Text Representations

Women in AI Academy and Consultancy

Nabanita Roy

WOMEN AI ACADEMY

FEATURE ENGINEERING FOR TEXTS

Machine Learning models quantifies everything.

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

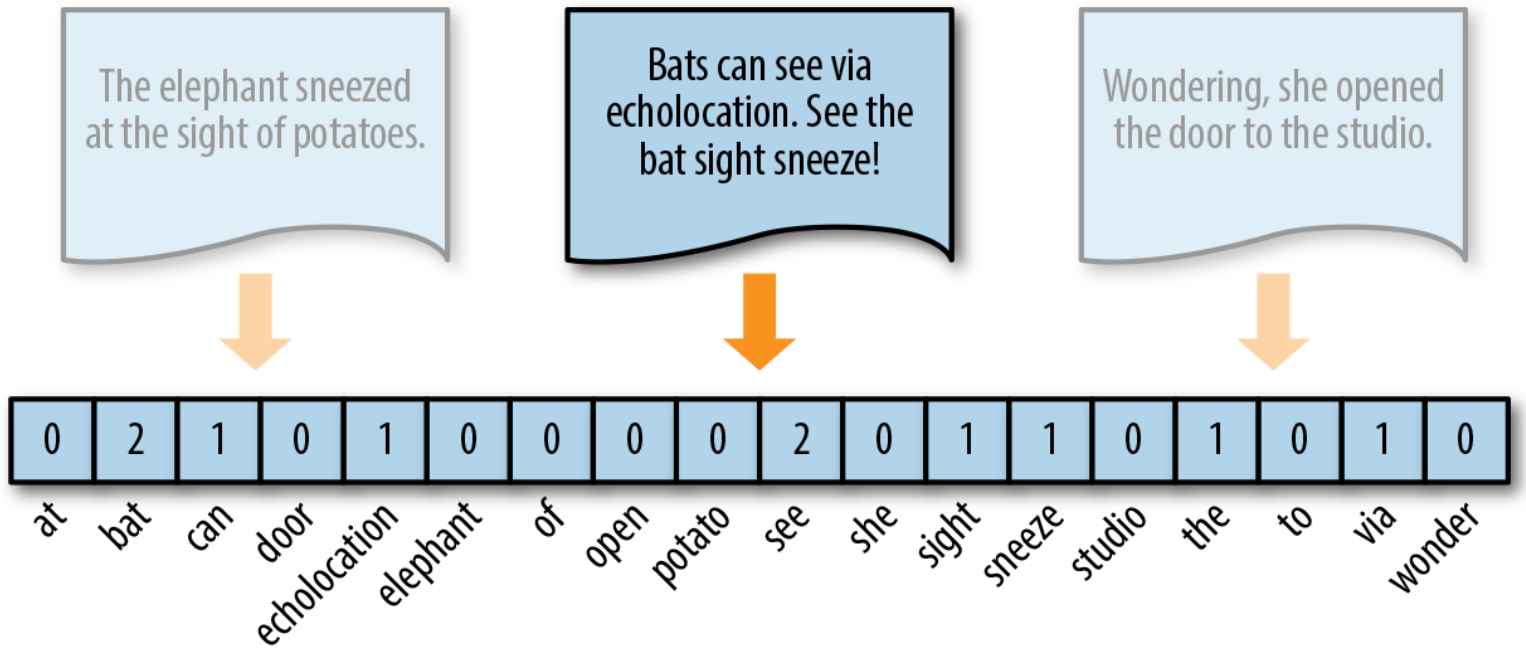


Image ref: https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-1-7ed0c7f3dfc5

WOMEN AI ACADEMY

ONE HOT REPRESENTATION

- A one hot vector is a vector whose elements are only 1 and 0. Therefore, it is a **Boolean Model**.
- For a finite set of vocabulary of size N, one **word** will be represented as a vector of N dimension where the value at the index for that **word** is 1.
- Context is lost and no frequency information

“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]

Statistical Model 2 - Bag of Words Model

- Position of the words is ignored and Frequency (the number of occurrences) of a token is considered
- In a Bag of Words or BoW, bag refers to an unordered list of words which allows multiple occurrences of the words

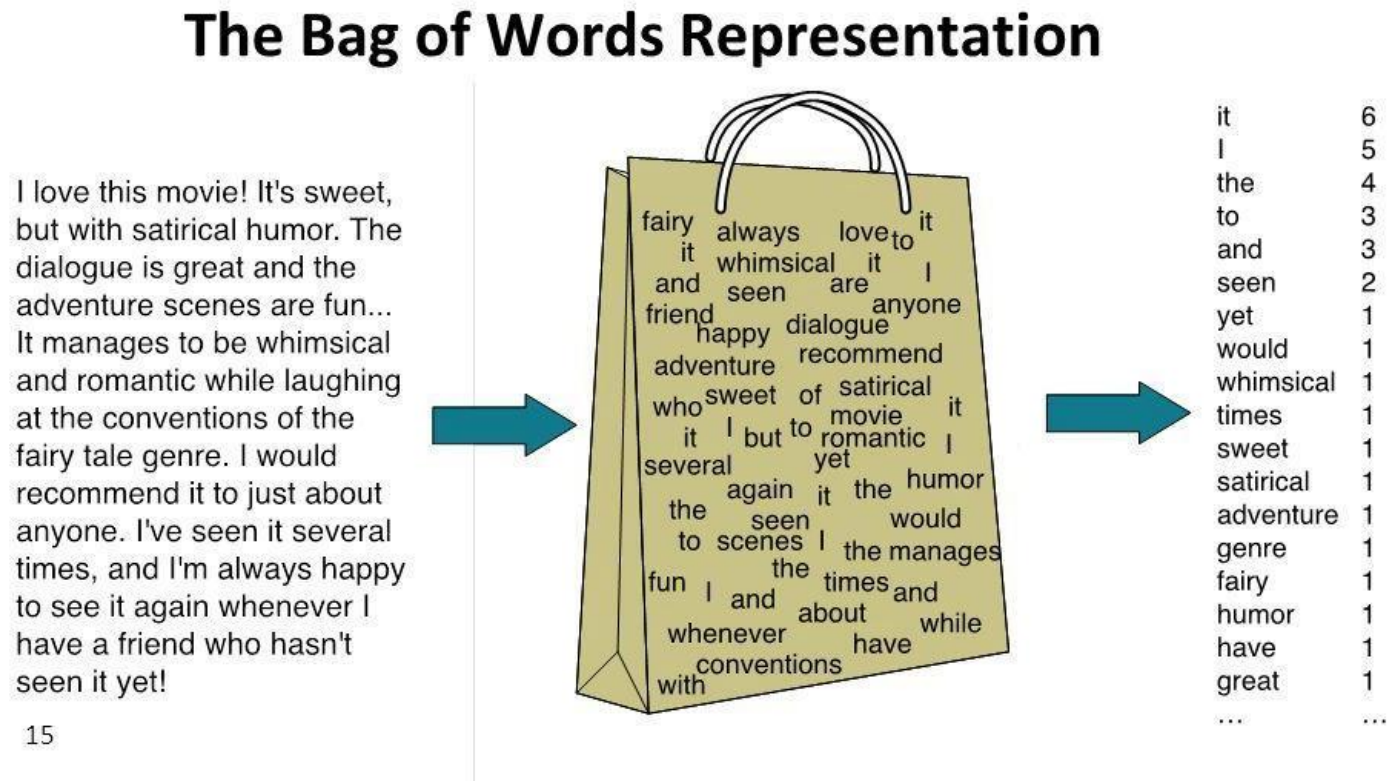
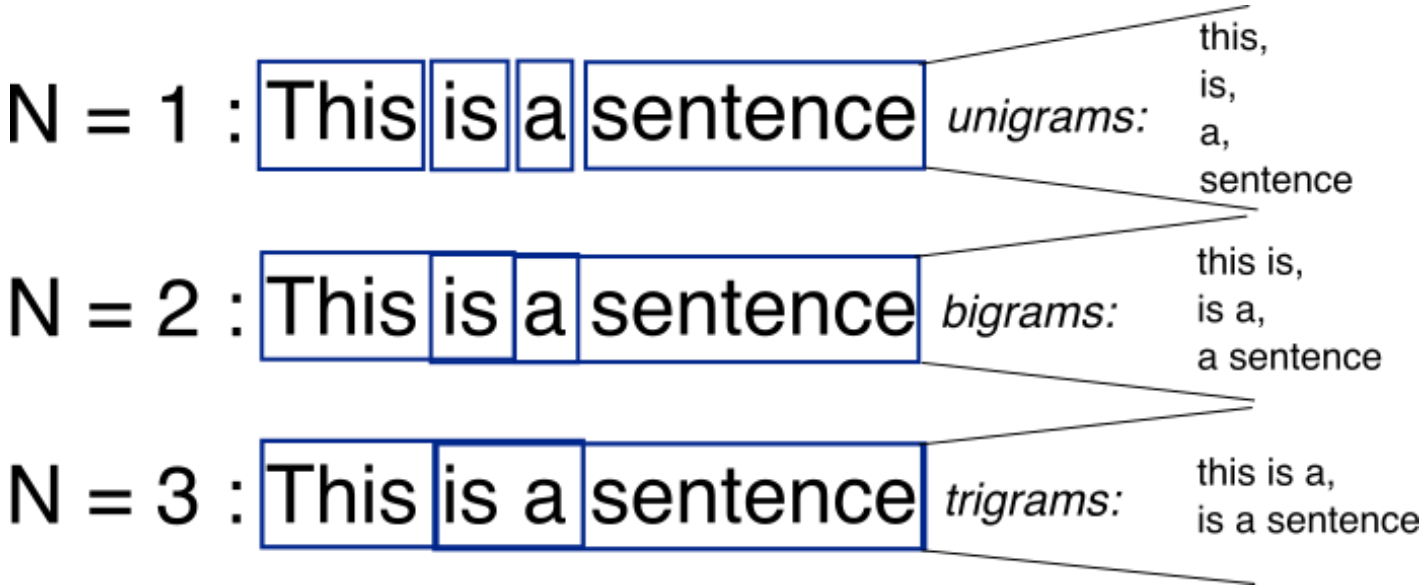


Image ref: https://web.cs.hacettepe.edu.tr/~ilyas/Courses/CMP711/lec05-TextClassificationNaiveBayes.pdf

Statistical Model II - Bag of N-Grams Model

Takes into account N tokens occurring in a sequence.



WOMEN AI ACADEMY

Bag of Words Model

PROBLEM: Highly frequent words start to dominate in the document (i.e. have larger score), but may not contain as much “informational content” to the model

TF – IDF Model

SOLUTION: Rescale the frequency of words by how often they appear in all documents, so that the scores for frequent words like “the” that are also frequent across all documents are penalized.

WOMEN AI ACADEMY

Statistical Model III

Term Frequency – Inverse Document Frequency

frequency of the word in the current document *how rare the word is across documents*

Documents are converted to vector models (or vectorized form) using the number of the times a token appears in **one** document and in **all** the documents.

Order is ignored.

Tom was chasing Jerry but **Jerry** very fast and hid in his little hole.

Tom was beaten up by **Bruno**, the dog. **Bruno** was angry and **Jerry** was hiding behind the lamp.

Tom and **Jerry** were fighting again.

WOMEN AI ACADEMY

TF-IDF Formula

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

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

Use log to dampen the effect of large corpus

(df+1) is used instead for terms that do not occur in the vocabulary to avoid Zero Division.

Image ref: http://filotechnologia.blogspot.com/2014/01/a-simple-java-class-for-tfidf-scoring.html

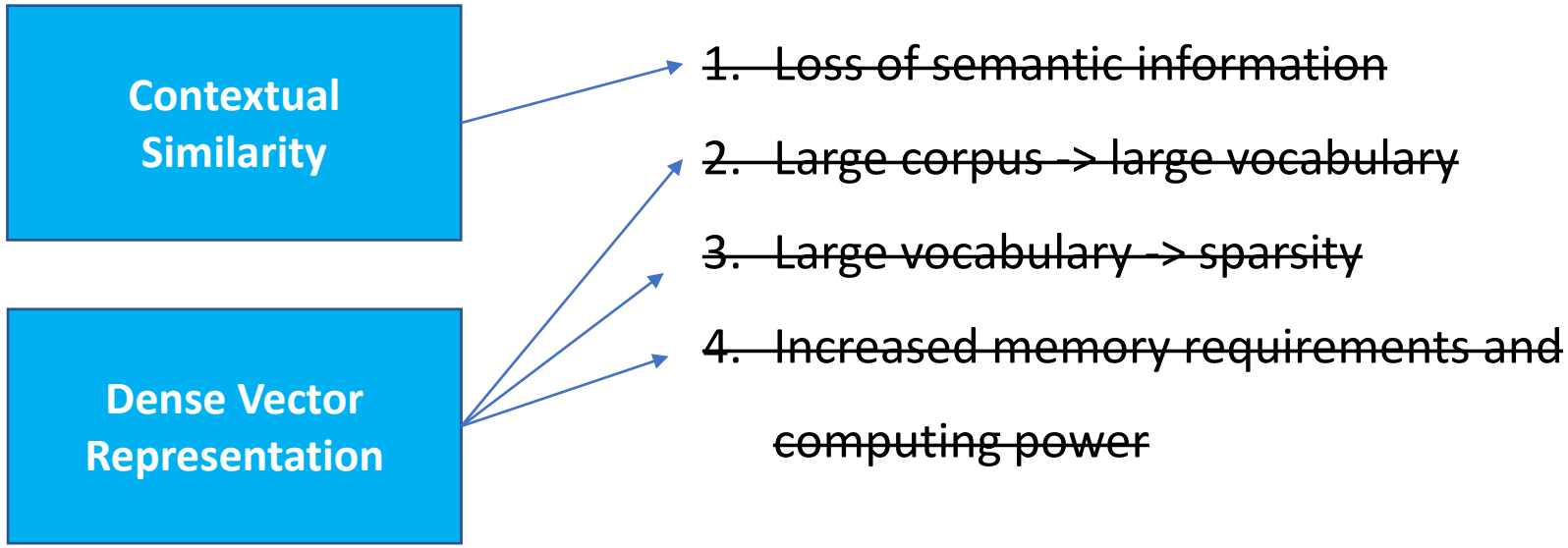
Drawback of Statistical Models

1. Loss of semantic information
2. Large corpus -> large vocabulary
3. Large vocabulary -> sparsity
4. Increased memory and computing power requirements

WOMEN AI ACADEMY

Word Embeddings and Neural Models

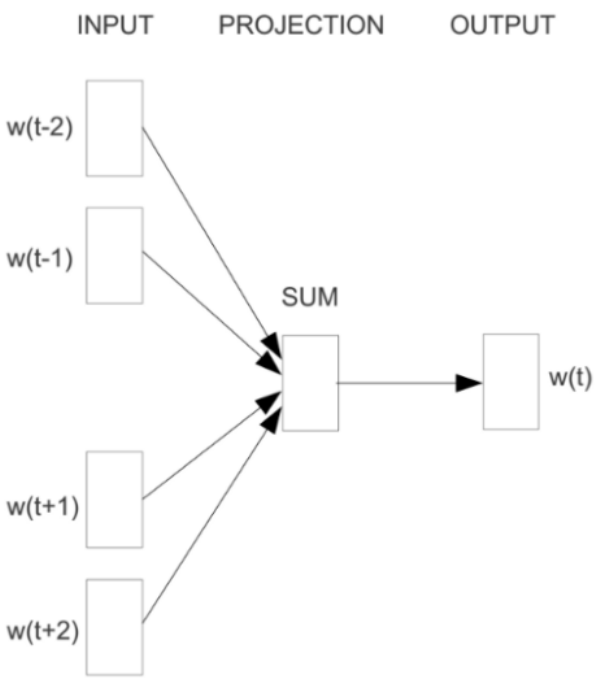
- An embedding is a dense vector of floating-point values.
- Word2Vec is a family of model architectures and optimizations that can be used to learn word embeddings from large dataset.



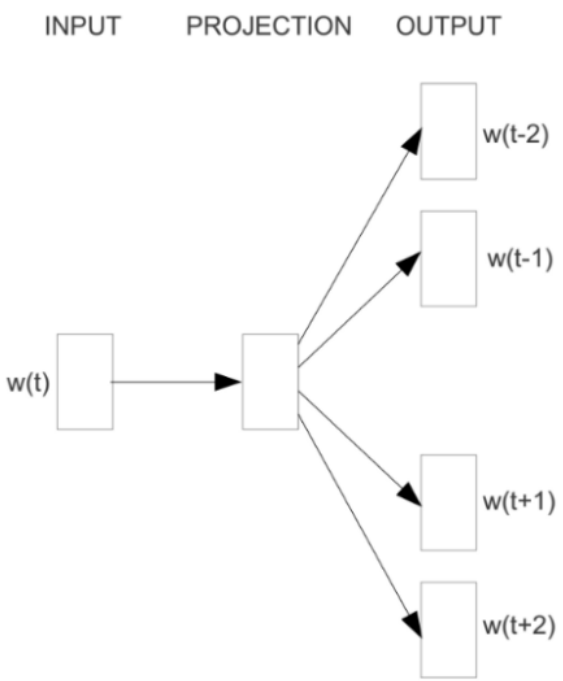
WOMEN AI ACADEMY

Word2Vec Model

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



CBOW



Skip-gram

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

WOMEN AI ACADEMY

Word2Vec

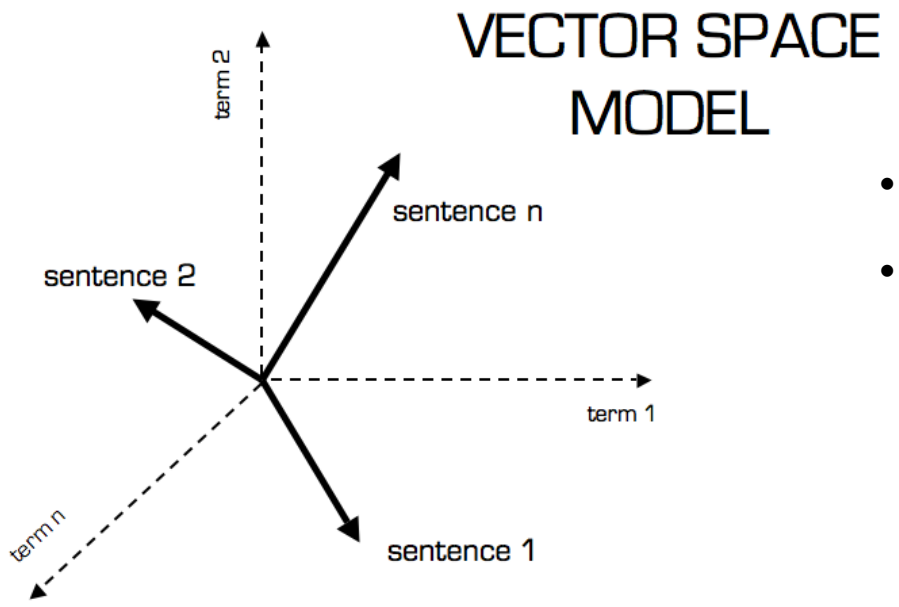
Window-based
->
Doesn't take overall corpus stats

GlobalVectors

The main idea behind the GloVe model is to focus on the co-occurrence probabilities of words within a corpus of texts in order to embed them in meaningful vectors.

WOMEN AI ACADEMY

Vector Space and Vector Similarity



- 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

Thank You
Q/A