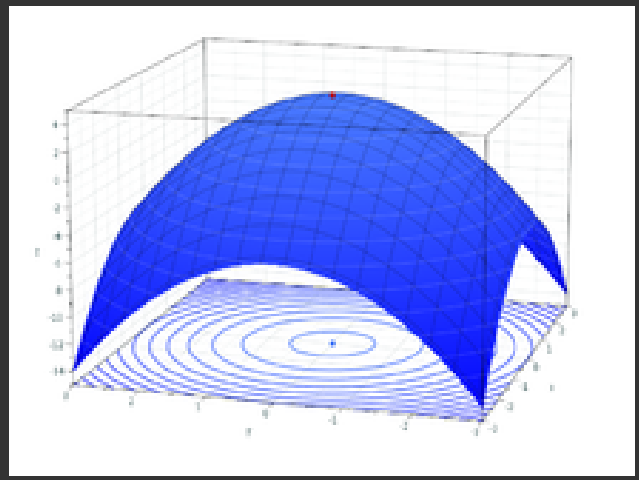
# Tutorial 3 - Vector Space Models

## COMP1046 - Maths for Computer Scientists

---

Dr. Ferrante Neri / Dr. Tony Bellotti



# Vector Space Model

A Vector Space Model is a way to represent documents numerically. Each document is coded as a vector of $n$ words. Each component of the vector is the frequency of a particular word in the document. Typically, connecting words such as "and" and "the" are ignored and "s" is removed from ends of words.

# Example

For example, consider this document,

*The best kinds of people are warm and kind,*
*They are always there and they never mind.*
*The best kinds of people smile and embrace,*
*They support you with strength and grace.*
*copyright "Alex"*

I can code it using the word list ("always", "best", "butter", "embrace", "grace", "kind", "never", "people", "smile", "strength", "support", "there", "they", "vector", "warm") as the vector

$$(1, 2, 0, 1, 1, 3, 1, 2, 1, 1, 1, 1, 1, 0, 1).$$

# Document encoding

If my domain language has $n$ words, I can express the space of all documents by the vector space $(\mathbb{R}^n, +, .)$. Some observations:-

- In reality, $n$ will be large (number of words in a language) and each document vector will be *sparse*, meaning most components will be zero.
- Using real numbers $\mathbb{R}$ allows for fractions and negative numbers, but this will be useful if we want to express comparisons between documents.

Answer the following questions about this vector space.

# Exercise 1

Consider the word list ("algebra", "cool", "cook", "learning", "linear"), so $n = 5$, and the following three documents:

- *We are learning linear algebra.*
- *Linear algebra is cool.*
- *I am learning to cook.*

*continued...*

# Exercise 1

(a) Code each document as vectors $\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}$ respectively.

(b) Compute $\mathbf{d_1} - \mathbf{d_2}$ and $\mathbf{d_3} - \mathbf{d_1}$.

(c) Are $\mathbf{d_1} - \mathbf{d_2}$ and $\mathbf{d_3} - \mathbf{d_1}$ elements in the vector space $(\mathbb{R}^n, +, .)$?

(d) If I introduce a $\mathbf{d_4}$ so that $\mathbf{d_3} - \mathbf{d_4} = \mathbf{o}$, what does this mean about these two documents?

(e) Give an example of a sentence, different from the ones above, that codes as $\mathbf{d_4}$ with the property $\mathbf{d_3} - \mathbf{d_4} = \mathbf{o}$.

(f) Compute $(\mathbf{d_1} - \mathbf{d_2}) \cdot (\mathbf{d_1} - \mathbf{d_2})$ and $(\mathbf{d_3} - \mathbf{d_1}) \cdot (\mathbf{d_3} - \mathbf{d_1})$. How could you interpret the values you have computed?

# Exercise 2

Let $U$ be the set of vectors representing documents that have the same proportion of words as a given document represented by the vector $\mathbf{d_1}$. The actual proportion may be different for different documents.

For example, coding using word list ("watch", "sea","men", "boat"), gives:-

- *the men watch the boats* codes as $(1, 0, 1, 1)$;
- *the watch men watch the boats and the men on the boats* codes as $(2, 0, 2, 2)$

so the vectors are proportional with a scalar 2.

# Exercise 2

(a) How can the condition that a document represented by vector $\mathbf{d_2}$ has the same proportion of words as $\mathbf{d_1}$ be expressed in vector notation?

(b) Give a formal definition of $U$ using set notation.

(c) Show whether or not $U$ is a vector subspace of $(\mathbb{R}^n, +, .)$.

# Exercise 3

Let $\mathbf{1_j}$ denote a vector of all zeroes except the $j$th component which is 1.
For example, with $n = 4$, $\mathbf{1_3} = (0, 0, 1, 0)$.
The vectors $\mathbf{1_1}, \cdots, \mathbf{1_n}$ form a basis for $(\mathbb{R}^n, +, .)$.

1. What scalars on the basis vectors are needed to express an arbitrary vector $(x_1, \cdots, x_n) \in \mathbb{R}^n$?
2. Suppose that the 1st and 2nd words commonly appear together in documents. It is then proposed to replace $\mathbf{1_2}$ with $\mathbf{b} = \mathbf{1_1} + \mathbf{1_2}$ in the basis. Show that this new set of vectors is still a basis, and show the scalars needed to express an arbitrary vector $(x_1, \cdots, x_n) \in \mathbb{R}^n$ in terms of the new basis.