**AI ML Internship Log**

# Day 4 - TF-IDF Feature Extraction

# Date - 12 June 2025

# Team Role - Member

# Project Title - Personality Prediction From Social Media

# **What I Did Today**

# Learned the concept of **TF-IDF (Term Frequency–Inverse Document Frequency)** and its role in feature extraction.

* Implemented TF-IDF using TfidfVectorizer from sklearn on 3 sample sentences.
* Successfully printed the TF-IDF matrix and interpreted the scores.
* Tried with and without stop\_words='english' to see the effect of removing common words.

1. **What I Learned**

* TF-IDF helps convert text into numerical vectors based on word importance.
* Rare, meaningful words are given higher scores.
* This is a **crucial step before feeding data to a machine learning model**.
* The output matrix is structured with:

Rows = Sentences

Columns = Words (in alphabetical order)

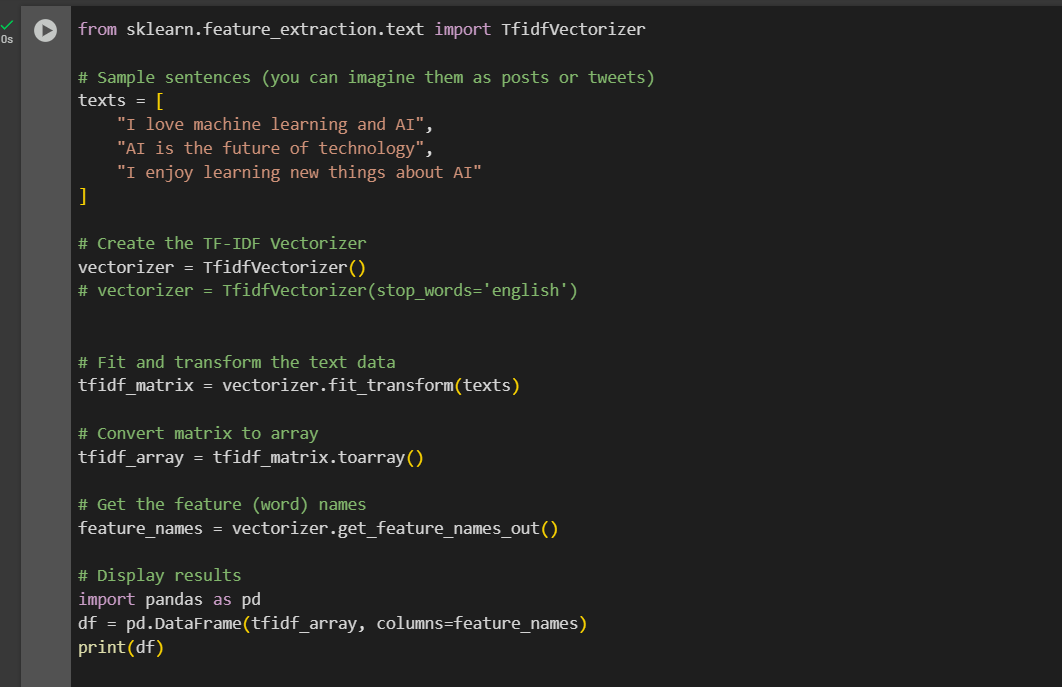
Values = Important score of the word in that sentence

1. **How I Felt**

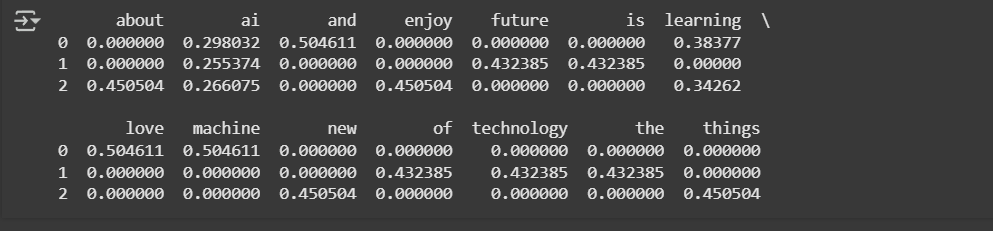
* Today was a really satisfying learning experience.
* Not only did I successfully run and tweak the TF-IDF code, but I also gained a clear understanding of **why this step matters**. I feel much more confident and technically stronger to proceed with the next stage of my project.

1. **Code Screenshot:**

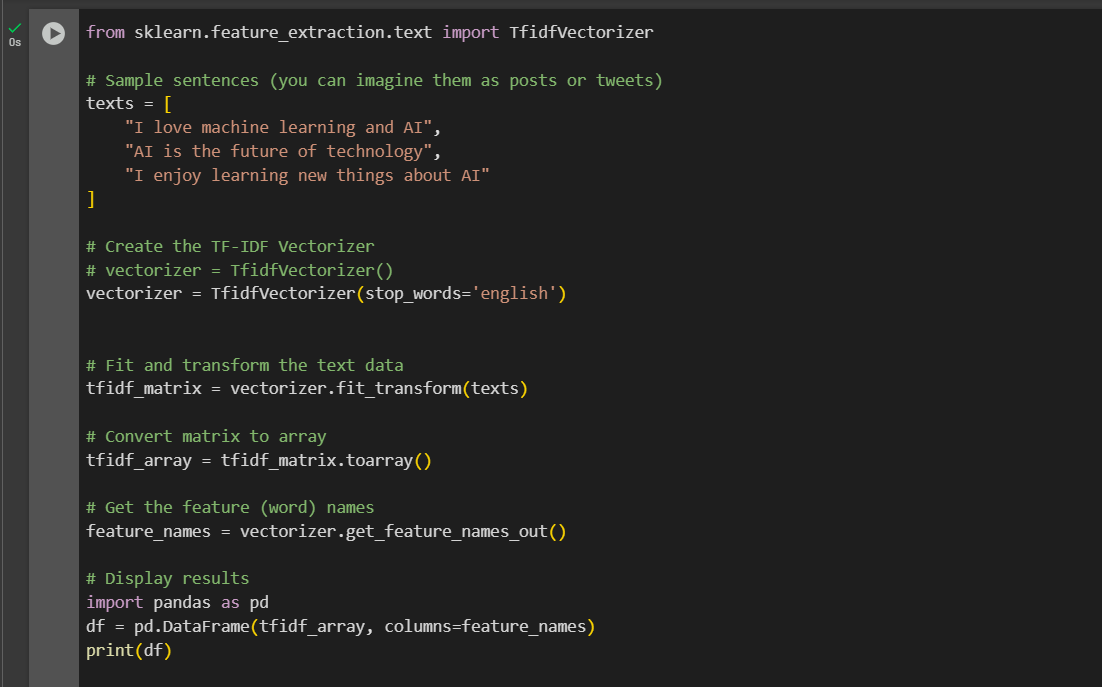
Without stop\_words=’english’



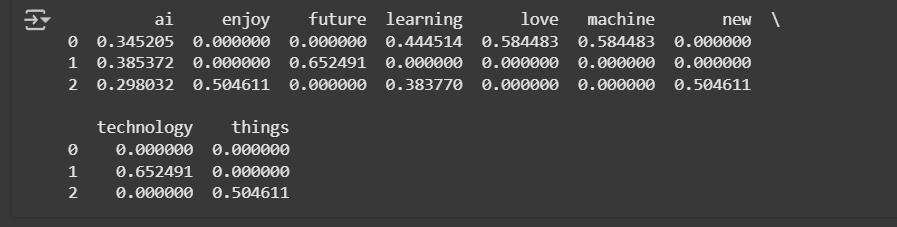
**Output:**

****

With stop\_words=’english’



**Output:**



1. **Output Observed:**

* Each sentence got converted into a vector of word importance scores.
* Words like "ai", "learning", "technology" showed up with meaningful TF-IDF values.
* Matrix shape: 3 rows (sentences) × N columns (words
* Words were **alphabetically ordered** in the output

1. **Conclusion:**

* Today’s work helped me bridge the gap between raw text and machine learning. I now understand how to convert cleaned words into numerical form using TF-IDF — a foundational step in any NLP pipeline.
* This not only improved my technical skills but also gave me confidence to move ahead with real data and model building. I feel more connected to the project and ready to take the next step in making meaningful predictions from social media content.

# 