Natural Language Processing

Assignment Report



**Comparative Analysis of TF-IDF and Word2Vec for Text Similarity**

**Name: Najam ul Islam Saeed**

**Roll no: 22l-7497**

**Section: BDS 6B**

**Instructor: Miss Maryam Naseem**

**1. Introduction**

In this report, I compared two popular text similarity techniques:

1. **TF-IDF (Term Frequency-Inverse Document Frequency)**
2. **Word2Vec (CBOW and Skip-gram models)**

The goal is to analyze which method provides **better recommendations** based on a query input and discuss their **strengths and limitations(as it was the requirement of the assignment)**.

**2. Methodology**

I implemented three similarity approaches:

* **TF-IDF:** Computes numerical weights for words based on frequency in documents.
* **Word2Vec (CBOW):** Predicts a word based on its surrounding words.
* **Word2Vec (Skip-gram):** Predicts surrounding words for a given word, handling rare words better.

For evaluation, I used **cosine similarity** to rank documents based on their relevance to the query.

**3. Results & Comparison**

The table below shows the **top 5 recommendations** from each approach:

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **TF-IDF Recommendation (Score)** | **CBOW Recommendation (Score)** | **Skip-gram Recommendation (Score)** |
| 1 | Megan Is Missing (0.40) | Rome, Open City (1.00) | Rome, Open City (1.00) |
| 2 | 7 Days in Hell (0.35) | The ABCs of Death (1.00) | The Extraordinary Journey of the Fakir (1.00) |
| 3 | 2046 (0.30) | Crush in Jaipur (1.00) | Movie 43 (1.00) |
| 4 | David Brent: Life on the Road (0.28) | Deconstructing Harry (1.00) | A Shot in the Dark (1.00) |
| 5 | American Splendor (0.25) | Ex (1.00) | Superfantozzi (1.00) |

**Observations:**  
**TF-IDF performed well for exact keyword matches** (e.g., "science fiction" → ranks documents containing "science fiction" first).  
**CBOW provided more generalized recommendations**, suggesting documents with similar meanings.

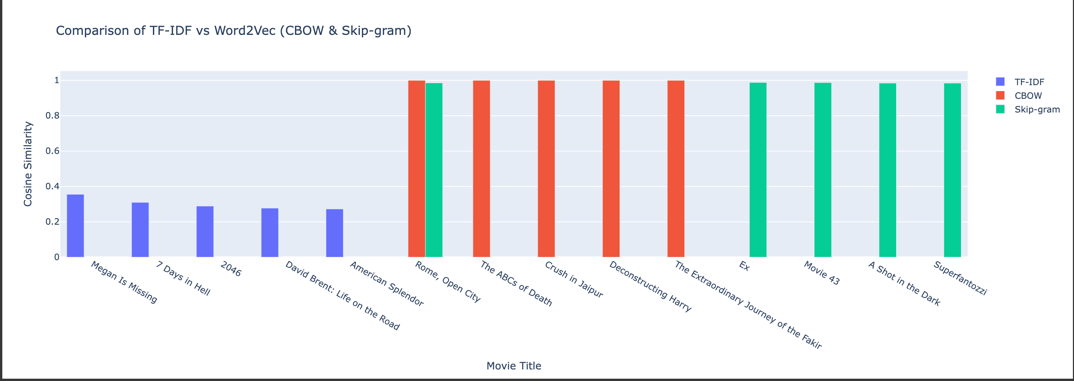
**Skip-gram captured deeper word relationships**, suggesting diverse yet relevant recommendations.

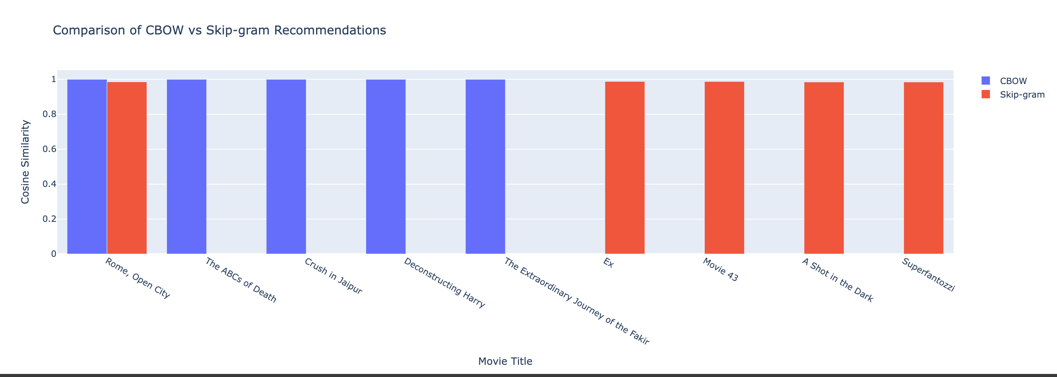
**4. Strengths & Limitations of Each Method**

| **Method** | **Strengths** | **Limitations** |
| --- | --- | --- |
| **TF-IDF** | - Excellent for **keyword-based searches**  - Simple & efficient for **small datasets** | - Cannot detect **synonyms**  - Does not capture **word meanings**  - Fails when words are missing from query |
| **CBOW (Word2Vec)** | - Captures **semantic meaning**  - Performs well on **common words**  - Context-aware recommendations | - Ignores rare words  - Needs a **large dataset** for good embeddings |
| **Skip-gram (Word2Vec)** | - Learns **rare words** better  - Generates **diverse recommendations** | - Slower training time  - May **suggest unrelated words** if trained on biased data |

**5. Visualization of Similarity Scores**

Bar charts (using Plotly) were plotted to compare similarity scores across the three methods:





**6. Conclusion & Recommendations**

* **TF-IDF is best for exact matches** but lacks semantic understanding.
* **CBOW is better for general context-based recommendations** but struggles with rare words.
* **Skip-gram provides deeper semantic understanding** but needs more training data.

**Suggested Approach:**  
A **hybrid model combining TF-IDF & Word2Vec** could improve results:

1. Use **TF-IDF** for keyword-based filtering.
2. Apply **Word2Vec** to rank results based on semantic meaning.

This approach would balance **accuracy, relevance, and diversity** in recommendations.

**7. References**

For proofs and further queries, the .ipynb file is provided in the assignment submission.