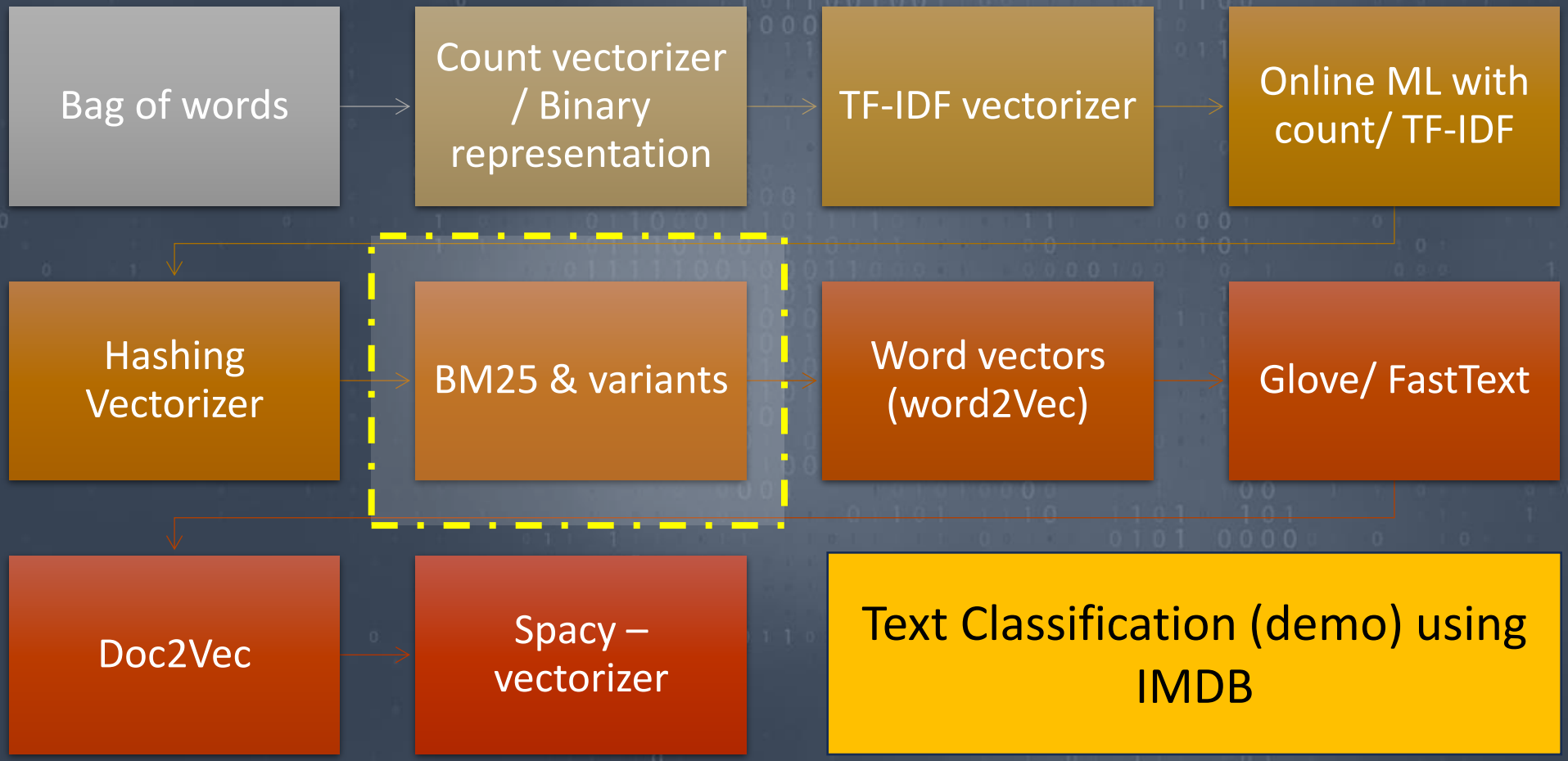
Vectorization & Embedding

"Textual Alchemy: Words to numbers"

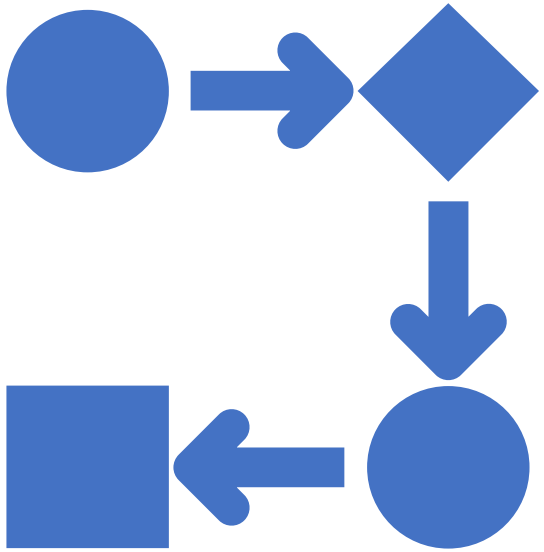
GROKKERS
AI FOR EVERYONE

Topics to cover



Vectorization

- is a transformative procedure.
- serves as the bridge between the rich, human-readable world of text and the numerical realm of machine learning algorithms.
- vectorization is the art and science of transforming the intricate patterns of language into structured numerical representations, or vectors,



Example

- Consider the following 3 documents:
  1. Document 1: "Machine learning is fascinating machine”
  2. Document 2: "Natural language processing is a key technology in AI."
  3. Document 3: "AI is shaping the future of technology."
- Tokenization
  - Document 1: ["machine", "learning", "is", "fascinating"]
  - Document 2: ["natural", "language", "processing", "is", "a", "key", "technology", "in", "ai"]
  - Document 3: ["ai", "is", "shaping", "the", "future", "of", "technology"]

Example

- Vocabulary Construction
  - Vocabulary: ["machine", "learning", "is", "fascinating", "natural", "language", "processing", "a", "key", "technology", "in", "ai", "shaping", "the", "future", "of"]

- Represent each document as a vector,
  - Document 1: [2, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
    - { represent the input doc -> vocab}
  - Document 2: [0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
  - Document 3: [1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1]

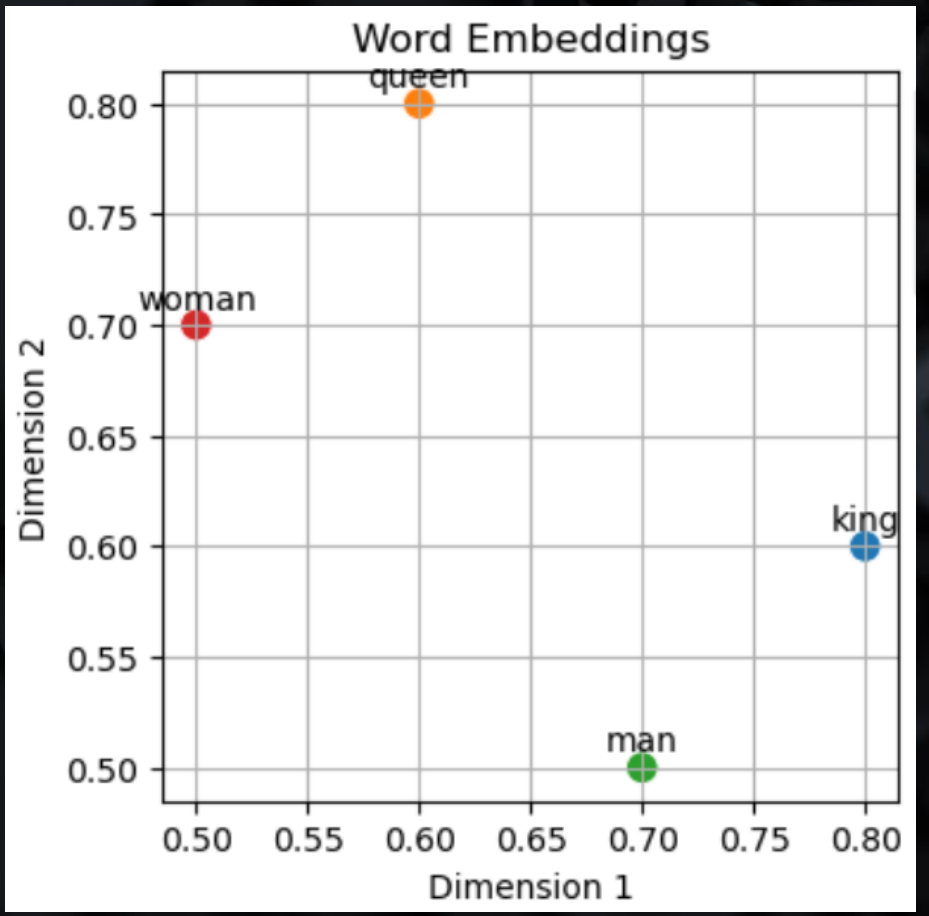
Embeddings

Embedding refers to the process of mapping words or phrases to continuous vector spaces.

capture semantic relationships between words and are often learned from large corpora using neural network models

Example

- Consider an example using a hypothetical 2-dimensional embedding space.
- words are represented as vectors with two values:
  - Word "king": [0.8, 0.6]
  - Word "queen": [0.7, 0.9]
  - Word "man": [0.6, 0.2]
  - Word "woman": [0.4, 0.7]
- words related in meaning or context are closer in the embedding space.
- For instance, "king" and “man" are closer to each other than "king" and “woman."
- Additionally, vector arithmetic can capture relationships such as "king - man + woman = queen."



comparison between vectorization and embeddings

Aspect	Vectorization	Embeddings
Definition	Converts text into numerical vectors based on word frequencies or counts.	Converts text into dense vectors that capture semantic meaning.
Dimensionality	Typically, high-dimensional and sparse.	Typically, lower-dimensional and dense.
Semantic Information	Does not capture semantic relationships or context.	Captures semantic relationships and context.
Context Awareness	Context-independent.	Context-aware.
Complexity	Simpler and easier to compute.	More complex and computationally intensive.
Example Techniques	Bag-of-Words (BoW), TF-IDF.	Word2Vec, GloVe, BERT, GPT.

Examples

Method	Example
Vectorization (BoW)	**Text:** "The cat sat on the mat."
	Vector: [1, 1, 1, 1] (where each dimension represents the presence of words like "the", "cat", "sat", "mat")
Vectorization (TF-IDF)	**Text:** "The cat sat on the mat."
	Vector: [0.5, 0.3, 0.7, 0.6] (weighted values based on frequency and importance in the document)
Embeddings (Word2Vec)	**Text:** "king"
	Vector: [0.23, -0.12, 0.56, ...] (dense vector capturing meaning; similar vectors for "queen", "man", "woman")
Embeddings (BERT)	**Text:** "bank" in "river bank"
	Vector: [0.21, -0.15, 0.34, ...]
	Text: "bank" in "financial bank"
	Vector: [0.19, -0.08, 0.40, ...] (contextual vectors representing different meanings of "bank")

GROKKERS

AI FOR EVERYONE