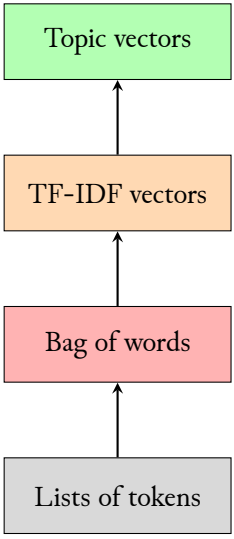
Topic Modelling

P. S. Langeslag

Models Building upon Models



Linear Discriminant Analysis (LDA)

- One-dimensional: classify as positive or negative by measuring along a single line
- Calculate the centroids of positive and negative data points and draw a line between them
- Set a threshold around the midway point for classification

Singular Value Decomposition (SVD, aka Principal Component Analysis/PCA)

SVD decomposes a matrix into three square matrices, one of which is diagonal.
(Lane et al. 111)

Singular Value Decomposition (SVD, aka Principal Component Analysis/PCA)

> SVD decomposes a matrix into three square matrices, one of which is diagonal.
> (Lane et al. 111)

- The three submatrices $U\Sigma V^T$ form the factors of the original matrix X
- Data (topics) not contributing to variance can be discarded ("truncated") from some of the submatrices for greater data separation.
- U is dimensioned for your vocabulary, and ordered by importance
- Σ is diagonal, and ordered by importance
- V^T is dimensioned for your document count
- SVD serves dimension reduction

→ See Steve Brunton's YouTube playlist for a better understanding of SVD!

Latent Semantix Analysis (or Indexing, LSA/LSI)

Uses truncated SVD to attain the greatest spread in word frequencies.

Latent Semantix Analysis (or Indexing, LSA/LSI)

Uses truncated SVD to attain the greatest spread in word frequencies.

> LSA uses SVD to find the combinations of words that are responsible, together, for the biggest variation in the data. You can rotate your TF-IDF vectors so that the new dimensions (basis vectors) of your rotated vectors all align with these maximum variance directions. The "basis vectors" are the axes of your new vector space [...]. Each of your dimensions (axes) becomes a combination of word frequencies rather than a single word frequency. So you think of them as the weighted combinations of words that make up various "topics" used throughout your corpus.
> (Lane et al. 112–113)

Latent Dirichlet Allocation

Like LSA, but assumes a Dirichlet distribution of word frequencies (a common Bayesian prior).

(See Lane et al. §4.5.)

Topic Vectors vs Word Embeddings

Both are described as company-they-keep approaches; but:

Topic Vectors

- Use SVD on BoW/TF-IDF models to infer topics
- Hence cannot rely on token context; they compare terms across documents

Word Embeddings

- Use context tokens to predict the target word, or vice versa (see next week)

Topic Vectors vs Word Embeddings

Both are described as company-they-keep approaches; but:

Topic Vectors

- Use SVD on BoW/TF-IDF models to infer topics
- Hence cannot rely on token context; they compare terms across documents

Word Embeddings

- Use context tokens to predict the target word, or vice versa (see next week)

The two concepts have their overlaps, so e.g. Jurafsky and Martin treat them in one chapter.

Bibliography

Brunton, Steve. “Singular Value Decomposition,” 2020–2022. https://www.youtube.com/playlist?list=PLMrJAkhIeNNSVjnsviglFoY2nXildDCcv.

Jurafsky, Dan, and James H. Martin. *Speech and Language Processing*. 3rd ed. draft., 2021. http://web.stanford.edu/~jurafsky/slp3/.

Lane, Hobson, Cole Howard, and Hannes Hapke. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*. Shelter Island, NY: Manning, 2019.