Review of Latent Semantic Analysis

CS410 Technology Review, Fall 2021

Robbie Li (NetID: robbiel2)

I. Introduction

In the world of topic modeling, there are 4 popular techniques that are widely used: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), and lda2vec (Xu, 2018). In this course (CS410), we covered pLSA and LDA in detail, but did not spend time diving into LSA and how it is different from pLSA and LDA. This technology review provides a detailed overview of LSA, outlines some of its strengths and limitations, and draws comparisons to pLSA and LDA.

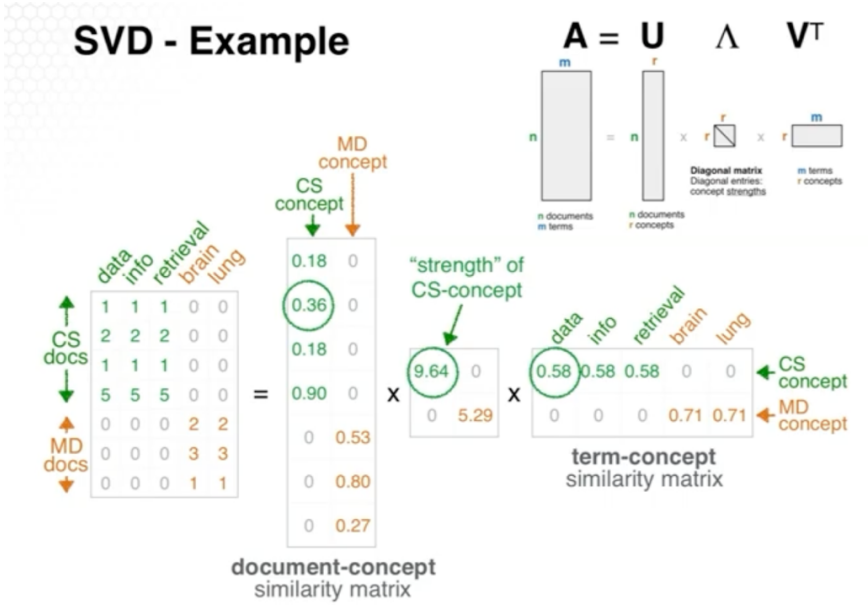
The practical use of Latent Semantic Analysis (also known as "Latent Semantic Indexing") was introduced into the public consciousness as a tool to automate the rote-grading of essays to free up teachers' time for other aspects of instruction (Associated Press, 2005). As with pLSA and LDA, the "Latent" part of its name refers to the hidden variables that help explain or categorize a document, which the technique aims to extract from a corpus in the form of topics.

II. Methodology

Both pLSA and LDA approach the problem of topic extraction from a probabilistic lens as a generative process. LSA differs is that it is powered by a generalizable tool in mathematics known as singular value decomposition (SVD). Singular value decomposition can be summarized as $A_{nm} = U_{nr} * \Lambda_{rr} * {V_{rm}}^{\mathrm{T}}$, where:

- A = a document-term matrix, *n* is # of documents, *m* is # of distinct terms
- U = a document-topic matrix, *r* is the # of concepts (topics)
- Λ = an ordered diagonal matrix of the *r* concepts
- V^T = the transpose of V, which is a term-concept matrix

SVD takes any input matrix and "decomposes" that matrix into the 3 separate matrices ($U, \Lambda, \text{and } V^T$). These matrices identify hidden variables (i.e., Λ) and explains the strength of relationship between these hidden variables to the rows of the original matrix (i.e., U) and to the columns of the original matrix (i.e., V^T). Below is a helpful illustration of SVD using a trivial example of 7 documents, 5 terms, and 2 concepts (Chau, 2020):



In this simple example, the A matrix maps out the counts of teach term for each document. In practical applications, these values are often replaced with TF-IDF scores to give weights to rarer words in the corpus. SVD provides insights into the topics through dimensionality reduction. The number of desired topics (*r*) is a parameter that is passed into the algorithm that can be fine-tuned depending on the use case. As in most matrix algorithms, the exact number of computations is numerous and involved, but this algorithm is widely available via data science libraries such as scikit-learn for Python.

III. Assessment

One interesting observation is that the underlying logic of SVD that's used in LSA can also be observed in pLSA (Xu, 2018):

$$P(D,W) = \sum_{Z} P(Z)P(D|Z)P(W|Z)$$

$$A \approx U_t S_t V_t^T$$

where $P(D,W)$ is the mathematical representation of pLSA, and A is for LSA ($S_t = \Lambda$). Another observation that helps to illustrate the value of LSA is that it works like an automatically constructed thesaurus in that it can find similar words to a search term and return relevant documents based on the shared concepts, even if the search term is not found in the documents themselves (Chau, 2020). Additionally, LSA also enables the clustering of documents based on the observed topics to help visualize popular topics in a corpus (Bhagwant, 2011).

There are a few limitations of LSA. First, it uses a bag-of-words representation of the corpus and assumes a single meaning for each word, which means any word-level ambiguity amongst the corpus (i.e., a word is used in multiple different meanings in different documents) will likely affect how strongly that word is associated to each of the discovered topics. Additionally, because it induces topics based on the text along, “None of its knowledge comes directly from perceptual information about the physical word, from instinct, or from experiential intercourse with bodily functions and feelings.” (University of Colorado at Boulder). Furthermore, LSA creates a sizeable input matrix that spans the semantic space, which makes applying SVD computationally intensive. This is mitigated by the fact that the input matrix is usually sparse, and usually only the top *k* ranking terms in each document are kept and the rest are set to 0. Lastly, LSA requires a large corpus to provide accurate results, often requiring more than 20K word types and more than 20K passages (Landauer & Dumais, 2008).

II. Conclusion

While Latent Semantic Analysis has found some practical applications, the world of topic mining has shifted towards more modern approaches that leverage probabilistic models (pLSA) and is able to incorporate prior knowledge (LDA and lda2vec). As research in generative models continues to expand due to its flexibility and use of more advanced statistical techniques, it’s unlikely that LSA will be heavily used other than as a baseline to compare other models to. However, its application of SVD to find latent topics laid important groundwork that other complex models have built on top of.

Works Cited

Associated Press. (2005, May 08). *Computers Grade Students' Writing.* Retrieved from WIRED: https://www.wired.com/2005/05/computers-grade-students-writing/

Bhagwant. (2011, August 27). *Latent Semantic Analysis (LSA) Tutorial.* Retrieved from TechnoWiki: https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

Chau, D. H. (2020, March 31). *Latent Semantic Indexing*. Retrieved from YouTube: https://youtu.be/M1duqgg8-IM

Landauer, T. K., & Dumais, S. (2008). *Latent semantic analysis.* Retrieved from Scholarpedia: http://www.scholarpedia.org/article/Latent_semantic_analysis

University of Colorado at Boulder. (n.d.). *Latent Demantic Analysis @ CU Boulder.* Retrieved from LSA @ CU Boulder: http://lsa.colorado.edu/

Xu, J. (2018, May 25). *Topic Modeling with LSA, PLSA, LDA & lda2Vec.* Retrieved from Medium: https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05