Collective Intelligence

Exercise (on June 24th, 2025)

(No submission needed)

Read this document and understand how to extract topics using NMF.

This document is based on:
https://www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf/

Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization is a statistical method that helps us to reduce the dimension of the input corpora or corpora. Internally, it uses the factor analysis method to give comparatively less weightage to the words that are having less coherence.

Some Important points about NMF:

1. It belongs to the family of linear algebra algorithms that are used to identify the latent or hidden structure present in the data.

2. It is represented as a non-negative matrix.

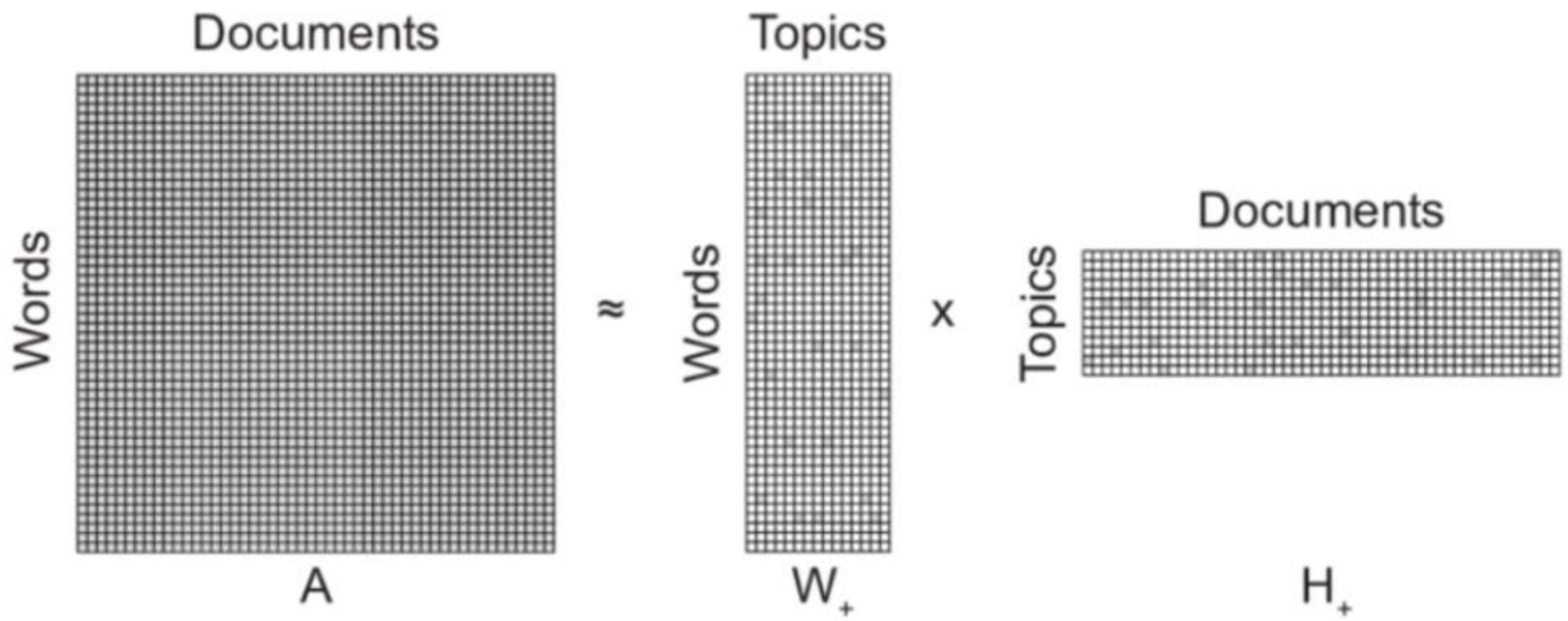
3. It can also be applied for topic modelling, where the input is the term-document matrix, typically TF-IDF normalized.

- **Input:** Term-Document matrix, number of topics.
- **Output:** Gives two non-negative matrices of the original n-words by k topics and those same k topics by the m original documents.
- In simple words, we are using linear algebra for topic modelling.

4. NMF has become so popular because of its ability to automatically extract sparse and easily interpretable factors.

Below is the pictorial representation of the above technique:

$$A \approx W \times H$$



As described in the image above, we have the term-document matrix (A) which we decompose it into two the following two matrices,

- **First matrix:** It has every topic and what terms in it,
- **Second matrix:** It has every document and what topics in it.

Other interesting application of NMF

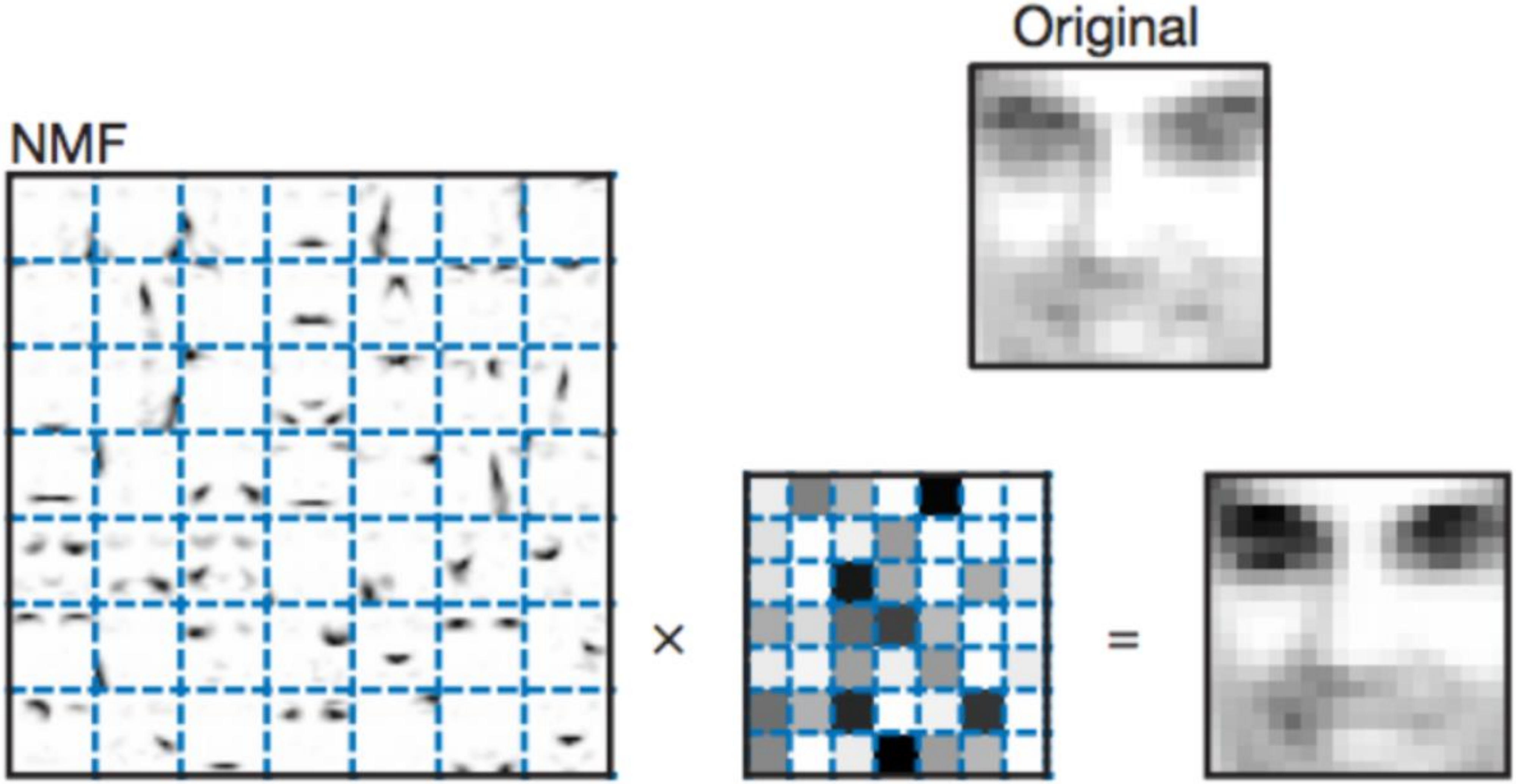


Image Processing uses the NMF. Let's look at more details about this.

Say we have a gray-scale image of a face containing *p* number of pixels and squash the data into a single vector such that the ith entry represents the value of the ith pixel. Let the rows of $X \in R^{(p \times n)}$ represent the *p* pixels, and the *n* columns each represent one image.

Now, in this application by using the NMF we will produce two matrices W and H. Now, a question may come to mind:

What exactly these matrices represent related to the given Use-Case?

Matrix W: The columns of W can be described as images or the basis images.

Matrix H: This matrix tells us how to sum up the basis images in order to reconstruct an approximation to a given face.

$$\underbrace{X(:,j)}_{j\text{th facial image}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{\text{facial features}} \underbrace{H(k,j)}_{\substack{\text{importance of features}\\ \text{in } j\text{th image}}} = \underbrace{WH(:,j)}_{\substack{\text{approximation}\\ \text{of } j\text{th image}}} .$$

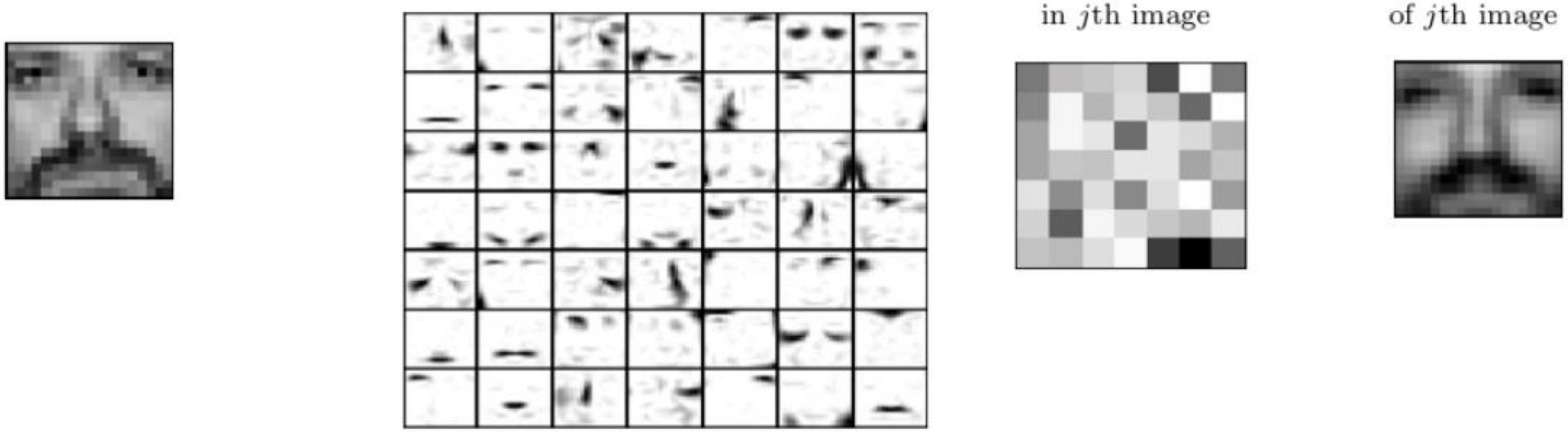


Figure 1: Decomposition of the CBCL face database, MIT Center For Biological and Computation Learning

In the case of facial images, the basis images can be the following features:

- Eyes,
- Noses,
- Mustaches,
- Lips,

And the columns of H represents which feature is present in which image.

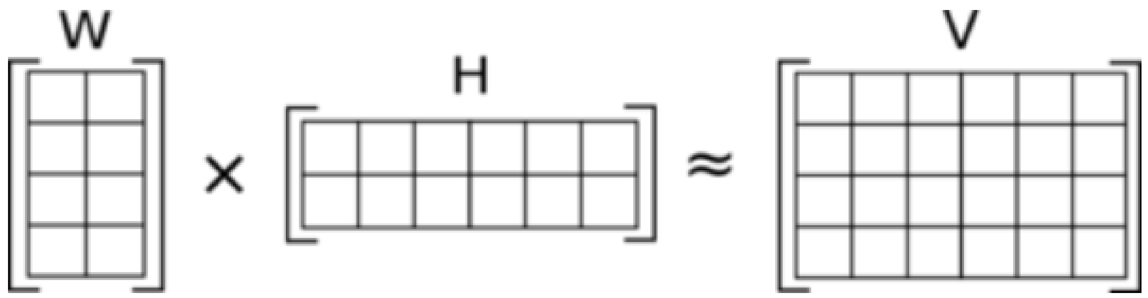
General case of NMF（Finding topics using NMF）

Let's have an input matrix V of shape m x n. This method of topic modelling factorizes the matrix V into two matrices W and H, such that the shapes of the matrix W and H are m x k and k x n respectively.

In this method, the interpretation of different matrices are as follows:

- **V matrix:** It represents the term-document matrix,
- **H matrix:** Each row of matrix H is a word embedding,
- **W matrix:** Each column of the matrix W represents the weightage of each word gets in each sentence i.e, semantic relation of words with each sentence.

But the main assumption that we have to keep in mind is that all the elements of the matrices W and H are positive given that all the entries of V are positive.



Let’s try to look at the practical application of NMF with an example described below:

Imagine we have a dataset consisting of reviews of superhero movies.

Input matrix: Here in this example, In the document term matrix we have individual documents along the rows of the matrix and each unique term along with the columns.

In case, the review consists of texts like Tony Stark, Ironman, Mark 42 among others. It may be grouped under the topic Ironman. In this method, each of the individual words in the document term matrix is taken into consideration.

While factorizing, each of the words is given a weightage based on the semantic relationship between the words. But the one with the highest weight is considered as the topic for a set of words. So this process is a weighted sum of different words present in the documents.