Text Representation

Dr. Abhishek Grover
Assistant Professor
Communication and Computer Engineering, LNMIIT

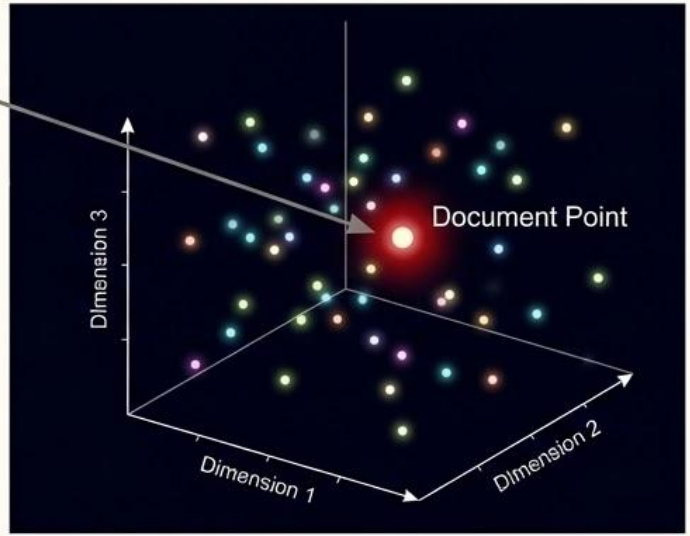
Table of Contents

The Bridge

- Computers are Numerical Engines
- Converting Text to Number
- **Vectorization**: The process of mapping words or phrases to real-valued vectors.
- Transform unstructured text into a structured numeric format
- Maintain as much syntactic (structure), semantic (meaning) and pragmatics (relationship) information as possible.

The Big Picture

- In NLP, a “Document” is just a point in a high-dimensional space.



One-Hot Encoding

- Each word in vocabulary is represented as a binary vector.
- Take a Vocabulary: Eg. [“Cat”, “Dog”, “Fish”]
- Assign a unique index to each word
- Represent each word as a vector of length N (Vocabulary size), where only the index of that word is 1 and all others are 0.
- Cat: [1,0,0]; Dog: [0,1,0]; Fish: [0,0,1]

Bag of Words (BoW) Model

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!

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
...	...

Bag of Words (BoW) Model

- BoW represents an entire document as a single vector.
- Vector length equals size of vocabulary.
- Each element in the vector corresponds to the frequency count of a specific word from the vocabulary.
- Intuition: Documents related to a specific topic will have specific words.
- Drawback: Ignores grammar and word order.

Algorithm

- Objective: To transform a collection of text documents (a Corpus) into a numerical matrix.
- Data Collection (Corpus): Gather all documents to train the model
- Preprocessing: Lowercasing, removing punctuation, characters and stopwords.
- Vectorization: For each document, create a vector of length V (number of unique words in corpus)
- Frequency counting: Iterate through each word in document and increment the count in the vector index

TF-IDF

- Normalizing word counts so as to give more weight to specific words
- Term Frequency (TF): Number of times a word appears in a document
- Inverse Document Frequency (IDF): A common word that appears in many documents should be penalized
- Effectively filters out common stop words and highlights unique meaningful words

TF-IDF

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

$$IDF(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}\right)$$

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

- The final TF-IDF score for term is stored in a vector to denote representation of a document.
- Bow Vector: [0,0,5,2]; TF-IDF Vector: [0.0,0.0,0.84,0.42]

Bag of N-Grams

- Features are contiguous sequences of n items (words) rather than just single words.
- Traditional BoW: Unigram (n=1). Eg. [“not”,”good”]
- Bag of Bigram: Eg. [“not good”]
- Bag of Trigram: Eg. [“is not good”]
- Tries to resolve issue of meaning but increases sparsity.

Co-occurrence Matrix

- Co-occurrence matrix (VxV) is for a corpus.
- Columns: Every word in vocabulary; Rows: Every word in vocabulary
- Cell value: Frequency of co-occurrence within the defined window.
- It aims to count frequency of terms that usually come together in a context window.

Comparison

BoW	TF-IDF
Raw frequency counts	Weighted importance scores
All words treated equally	Rare words given higher weightage
Requires manual stop words	Naturally penalizes common words
Basic Text classification (Spam v/s Not spam)	Information retrieval, topic identification, keyword extraction

Issues

- Sparsity
- Loss of Semantic Similarity
- Disregard for word order
- Fixed Vocabulary
- Frequency Bias

Transition to Word Embedding

- “Sparse to Dense”: Instead of vectors with 50,000 dimensions (mostly zeros), we move to “Dense Embeddings” with only 100-300 dimensions
- Eg. Word2Vec, Glove or FastText
- Words used in similar contexts needs to be mapped to nearby points in vector space.
- A vector space which allows us to do word arithmetic: King-Man+Woman=Queen)