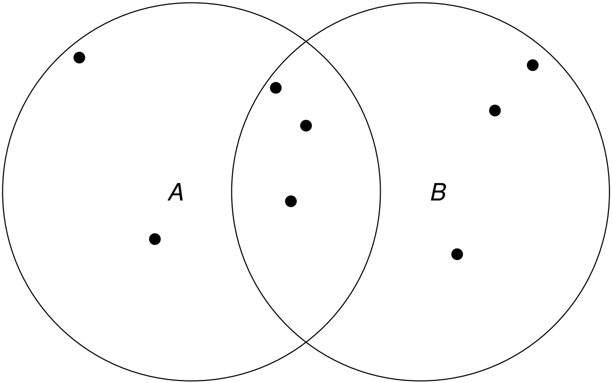


Compare and Contrast





Venn Diagram



Similarity as Set intersection

Many similarity assessments can be framed as set intersection.

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

Refinements

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

Comparing Documents

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.

Word Vectors

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

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

Feature vectors

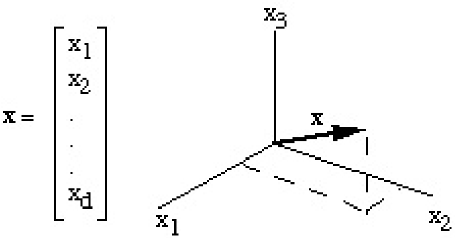
A *feature vector* is an n-dimensional vector of *features* that represent some object.

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

- 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)

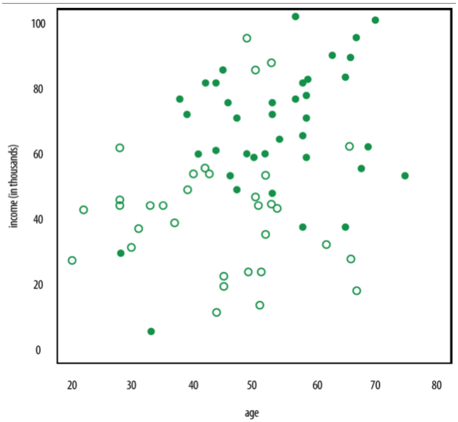
A vector locates an object (document, person, . . .) as a point in *n*-space. The angle of the vector in that space is determined by the relative weight of each term.

Feature vectors and vector space



Credit as a function of age and income

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



Vector space model for documents

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

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

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

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

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.

Term weighting

- The basic elements used in term weighting are:
  - f_d, the number of terms contained in document d
  - $f_{d,t}$, the frequency of term t in document d (TF)
  - f_{ave}, the average number of terms contained in a document
  - N, the number of documents in the collection
  - f_t, the number of documents containing term t (DF)
  - F_t, the total number of occurrences of t across all documents
  - n, the number of indexed terms in the collection

These statistics are sufficient for computation of the similarity functions underlying highly effective search engines.

Intuitions behind Term Weighting

The two basic observations we wish to proceduralise in the form of term weights are:

1. terms that occur frequently in a given document have high utility:

   $w_{d,t} \propto f_{d,t}$

2. terms that occur in a wide variety of documents have low utility:

   $w_t \propto \frac{1}{f_t}$

Models which weigh up these two are referred to as **TF-IDF** (term frequency–inverse document frequency) models

The "classic" TF-IDF formulation is:

$$w_{d,t} = f_{d,t} \times \log \frac{N}{f_t}$$

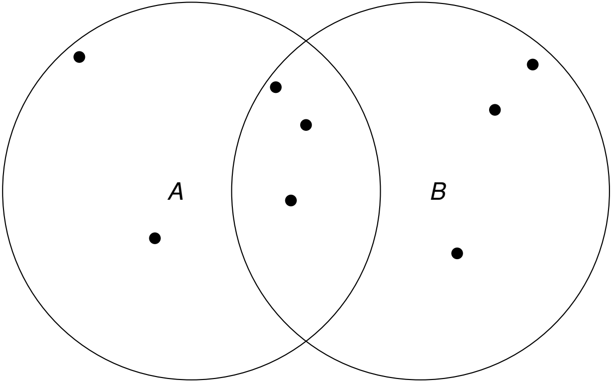
Measuring Similarity

We have discussed similarity at an intuitive level.

How do we measure similarity quantitatively?

Jaccard Similarity

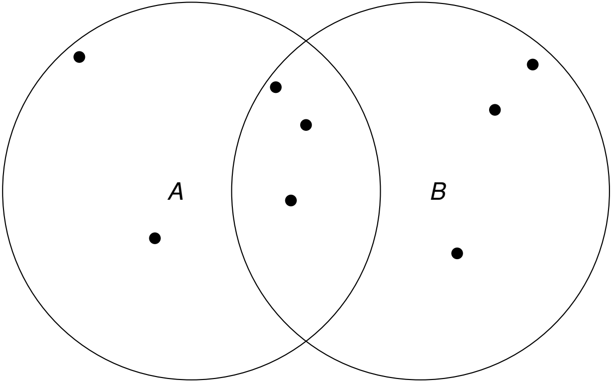
$$\frac{|A \cap B|}{|A \cup B|}$$



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

Dice Similarity

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



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

Similarity vs Distance

What is the relationship between similarity and distance?

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.

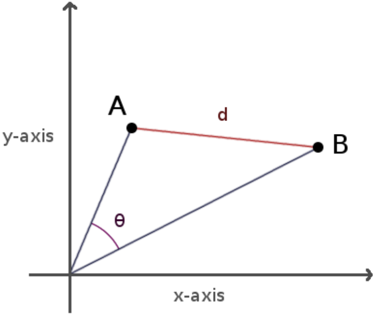
$$d(x, y) = d(y, x)$$

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

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:

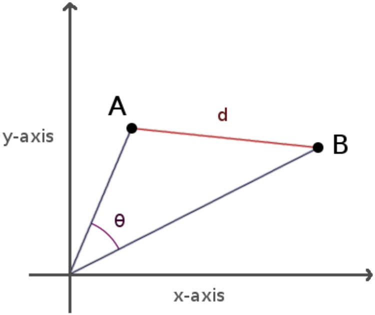


In n-dimensional space:

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

Cosine Distance

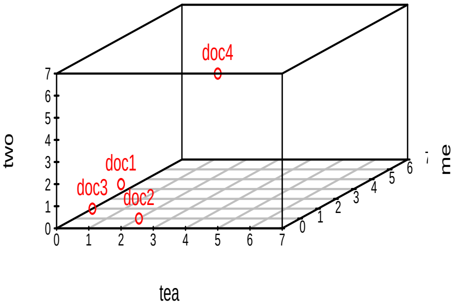
Given two items A and B, and their corresponding feature vectors $\vec{a}$ and $\vec{b}$, respectively, we can calculate their similarity via their *vector cosine* (the cosine of the angle θ between the two vectors):



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

"Long" documents & Euclidean distance

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

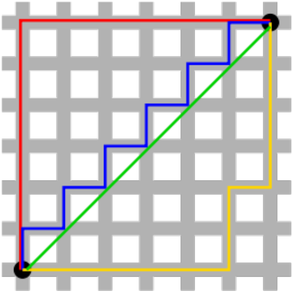


- Doc4, like Doc1, is all about "tea" and "two".
- But because it is longer, it is in a space by itself.

Manhattan Distance

[“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:



In n-dimensional space:

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

Probabilistic measures

Relative entropy:

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

or alternatively *skew divergence*:

$$s_\alpha(x, y) = D(x \mid\mid \alpha y + (1 - \alpha)x)$$

or *Jensen-Shannon divergence*:

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

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

NB: Probability will be reviewed next lecture!

Summary

- 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`