Topic Modeling using Non-negative Matrix Factorization (NMF)

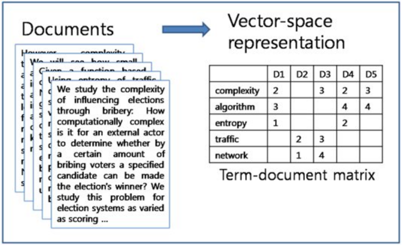
Tweet topic analysis

Praveen Gowtham

Previously in the land of NLP

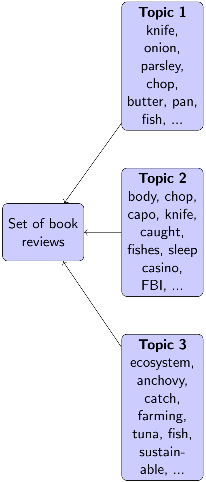
Text Processing Flow

Set of documents → Tokenize → Remove stop words → Lemmatize or stem → TF-IDF or Count Vectorizer → Term-frequency matrix (Bag of words)



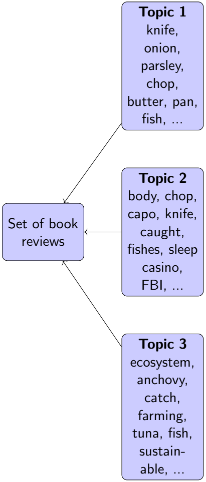
Previously: Naive Bayes to classify emails as spam or not.

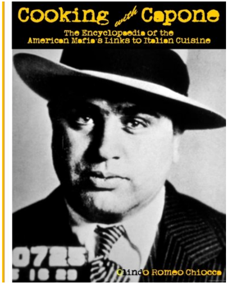
Other tasks?

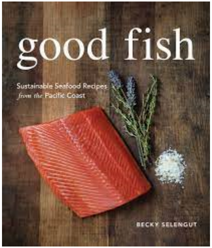


- Collection of book reviews: combinations of 3 topics.
- Certain word sets feature heavily in each topic.
- Take a book review: get combination of topics.

Other tasks?


Topic 1
knife, onion, parsley, chop, butter, pan, fish, ...
Topic 2
body, chop, capo, knife, caught, fishes, sleep casino, FBI, ...
Topic 3
ecosystem, anchovy, catch, farming, tuna, fish, sustain-able, ...
Set of book reviews


Cooking with Capone
The Encyclopaedia of the American Mafia's Links to Italian Cuisine
Olindo Romeo Chiocca


good fish
Sustainable Seafood Recipes from the Pacific Coast
BECKY SELENGUT

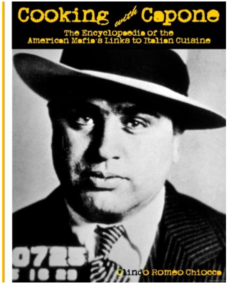
Other tasks?

Topic 1
knife, onion, parsley, chop, butter, pan, fish, ...

Topic 2
body, chop, capo, knife, caught, fishes, sleep casino, FBI, ...

Topic 3
ecosystem, anchovy, catch, farming, tuna, fish, sustain-able, ...

Set of book reviews



$$0.5\,Topic1 + 0.5\,Topic2$$

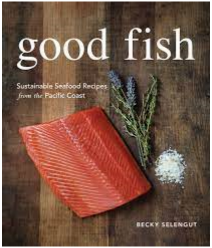
Other tasks?

Topic 1
knife, onion, parsley, chop, butter, pan, fish, ...

Topic 2
body, chop, capo, knife, caught, fishes, sleep casino, FBI, ...

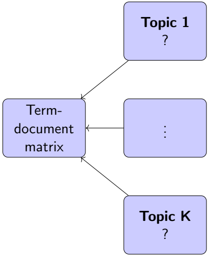
Topic 3
ecosystem, anchovy, catch, farming, tuna, fish, sustain-able, ...

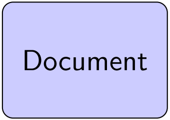
Set of book reviews



$$0.7\,Topic1 + 0.3\,Topic3$$

Topic Modeling: the problem

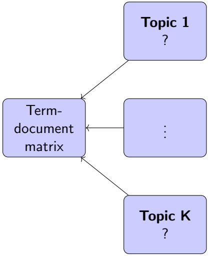


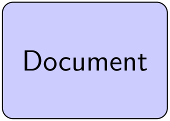


$$?Topic1 + \ldots + ?TopicK$$

- K "latent" topics.
- Unknown word distribution for topics.
- Document topic breakdown: unknown.

Topic Modeling: the problem





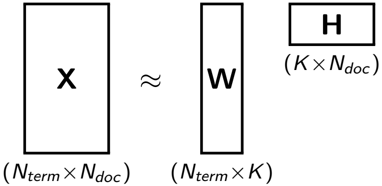
$$?Topic1 + \ldots + ?TopicK$$

- K "latent" topics.
- Unknown word distribution for topics.

- Document topic breakdown: unknown.

Goal is to learn both sides of this at the same time.

Non-negative Matrix Factorization: NMF



Definitions

X: word frequency for the documents (our BoW matrix)
W: word distribution for each topic.
H: weight of each topic in a document

Non-negative Matrix Factorization: NMF

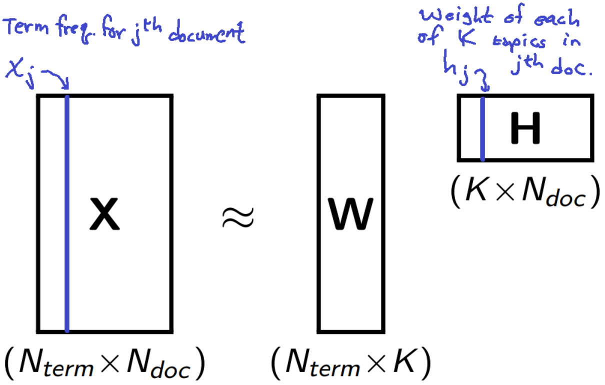
$$\underset{(N_{term} \times N_{doc})}{\boxed{\mathbf{X}}} \approx \underset{(N_{term} \times K)}{\boxed{\mathbf{W}}} \; \underset{(K \times N_{doc})}{\boxed{\mathbf{H}}}$$

Conditions

- Assumes data generated by K topics.
- **W** and **H** are non-negative matrices.

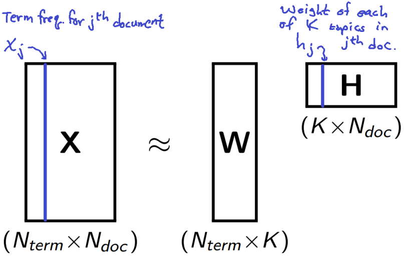
NMF: Some Intuition

Each column of **X**: term frequencies for a given document.



NMF: Some Intuition

Each column of **X**: term frequencies for a given document.



Short form:

$$x_j \approx \mathbf{W} \begin{bmatrix} h^{doc(j)}_{topic(1)} \\ \vdots \\ h^{doc(j)}_{topic(K)} \end{bmatrix}$$

NMF: Some Intuition

$$x_j \approx \begin{bmatrix} | & | & & | \\ w^{topic(1)} & w^{topic(2)} & \dots & w^{topic(K)} \\ | & | & & | \end{bmatrix} \begin{bmatrix} h^{doc(j)}_{topic(1)} \\ \vdots \\ h^{doc(j)}_{topic(K)} \end{bmatrix}$$

NMF: Some Intuition

Expanding it:

$$x_j \approx h_{topic(1)}^{document(j)} * \begin{bmatrix} w_1^{topic(1)} \\ w_2^{topic(1)} \\ \vdots \\ w_{N_{term}}^{topic(1)} \end{bmatrix} + \ldots + h_{topic(K)}^{document(j)} * \begin{bmatrix} w_1^{topic(K)} \\ w_2^{topic(K)} \\ \vdots \\ w_{N_{term}}^{topic(K)} \end{bmatrix}$$

In words:

Tries to model term frequencies for each document as weighted sum of word distributions for each topic.

Finding W and H

One possible way: minimize squared loss error for each element.

$$L = \sum_{ij} |X_{ij} - (\mathbf{WH})_{ij}|^2$$

subject to all elements of W and H ≥ 0.

Result

Minimizing loss subject to constraint:

- Often leads to topics that are interpretable. (**W** matrix)
- Topic breakdown for each document. (**H** matrix)

D.D. Lee and H.S. Seung, Nature **401**, 789 (1999)

Case Study: COVID-19 Tweet Analysis

Think tank hires you:

- Want to know trending concerns about Covid-19.
- Higher level analytics on these concerns.
  - Distribution of concerns/issues.
  - Concerns/issues that go hand-in-hand
  - Issue importance time trends.
- Take to the Twitter-verse.

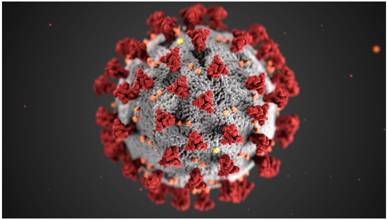


Figure: COVID-19

Starting point

Topic modeling of Covid-19 tweets using NMF.