Word embedding

What is Word Embedding?

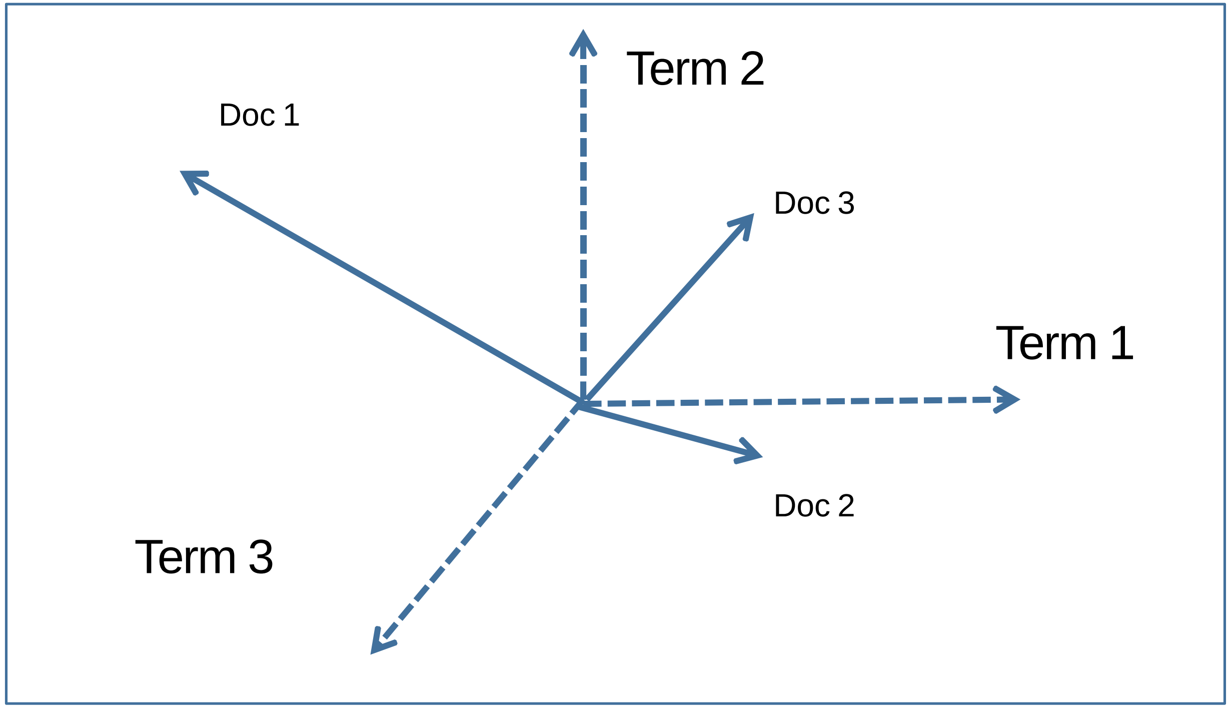
- Natural language processing (NLP) models do not work with plain text. So, a numerical representation was required.
- Word embedding is a class of techniques where word is represented as a real value vectors.
- It is a representation of word in a continuous vector space.
- It is a dense representation in a vector space.
- It can be represented in smaller dimension compared to sparse representation like one-hot encoding.
- Most of the word embedding method is based on “distributional hypothesis” by Zelling Harris.

What is word embedding? continued

- The Distributional Hypothesis is that words that occur in the same contexts tend to have similar meanings. (Harris, 1954)
- Word embeddings are designed to capture the similarity between representation like: meaning, morphology, context etc.
- The captured relationship helps us to work on downstream NLP task like chat-bot, text summarization, information retrieval etc.
- It is generated by co-occurrence matrix, dimensionality reduction and neural networks.
- It can be broadly categorized in two parts: frequency-based embeddings and prediction-based embeddings.
- The earliest work to give a vector representation was vector space model used in information retrieval task.

Vector space model

- A document was represented in a vector space.
- The dimensionality of vector space is of size of unique words in corpora.



	Term 1	Term 2	Term 3
Doc 1	0	5	5
Doc 2	2	0	1
Doc 3	3	3	0

- Hypothetical corpora with three words represented as dimension.
- Three doc projected in the vector space as per their term frequency

Vector space model continued

- The document got a numerical vector representation in a vector space represented by words.
- E.g.
  - Doc 1 -> [0, 5, 5]
  - Doc 2 -> [2, 0, 1]

- This representation is sparse in nature. Because, in real life scenario the dimensionality of a corpus shoots up to millions.
- It is based on term frequency.
- TF-IDF normalization is applied to reduce the weightage of frequent words like ‘the’, ‘are’ , and etc.
- One-hot encoding is a similar technique to represent a sentence/document in vector space.
- This representation gather limited information and fails to capture the context of the word.

Co-occurrence matrix

- It is applied to capture the neighbouring word that appeared with the word under consideration. A context window is considered to calculate co-occurrence.
- E.g.:
  - India won the match. I like the match.
  - Co-occurrence matrix for above two sentence for context window of 1.

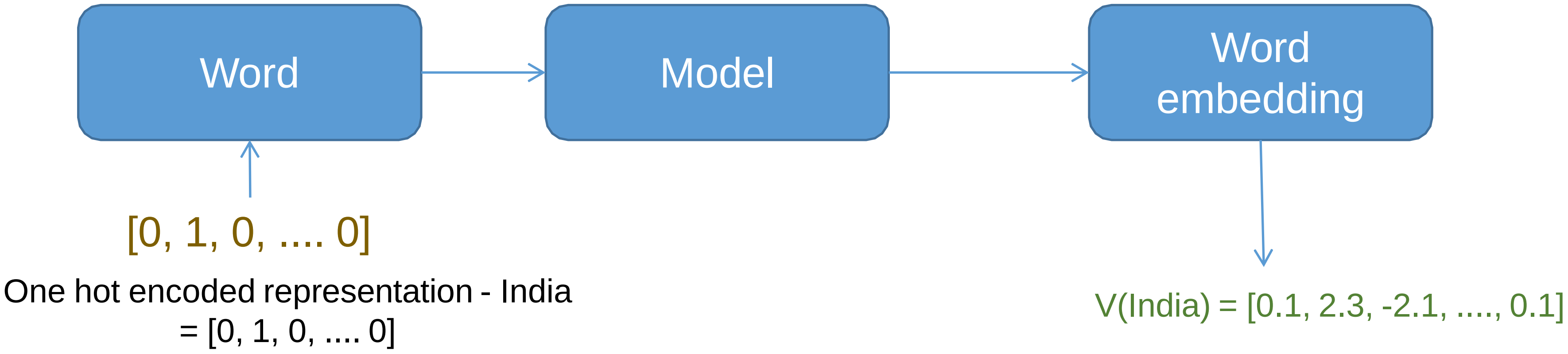
	India	won	the	match	I	like
India	1	1	0	0	0	0
won	1	1	1	0	0	0
the	0	1	1	1	0	1
match	0	0	1	1	0	0
I	0	0	0	0	1	1
like	0	0	1	0	1	1

Co-occurrence matrix continued

- Representations like One-hot encoding, Count based method and co-occurrence matrix based methods are very sparse in nature.
- Either context was limited or absent all together.
- Single representation for word in every context.
- Relation between two words like: semantic reasoning is not possible with this representation.
- Context is limited but predetermined.
- Long term dependencies are not captured.

Prediction based word embeddings

- It is a method to learn dense representation of word from a very high dimensional representation.



Word

Model

Word embedding

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

One hot encoded representation - India = [0, 1, 0, ... 0]

V(India) = [0.1, 2.3, -2.1, ..., 0.1]

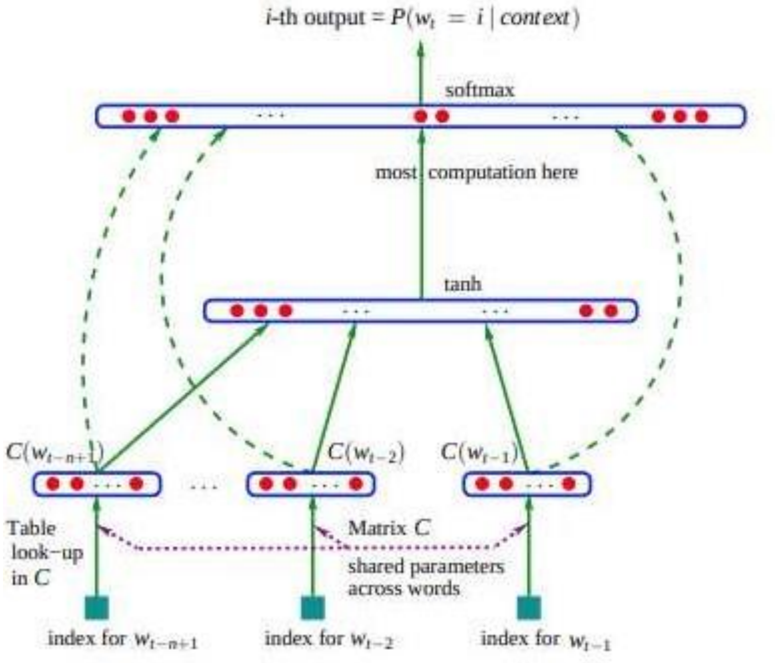
- It is a modular representation, where a sparse vector is fed to generate a dense representation

Language modelling

- Word Embedding models are very closely related to Language modelling.
- Language modelling tries to learn a probability distribution over the words in Vocabulary (V)
- Prime task of language model to calculate the probability a word W_i given the previous (N-1) words, mathematically $P(W_i|W_{i-1},\dots W_{i-n+1})$
- Probabilities over n-gram is calculated by frequency by constituent n-gram.
- In Neural network as well we achieve the same using softmax layer.
  - We calculate the log probability of W_i and normalize it with the sum of the probablities over all the words.
  - $P(W_i|W_{i-1},\dots W_{i-n+1}) = \frac{exp(h^T V'_{W_i})}{\sum_{W_{i\in V}} exp(h^T V'_{W_i})}$
  - In this case, h is the representation from hidden layer and V^i_W is the embedding of the word.
  - The inner product of $h^T V'_{W_i}$ generate the log probability of word W_i.

Classical Neural language model

- It was proposed by Bengio et al., 2003



- It consists of one layer feed-forward neural network to predict next word in sequence.
- The model tries to maximize the probability as computed by softmax.

- Bengio $L = \frac{1}{T}\sum_t \log f(w_t, w_{t-1}, \cdots, w_{t-n+1}; \theta)$ concepts
  - Embedding layer: a layer that generates word embeddings by multiplying an index vector with a word embedding matrix.

Classical Neural language model continued

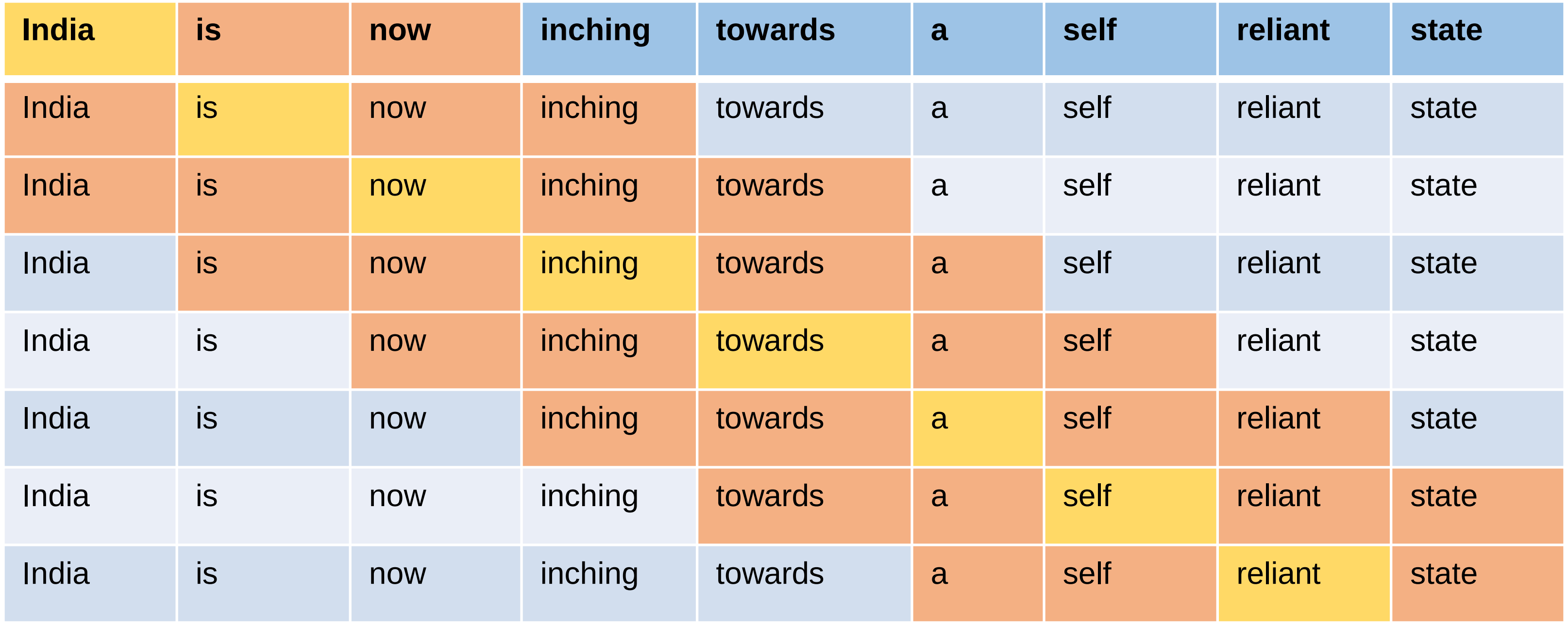
- Intermediate layers: One or more layers that produce an intermediate representation of the input, e.g. a fully-connected layer that applies a non-linearity to the concatenation of word embeddings of n previous word
- Softmax Layer: the final layer that produces a probability distribution over words in V
- Intermediate layer can be replaced with LSTM.
- The network has computational complexity bottleneck due to softmax layer, in which probability over the set of vocabulary needs to be computed.
- Neural based work embedding model made a significant progress with Word2vec model proposed by Mikolov et.al. in 2013

Word2Vec

- It was proposed by Mikolov et.al. in 2013.
- It is a two layer shallow neural network trained to learn the contextual relationship.
- It places contextually similar word near to each other.
- It is a co-occurrence based model.
- Two variants of the model was proposed
    - Continuous bag of words model (CBOW)
        - Given the context word, predict the center word
        - Order of context words are not considered, so this representation is similar to BOW.
    - Skip-gram model

What does context mean?

- Context is co-occurrence of the words. It is a sliding window around the word under the consideration.



India	**is**	**now**	**inching**	**towards**	**a**	**self**	**reliant**	**state**
India	is	now	inching	towards	a	self	reliant	state
India	is	now	inching	towards	a	self	reliant	state
India	is	now	inching	towards	a	self	reliant	state
India	is	now	inching	towards	a	self	reliant	state
India	is	now	inching	towards	a	self	reliant	state
India	is	now	inching	towards	a	self	reliant	state
India	is	now	inching	towards	a	self	reliant	state

Window size = 2, Yellow patches are words are in consideration, orange box are the context window