# Semantic Search in Articles Using NLP — Final Report

## Objective

This project demonstrates how to implement **semantic search** on article-like text data using a combination of **TF-IDF**, **BM25**, and **Dense Retrieval** techniques.

We compare the effectiveness of sparse (TF-IDF, BM25) and dense (transformer-based) methods in retrieving the most semantically relevant sentences to a natural-language query.

## Dataset Description

* **Source**: An excerpt from the movie *Interstellar*'s Wikipedia-style summary
* **Content**: 17 sentences about the movie, including plot, production, and release details
* **Format**: The text is preprocessed into a list of sentences (documents)

## Methods & Results

### Method 1: TF-IDF (Sparse Embedding)

* **Tool**: TfidfVectorizer from sklearn
* **Process**:
  + Used bigrams (1-2 n-grams)
  + Removed English stopwords
  + Extracted top 5 keywords per document
* **Query**: "how precise was the science ?"
* **Search**: Used cosine similarity to find most relevant sentences

**Result (Top 5 TF-IDF Matches)**:

Doc 15: It received acclaim for its performances, direction, screenplay, musical score, visual effects, ambition, themes, and emotional weight

Doc 16: It has also received praise from many astronomers for its scientific accuracy and portrayal of theoretical astrophysics

Doc 17: Interstellar was nominated for five awards at the 87th Academy Awards, winning Best Visual Effects, and received numerous other accolades

Doc 12: Interstellar premiered on October 26, 2014, in Los Angeles

Doc 14: The film had a worldwide gross over $677 million (and $773 million with subsequent re-releases), making it the tenth-highest grossing film of 2014

### Method 2: BM25 (Lexical Search)

* **Tool**: rank\_bm25
* **Process**:
  + Custom tokenizer removing punctuation and stopwords
  + Ranked documents based on BM25 scores
* **Query**: "how precise was the science ?"

**Result (Top 5 BM25 Matches)**:

Score: 1.669 It has also received praise from many astronomers for its scientific accuracy and portrayal of theoretical astrophysics

Score: 1.500 Interstellar was nominated for five awards at the 87th Academy Awards, winning Best Visual Effects, and received numerous other accolades

Score: 1.414 Interstellar uses extensive practical and miniature effects and the company Double Negative created additional digital effects

Score: 1.298 It received acclaim for its performances, direction, screenplay, musical score, visual effects, ambition, themes, and emotional weight

Score: 1.169 Brothers Christopher and Jonathan Nolan wrote the screenplay, which had its origins in a script Jonathan developed in 2007

### Method 3: Dense Retrieval (Semantic Embedding)

* **Tool**: sentence-transformers + faiss
* **Model**: all-MiniLM-L6-v2
* **Process**:
  + Sentences and queries embedded into vector space
  + FAISS used for fast nearest-neighbor search
* **Query**: "how precise was the science"

**Result (Top 5 Dense Embedding Matches)**:

1. It has also received praise from many astronomers for its scientific accuracy and portrayal of theoretical astrophysics

2. Interstellar uses extensive practical and miniature effects and the company Double Negative created additional digital effects

3. It received acclaim for its performances, direction, screenplay, musical score, visual effects, ambition, themes, and emotional weight

4. Brothers Christopher and Jonathan Nolan wrote the screenplay, which had its origins in a script Jonathan developed in 2007

5. Interstellar was nominated for five awards at the 87th Academy Awards, winning Best Visual Effects, and received numerous other accolades

## Tools & Libraries

* Python 3
* scikit-learn (TF-IDF & cosine similarity)
* rank\_bm25 (BM25 lexical search)
* sentence-transformers (Dense embedding)
* faiss (Efficient vector search)
* pandas, numpy, tqdm

## Key Insights

| **Method** | **Strength** | **Weakness** |
| --- | --- | --- |
| **TF-IDF** | Fast, interpretable keywords | Struggles with synonyms & context |
| **BM25** | Effective lexical ranking | Lacks semantic understanding |
| **Dense Embeddings** | Best semantic match | Requires more compute and model loading |

Dense retrieval consistently provides the most **contextually accurate** results, especially for queries involving **abstract meaning** (e.g., "precision of science").

## Reflection Questions

**Q1: What was the biggest challenge?**  
Integrating multiple search paradigms (TF-IDF, BM25, dense) into a consistent pipeline and ensuring fair comparison.

**Q2: What did you learn from the project?**  
That modern sentence embeddings significantly outperform traditional lexical methods in capturing **semantic nuances** in natural language.