Classifying Documents Using Nonnegative Matrix Factorization

Brian de Silva
Department of Applied Mathematics, University of Washington

Objectives

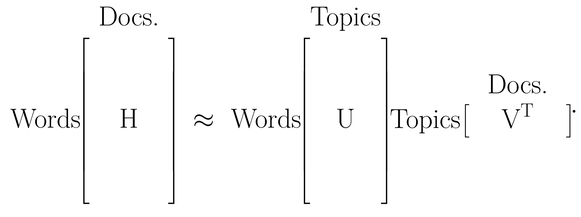
- Classify text documents based on subject matter, author, etc.
- Explore important features (topics)

Text Documents

We use the **bag-of-words** approach to model text documents as vectors with nonnegative entries corresponding to the number of times words appear in the documents. We use a tool called **term frequency-inverse document frequency** to reweight the entries of the bag-of-words vectors so they reflect the importance of a word in a document. Then we concatenate the tf-idf reweighted vectors for all the documents from a corpus as columns in one histogram matrix, H.

Nonnegative Matrix Factorization (NMF)

The **Nonnegative Matrix Factorization** of a matrix H is an approximate decomposition of H into two (often low-rank) matrices with nonnegative entries $H \approx UV^T$:



The columns of U give the "topics" present in the documents of the corpus and the columns of V^T give the coordinates of each document in this topic basis.

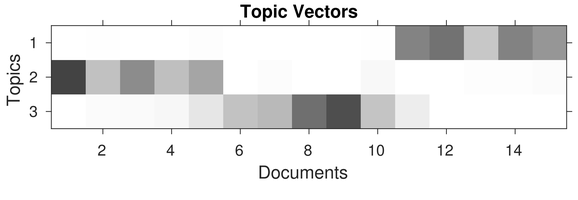


Figure 1: Topic vectors for the historical texts (perfect clustering)

Gram Matrix

For our purposes the (normalized) Gram matrix G is given by $G = \frac{1}{\|U\|_2^2} U^T U$. Here $G_{i,j}$ is the dot product of scaled versions of topics i and j hence it gives a measure of their similarity. Left-multiplying V by G modifies the topic vectors so that the weight of their entries are distributed amongst topics with similar compositions. This means we no longer need to know the number of topics to use in advance.

Algorithm Overview

1. Construct dictionary and histogram matrix H
2. Use NMF to approximately decompose H into topics and topic vectors
3. Apply Gram reweighting to topic vectors
4. Cluster the documents using their topic representations using k-means

Key Observation

Using the Gram matrix to reweight the topic vectors we incorporate the similarity between topics into the model, significantly improving performance and reducing the amount of required supervision.

Topics

Topic 1	Topic 2	Topic 3
lincoln	col	rejoined
douglas	capt	aug
political	gen	sick
slavery	rankin	enlisted
southern	lexington	fort

Table 1: Top five words in each topic (historical documents)

Topic 1	Topic 4	Topic 6
fig	acid	selection
cuticle	uric	hybrids
dragonfly	foods	varieties
cockroach	meat	sterility
wings	extract	wax

Table 2: Top five words in each topic (science documents)

Data Set

We gathered the following set of text documents from Project Gutenberg to test our algorithm:

- Historical texts
  - *History of the Seventh Ohio Volunteer Cavalry* by R. C. Rankin
  - *History of Company E of the Sixth Minnesota Regiment of Volunteer Infantry* by Alfred J. Hill
  - *Abraham Lincoln and the Union A Chronicle of the Embattled North* by Nathaniel W. Stephenson
- Scientific texts
  - *On The Origin of Species* by Charles Darwin
  - *The Life-Story of Insects* by Geo. H. Carpenter
  - *The Chemistry of Food and Nutrition* by A. W. Duncan

Every text was broken up into five equi-length text files. Some books were much longer than others. The dictionary was created directly from the corpus itself and well known "stopwords" that were likely to appear across almost all documents were removed (e.g. "the" or "a").

Clustering Performance

The k-means algorithm's clustering accuracy depended upon multiple factors, including the number of topics used, the number of clusters, the similarity measure used, the corpus, and the NMF.

Data set	Topics	Purity	Similarity
Science	2	1	Cosine
Science	3	1	Cosine
Science	10	1	Cosine
Historical	2	0.88	Cosine
Historical	3	1	Cosine
Historical	10	1	Cosine
All	2	0.73	Cosine
All	6	1	Cosine
All	10	1	Cosine

Table 3: Summary of clustering performance

Topic Vector Plots

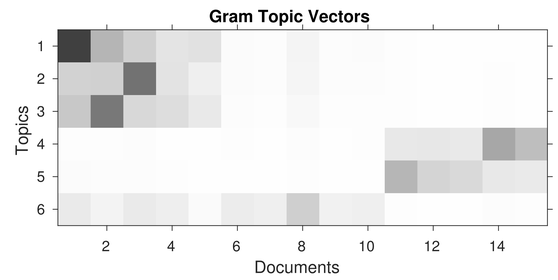


Figure 2: Gram-reweighted topic vectors for the scientific texts

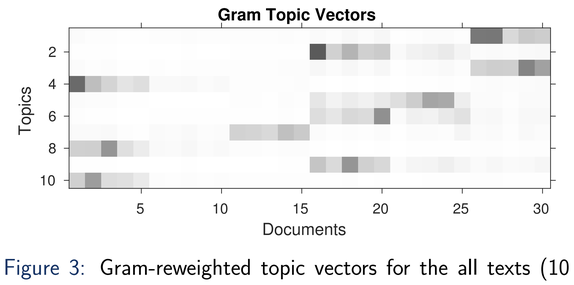


Figure 3: Gram-reweighted topic vectors for the all texts (10 topics)

Conclusions

- NMF provides a representation of documents in which they are easy to cluster and automatically provides insight into their content
- Gram matrix reweighting reduces supervision required for clustering

Acknowledgements

All of the text documents used for this project were obtained from Project Gutenberg: `https://www.gutenberg.org`. The stopwords removed from the documents were from the "Long Stopword List" at `http://www.ranks.nl/stopwords`

Contact Information

- Email: bdesilva@uw.edu