PTC | POLYTECHNIC COLLEGE SURINAME

UNIVERSITY OF APPLIED SCIENCES

QUALITY SYSTEM CERTIFICATION DNV ISO 9001

Word2Vec, Doc2Vec and other Word Embeddings

NLP Techniques

Tuesday, November 18, 2025

Word Embeddings Overview

- Introduction to Word2Vec, Doc2Vec and other embedding techniques.
- Word embeddings convert words into numeric vectors that capture meaning.

Word2Vec

- Predicts surrounding words (Skip-gram) or predicts the word from context (CBOW).
- Learns semantic relationships: king - man + woman ≈ queen.
- Efficient and widely used for NLP tasks.

Applying Word2Vec

Steps:

1. Prepare text corpus.
2. Train using CBOW or Skip-gram.
3. Use model.wv to get word vectors.

Applications: similarity, clustering, classification.

Doc2Vec

- Extension of Word2Vec for whole documents.
- Adds document ID vectors during training.
- Produces fixed-length vectors for sentences, paragraphs, or documents

Applying Doc2Vec

Steps:

1. Tag each document (TaggedDocument).
2. Train the Doc2Vec model.
3. Infer vectors for unseen documents.

Use cases: document similarity, classification, clustering.

GloVe

- Global Vectors for Word Representation.
- Uses word co-occurrence statistics over entire corpus.
- Captures global patterns better than Word2Vec.

FastText

- Extension of Word2Vec.
- Represents words as character n-grams.
- Handles misspellings and rare words better.
- Useful for morphologically rich languages.

WordRank

- Ranking-based embedding method.
- Uses ranking loss and focuses on improving similarity ranking.
- Often yields better performance on similarity tasks.

Varembed

- Variational Embedding method.
- Learns distributions instead of fixed vectors.
- Captures uncertainty in word meaning (polysemy).

Poincaré Embeddings

- Embeds words into hyperbolic space.
- Great for hierarchical relationships (e.g., taxonomy trees).
- Useful when relationships are not linear.