E-commerce product recommendation system

Mathematical Modules

USED:

TF-IDF Vectorizer

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (or corpus). It helps in transforming text data into numerical data, which is useful for machine learning algorithms.

Key Concepts

1.Term Frequency (TF): Measures how frequently a word appears in a document. The more frequent a word is in a document, the higher its TF value.

$$TF = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

Inverse Document Frequency (IDF): Measures the importance of a word. Words that are common across all documents have lower IDF values. It helps to reduce the weight of common words like "the", "is", etc.

$$IDF = \text{Log}\ \frac{\text{(Total number of documents)}}{\text{(Number of documents containing the term t)}}$$

TF-IDF Score: Combines TF and IDF. A higher TF-IDF score indicates the word is important in that document but not common across documents.

$$TF\text{-}IDF = TF \times IDF$$

TFIDF Vectorizer Steps

Diagrams

1. Workflow of TF-IDF Vectorizer

Documents (Text Data)
|
v
Compute Term Frequency (TF)
|
v
Compute Inverse Document Frequency (IDF)
|
v
Calculate TF-IDF Scores
|
v
Transform Text into Numerical Vectors

Steps:

1.Compute TF: For each document, count the frequency of each term.

2.Compute IDF: For each term, calculate the inverse document frequency.

3.Compute TF-IDF: Multiply the TF by the IDF for each term in each document.

Example Products and Their TF-IDF Transformation

Assume we have the following documents:

1. **Document 1:** "I love machine learning"
2. **Document 2:** "Machine learning is fun"
3. **Document 3:** "Deep learning with TensorFlow"

For the term "machine" in Document 1:

- **TF("machine") in**
- **Document 1 = 1/4 (since there are 4 terms in Document 1)**
- **IDF("machine") = log(3/2) (since "machine" appears in 2 out of 3 documents)**
- **TF-IDF("machine") = (1/4) * log(3/2)**

Score = 90.0 (most important word)

Cosine Similarity

Cosine Similarity is a metric used to measure how similar two vectors are, regardless of their magnitude. It calculates the cosine of the angle between two non-zero vectors in a multi-dimensional space.

$$\text{Cosine Similarity} = \cos(\theta) = A \cdot B \ / \ \|A\|\|B\|$$

Assume we have the following document vectors after applying TF-IDF:

1. **Document 1 Vector:** [0.2,0.4,0.4,0.1]
2. **Document 2 Vector:** [0.1,0.3,0.4,0.4]

Calculate Dot Product:

$A \cdot B = (0.2 \times 0.1) + (0.4 \times 0.3) + (0.4 \times 0.4) + (0.1 \times 0.4)$

$0.02 + 0.12 + 0.16$

Cosine Similarity

Cosine Similarity=cos(θ)=A·B / ‖A‖‖B‖

2. **Calculate Magnitudes:**

$$\|\mathbf{A}\| = \sqrt{(0.2)^2 + (0.4)^2 + (0.4)^2 + (0.1)^2} = \sqrt{0.04 + 0.16 + 0.16 + 0.01} = \sqrt{0.37}$$

$$\|\mathbf{B}\| = \sqrt{(0.1)^2 + (0.3)^2 + (0.4)^2 + (0.4)^2} = \sqrt{0.01 + 0.09 + 0.16 + 0.16} = \sqrt{0.42}$$

3. **Calculate Cosine Similarity:**

$$\text{Cosine Similarity} = \frac{0.34}{0.61 \times 0.65} \approx \frac{0.34}{0.3965} \approx 0.86$$

Collaborative filtering

Collaborative filtering is a technique used in recommendation systems to make predictions about a user's preferences based on the preferences of many users.

User-Item Rating Matrix

	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6	Product 7	Product 8	Product 9
User 1	5	4	0	0	1	1	0	0	0
User 2	0	5	4	0	0	0	2	0	0
User 3	1	0	0	4	4	0	0	5	0
User 4	0	0	0	0	5	4	0	0	0
User 5	0	0	5	4	0	0	0	0	4

Steps:

1.Compute Similarity Between Users: Use a similarity measure like cosine similarity or Pearson correlation to find users who have similar preferences.

2.Select Neighbors: Choose a subset of similar users (neighbors).

3.Recommend Items: Recommend items that the neighbors have rated highly but the target user hasn't rated yet.

- **User 1 and User 2:**

$$\text{Cosine Similarity} = \frac{(5*0+4*5+0*4+0*0+1*0+1*0+0*2+0*0+0}{\sqrt{(5^2+4^2+1^2+1^2)} \times \sqrt{(5^2+4^2+2^2)}}$$

$$\text{Cosine Similarity} = \frac{20}{\sqrt{43} \times \sqrt{45}} \approx 0.67$$

- **User 1 and User 3:**

$$\text{Cosine Similarity} = \frac{(5*1+4*0+0*0+0*4+1*4+1*0+0*0+0*5+0}{\sqrt{(5^2+4^2+1^2+1^2)} \times \sqrt{(1^2+4^2+4^2+5^2)}}$$

$$\text{Cosine Similarity} = \frac{9}{\sqrt{43} \times \sqrt{42}} \approx 0.21$$

- **User 1 and User 4:**

$$\text{Cosine Similarity} = \frac{(5*0+4*0+0*0+0*0+1*5+1*4+0*0+0*0+0}{\sqrt{(5^2+4^2+1^2+1^2)} \times \sqrt{(5^2+4^2+2^2)}}$$

$$\text{Cosine Similarity} = \frac{9}{\sqrt{43} \times \sqrt{45}} \approx 0.2$$

- **User 1 and User 5:**

$$\text{Cosine Similarity} = \frac{(5*0+4*0+0*5+0*4+1*0+1*0+0*0+0*0+0}{\sqrt{(5^2+4^2+1^2+1^2)} \times \sqrt{(5^2+4^2+4^2)}}$$

$$\text{Cosine Similarity} = \frac{0}{\sqrt{43} \times \sqrt{57}} = 0$$

User 1 is most similar to User 2, with a similarity of 0.67.