Module 6

Non Negative Matrix Factorization

Understanding Non-Negative Matrix Factorization

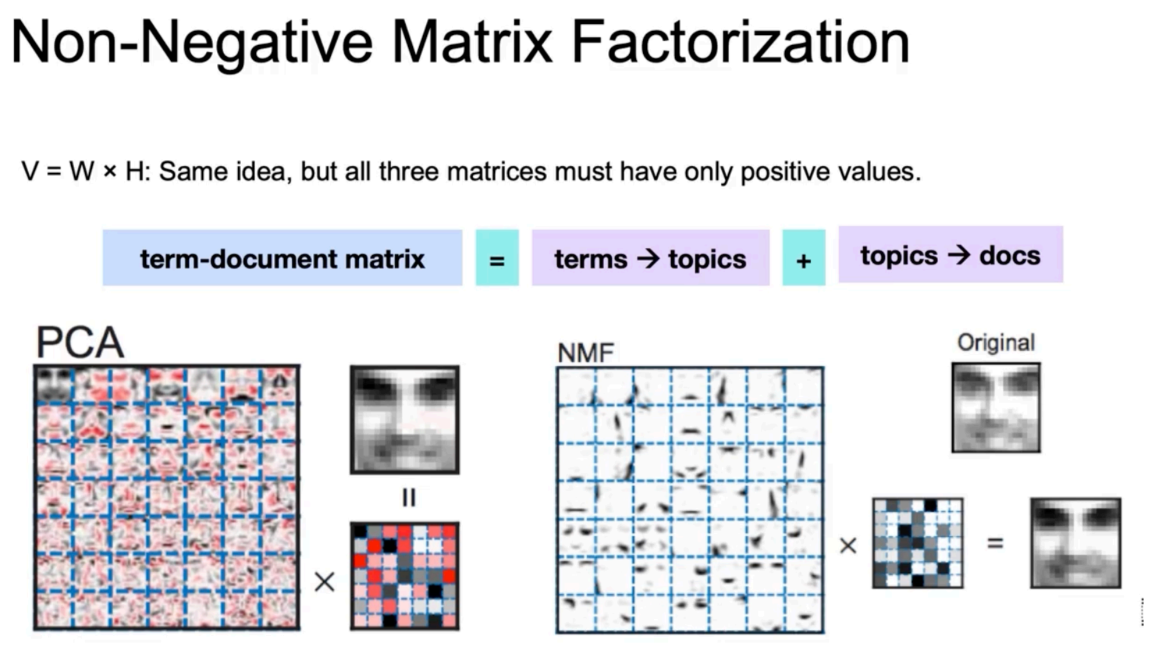
- NMF decomposes a matrix of positive values into two matrices (W and H), also containing only positive values.
- It is particularly useful for datasets like word counts or image pixels, where negative values do not exist.

Applications and Benefits of NMF

- NMF is effective in areas such as word recognition, image processing, text mining, and handling non-interpretable data like video and music.
- The additive nature of NMF allows for more intuitive interpretations of the components, as they can be seen as building blocks that combine to recreate the original data.

Comparison with Other Dimensionality Reduction Techniques

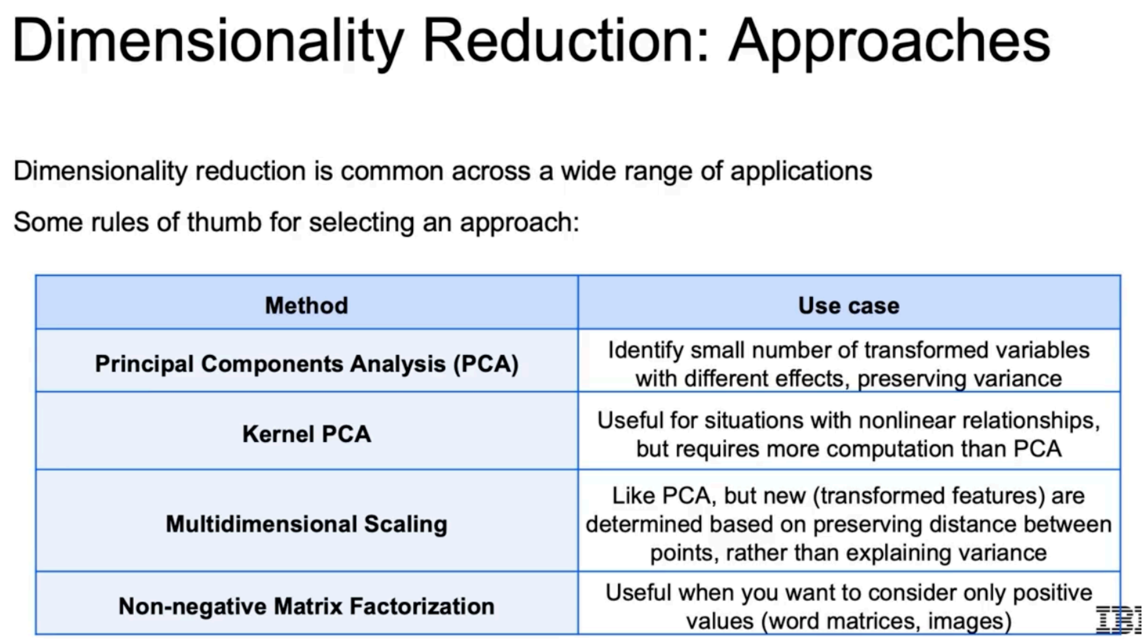
- Unlike PCA, which can involve negative values and complex combinations of features, NMF maintains a straightforward additive approach.
- NMF is sensitive to initialization, and the choice of the number of topics or components can significantly affect the results.



Non-Negative Matrix Factorization

V = W × H: Same idea, but all three matrices must have only positive values.

term-document matrix = terms → topics + topics → docs



Dimensionality Reduction: Approaches

Dimensionality reduction is common across a wide range of applications

Some rules of thumb for selecting an approach:

Method	Use case
Principal Components Analysis (PCA)	Identify small number of transformed variables with different effects, preserving variance
Kernel PCA	Useful for situations with nonlinear relationships, but requires more computation than PCA
Multidimensional Scaling	Like PCA, but new (transformed features) are determined based on preserving distance between points, rather than explaining variance
Non-negative Matrix Factorization	Useful when you want to consider only positive values (word matrices, images)

Non Negative Matrix Factorization Notebook - Part 1

Data Preparation

- The BBC dataset is pre-processed into a sparse matrix format, which minimizes memory usage by only storing non-zero values.
- The data includes `bbc.terms` (a list of words) and `bbc.docs` (a list of articles categorized by topic).

Sparse Matrix Construction

- The sparse matrix is constructed by converting a list of strings into tuples, where each tuple contains a word ID, article ID, and the count of occurrences of that word in the article.
- NumPy and Pandas are used to create the sparse matrix, which allows for efficient data handling in natural language processing tasks

Non Negative Matrix Factorization Notebook - Part 2

Understanding Non-Negative Matrix Factorization

- NMF is used to break down a document-word matrix into two matrices: one representing the relationship of words to topics and the other for reconstructing documents from these topics.

- The number of components is set to five, corresponding to the original topics in the documents.

Implementation Steps

- The NMF model is initialized with random values and applied to the sparse matrix of word frequencies.
- The output is a DataFrame that summarizes the relationship between articles and the identified topics.

Analyzing Topics and Words

- A new DataFrame is created to show how words contribute to each topic, allowing for a clearer understanding of the topics.
- The original topics (business, entertainment, politics, sports, tech) are compared to the new topics derived from the NMF process, revealing the dominant topics for each article.