

The background consists of several large, overlapping triangles in various colors: red, orange, yellow, green, blue, and purple. The triangles are separated by thin white lines, creating a dynamic, geometric pattern.

Embedding Space

Sadegh Jafari

- 
- The background of the slide is composed of several large, overlapping triangles in various colors: red, orange, yellow, green, blue, and purple. These triangles are arranged in a way that they fill the entire frame, with some overlapping others, creating a dynamic and colorful geometric pattern.
- What is embedding Space?

One hot embedding or encoding?

- One hot encoding does not tell anything about the semantics of the items but embedding will group co-occurring items together in the representation space. So, word embedding is a better solution than one-hot encoding as it gives better results.

example

One-hot encoding

	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0

...

...

- 
- One hot pros and cons??

TF-IDF

- TF (Term Frequency)
- IDF (Inverse Document Frequency)

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

example

[He is Walter],

[He is William],

[He isn't Peter or September]

example

[0.33, 0.33, 0.33],

[0.33, 0.33, 0.33],

[0.20, 0.20, 0.20, 0.20, 0.20]

example

“He”: $\text{Log}(3/3) = 0$,

“is”: $\text{Log}(3/2) : 0.1761$,

“or, Peter, ..”: $\log(3/1) : 0.4771$

example

[1. , 1.1761 , 1.4771 , 0. , 0. , 0. , 0. , 0.],

[1. , 1.1761 , 0. , 1.4771 , 0. , 0. , 0. , 0.],

[1. , 0. , 0. , 0. , 1.4771 , 1.4771, 1.4771 , 1.4771],

- 
- TF-IDF pros and cons??


Word2Vec?

- Word2vec is a technique for natural language processing (NLP) published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text.

- 
- Word2Vec pros and cons??

Contextualized embedding?

- Contextual embeddings assign each word a representation based on its context, thereby capturing uses of words across varied contexts and encoding knowledge that transfers across languages.

- 
- The background of the slide is composed of several large, overlapping, semi-transparent geometric shapes in various colors: red, orange, yellow, green, blue, and purple. These shapes are arranged in a way that creates a dynamic, abstract pattern. The central area of the slide is white, providing a clear space for the text.
- Contextualized embedding
pros and cons??