PyCon SriLanka

Word Embeddings for Text Classification

Nabanita Roy

Data Scientist @ ACI Worldwide | Education Lead @ Women in AI Ireland

Ireland

PyCon SriLanka

AGENDA

- Machine Learning for Texts and Natural Languages
- Transformation of Texts to Numbers
- A Note on Vector Space Model
- Traditional Text Representation
- Word Embeddings
- Introduction to Word2Vec
- CBOW vs SkipGram
- Demo

What is Machine Learning?

What is Natural Language Processing?

Where does NLP fit in the AI ecosystem?

Why Transform Texts to Numbers

Machine Learning models quantifies everything.

Therefore, numeric representation for texts is required as input an ML model.

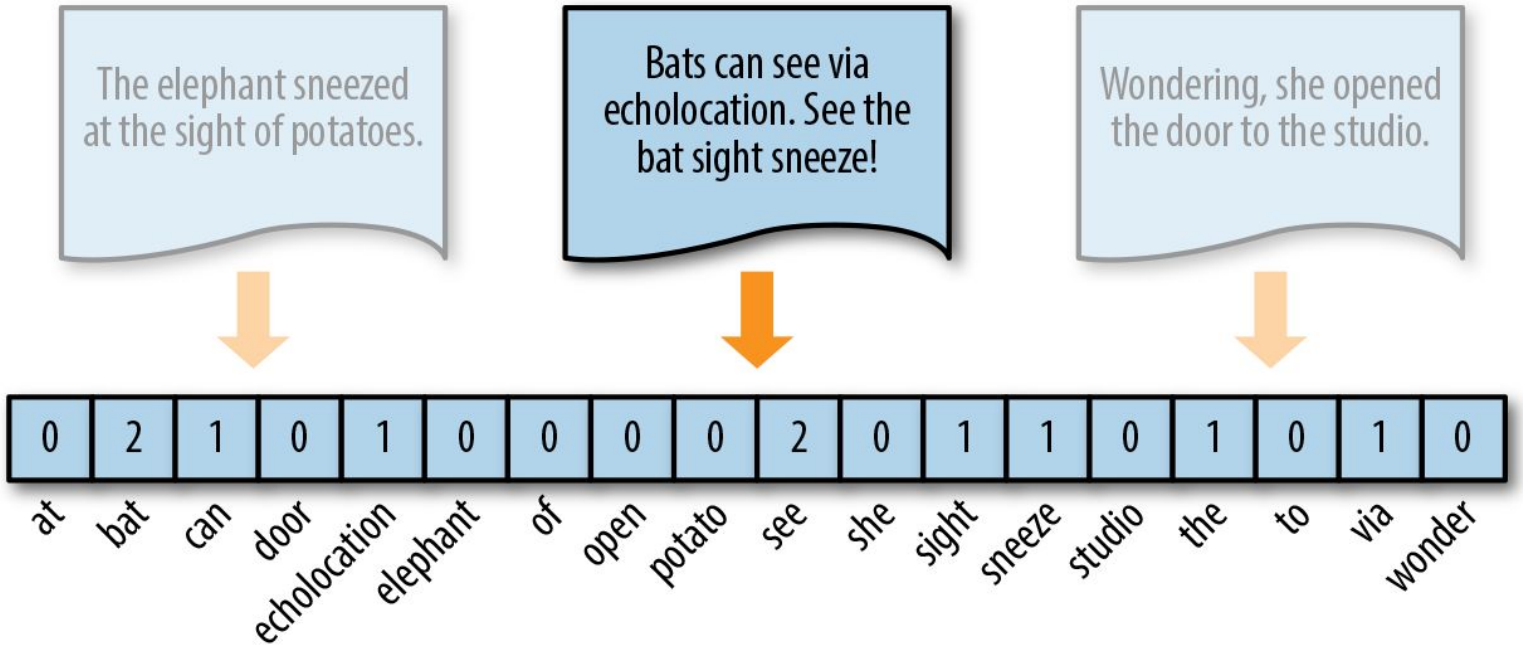


Image ref: https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-1-7ed0c7f3dfc5

A Note on Vector Space Model

- A vector is an object that has both a magnitude and a direction
- Feature vector is an n-dimensional vector of numerical features
- Similarity measures for Text Analysis – Cosine Similarity

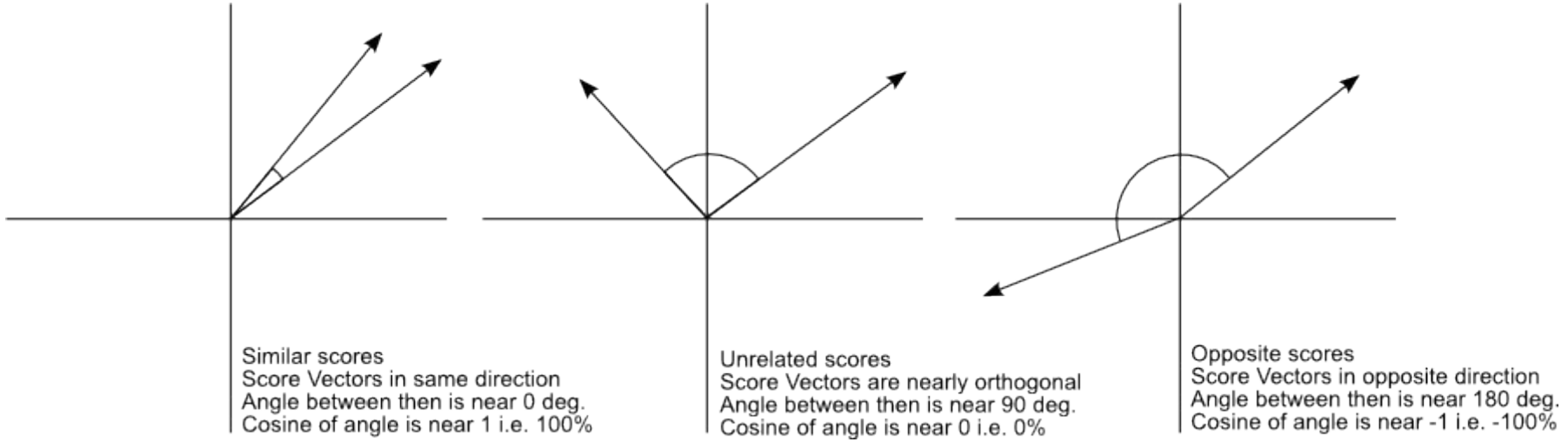
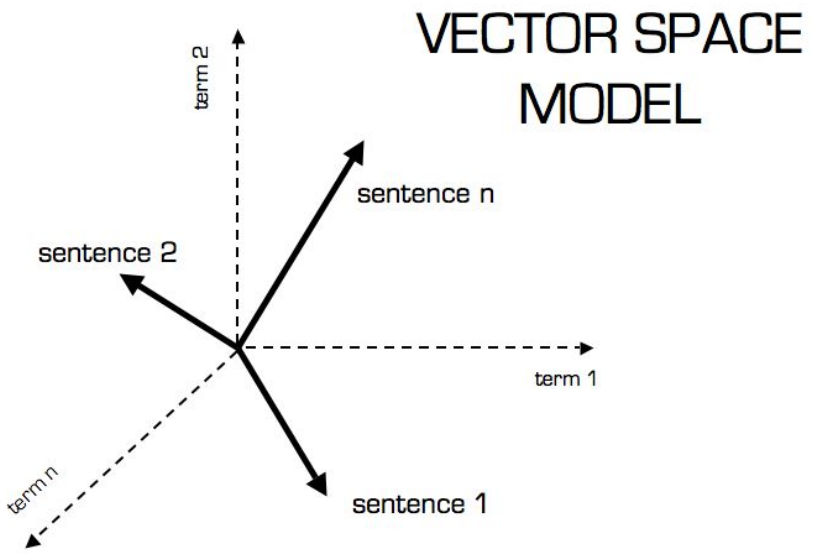


Image ref: https://github.com/sagarmk/Cosine-similarity-from-scratch-on-webpages

A Note on Vector Space Model



- Represent text as vectors in a vector space
- Similarity techniques allows us to identify the terms which occur in similar contexts
  - ✔ Euclidean Distance
  - ✔ Cosine Similarity
- The distances in vector space isn't semantic distance

Traditional Text Representation Models

"The mouse was chasing the cat"

vocab = ['the', 'was, 'mouse', 'chase', 'cat']

mouse = [0, 0, 1, 0 , 0]
cat = [0, 0, 0, 0 , 1]
chase = [0, 0, 0, 1, 0]

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

15

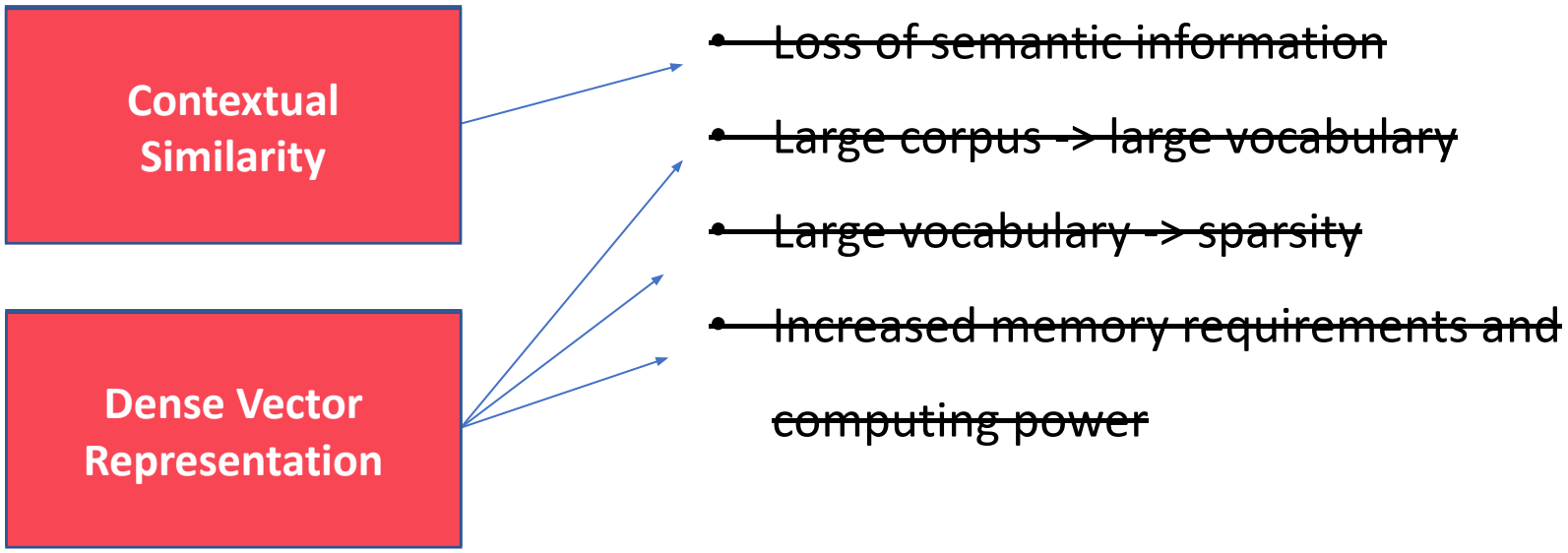
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Drawbacks

- Loss of contextual information
- Large corpus => large vocabulary
- Large vocabulary => sparsity
- Increased memory & computing power requirements

Image ref: https://web.cs.hacettepe.edu.tr/~ilyas/Courses/CMP711/lec05-TextClassificationNaiveBayes.pdf

Word Embeddings and Neural Models



Word Embeddings and Neural Models

- An embedding is a dense vector of floating-point values
- Distributional Hypothesis – Words with similar meaning occur in similar contexts

I am at **PyCon SriLanka**

I am at **PyLadies meet-up**

conference, python, programming

I am at my **home**

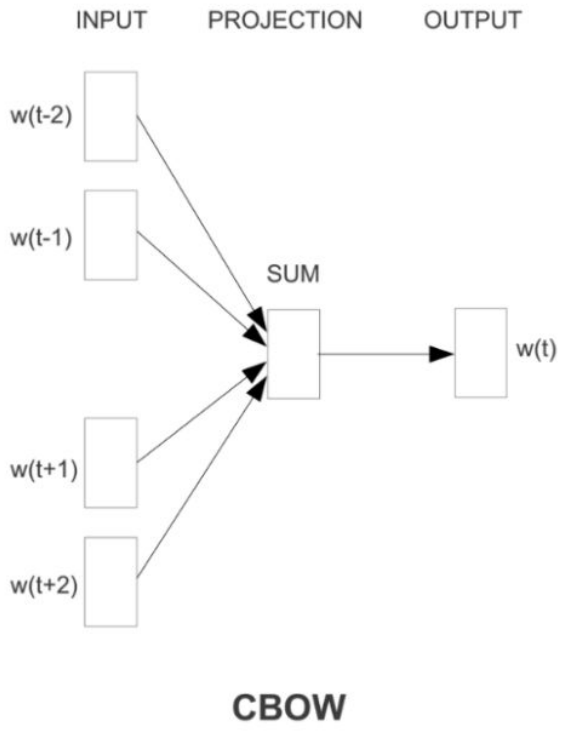
I am in my **room**

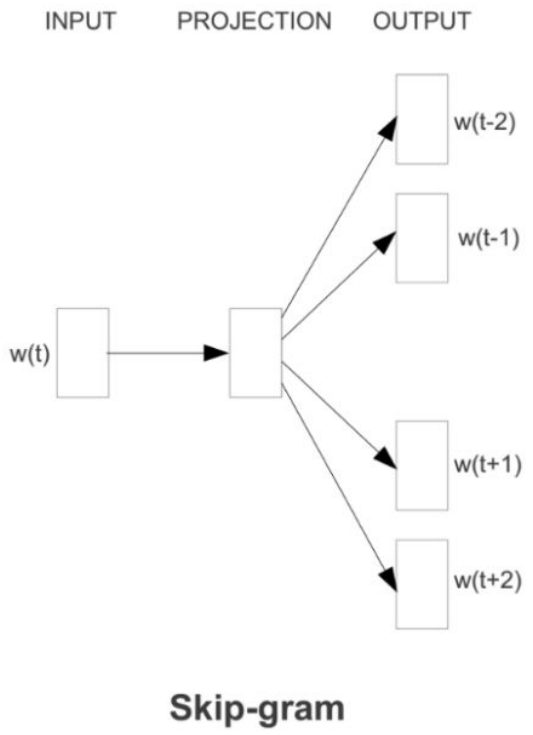
house, bed, kitchen, address

Word2Vec Model

Word2Vec is a family of model architectures and optimizations that can be used to learn word embeddings from large dataset.

predicts the middle word **based on** surrounding context words.





predict **words within a certain range before and after the current word** in the same sentence

Image ref: Efficient Estimation of Word Representations in Vector Space", 2013

How are Word Embeddings Used?

- Input to machine learning models – Embedding Layer (Keras)

- Transfer Learning –
  - Google News - https://code.google.com/archive/p/word2vec/
  - GloVe - https://nlp.stanford.edu/projects/glove/
  - FastText - https://fasttext.cc/docs/en/english-vectors.html

- Visualize and Identify patterns in corpus

Demo using Gensim