ECE 532 Final Project: Singular Value Decomposition (SVD) in Natural Language Processing (NLP)

Dayuan Jian: [jian4@wisc.edu](mailto:jian4@wisc.edu) 9075253105

Deming Kong: [dkong25@wisc.edu](mailto:dkong25@wisc.edu) 9075639790

Yuhao Liu: [liu697@wisc.edu](mailto:liu697@wisc.edu) 9077738517

1 Abstract

In class, we have seen several applications of SVD in machine learning. We have seen that SVD technique can be used to reduce noise and computing effort. SVD technique can achieve these objectives because it only takes important features and discards trivial features.

In this project, we are going to introduce the application of SVD in natural language processing (NLP). NLP is a broad topic and in this project, we are going to see how we can categorize documents using SVD.

This project will introduce the mathematical foundations in NLP. We will also talk about how to convert a document into a vector. Then, we will present how we can measure the similarities between documents. SVD technique will be presented in the activity to facilitate the understanding of the materials.

1.1 Learning Objectives

* Understand what is NLP and its importance
* Understand the basic processes and concepts of NLP
* Ability to use SVD technique to categorize documents

2 Background

In class, we have seen the applications of SVD decomposition, mainly used to reduce the effect of noise and reduce the computing effort. SVD decomposition is writing an n-by-m matrix A into the multiplication of left singular matrix, singular values, and right singular matrix. We only take the k most important singular values and singular vectors in our computation. [1] In this project, we are going to see the effect of SVD decomposition in natural language processing.

Natural language processing (NLP) is a technique that can be used to understand human languages. Generally speaking, NLP includes the following steps: processing of raw data, representing words as numbers, and representing the document by a vector.[2]

When we get our raw data, for example, a web page, we first need to remove everything we do not need, like HTML tags, encoding tags. Then, we have to figure out the boundaries of words and remove words without any meanings, like a, an, the. After this, we also need to change all characters to lowercase. Finally, every word should be changed to its original form, for example, eating/eats should be replaced by “eat”. This process is called the processing of raw data. [2]

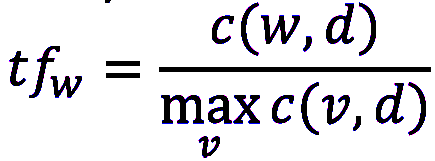
Given a file which has been pre-processed, we can represent the document by a vector:

𝑣𝑑 = [𝑐(𝑤1,𝑑), 𝑐(𝑤2,𝑑), …, 𝑐(𝑤m,𝑑)] /𝑍𝑑

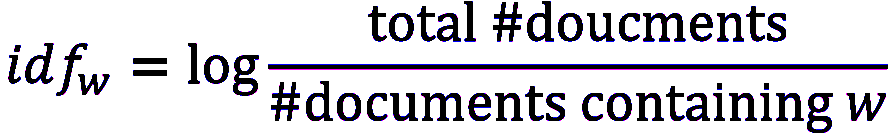
Where the document is denoted by d, the number of words w in d is denoted by c(w,d), and Zd is the total number of words in d.[2]

Another way to represent a document by a vector is to use tf-idf:

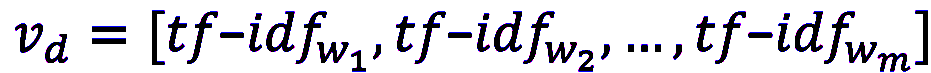
Normalized term frequency



Inverse document frequency



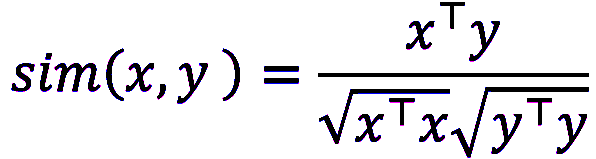
Document vector





[2]

Now, we have vectors instead of documents and we can apply our knowledge onto these vectors to get the information we want from the documents. One piece of information we are going to discuss in this project is to figure out whether two documents belong to the same category or a different one. If we take the dot product of two unit vectors, we will know the angle between these two vectors. If the same words appear again and again in two different documents, we will know that these two documents may talk about the same topic by taking the dot product of the two document vectors. So, we define the cosine of the angle to be the similarity of two documents. It can be computed using the following equation:



[2]

Since a document can contain some very common words, we can use SVD technique to avoid the influence of those words. The remaining activities are going to help you understand natural language processing and see the application of SVD in NLP.

Ethical issues:

The NLP is based on processing human language, and the human language covers every aspect of human beings, which includes ethics. It is then unavoidable to discuss whether there are ethical issues related to NLP.

People can be biased. However, NLP is a machine based process, it might easily misinterpret or be manipulated by biased or false information such as adversarial information, propaganda, fake news.[3] Moreover, it can also cause privacy issues when it comes to using NLP to analyze demographic inference or personal profiling. Lastly, when the resource is limited for NLP, its performance is still questionable. For example, the applications of NLP in disaster response, disease monitoring, and medical applications.

3 Warm-Up

**Problem 1:**

a): Please open and read the following files inside q1 folder: *doc1.txt, doc2.txt, and doc3.txt.* These three files should correspond to a topic respectively. Can you tell if any two files are related in terms of topics? Then open *SVD in NLP.ipynb* file and run through the *1 a)* cell.

b): Run the cell *1 b) build TF-IDF expressions*. Here, each column represents a given file. For instance, the first column, doc1, represents the frequency of every word listed in the corpora in doc1. Now, with the knowledge you learned in the background section, explain the meaning of entries inside this matrix. Why do some entries have value 0? What does a row of this matrix represent?

c): Run the cell *1 c) SVD*, complete the missing codes that compute the SVD of the tf\_idf matrix. If we call Matrix U the topic-term matrix, s the topic importance, and VT the topic-document matrix. In the next missing 2 lines, multiply U and VT with s. Run the codes to obtain the results and answer the following questions:

i) Which topic is doc1, doc2, doc3 in?

ii) Why does each doc only pertain to 1 topic? Refer back to the tf-idf matrix for answers.

4 Main Activity

**Problem 2:**

Problem 2 uses 3 new files in folder q2. Complete the code that calculates the cosine similarity between two vectors. Refer to the background section if needed. Run the code and comment on the results. Is doc 1 more related to doc2 or doc3?

**Problem 3:**

a): For this question, we are using 3 sets of documents corresponding to three main topics: COVID-19, politics, and tesla, respectively. The files are named in this order (doc 1-10 are for topic 1, and so on). Complete the missing codes to build SVD. You may load the topic\_doc.csv file into Excel, convert scientific numbers to regular numbers, and examine the patterns.

b): we pick a certain file to compute its cosine similarity with other files. Here, doc1 is chosen. Complete the missing codes that compute its similarities with files within the same main topic (intra-group) and similarities with files outside its main topic. Compare the average results.

5 Reference

[1] ECE 532 lecture notes

[2] Prof. Yingyu Liang, CS 540, *Natural Language Processing Basics*

[3] J. L. Leidner and V. Plachouras, “Ethical by Design: Ethics Best Practices for Natural Language Processing,” *ACL Anthology*. [Online]. Available: https://www.aclweb.org/anthology/W17-1604/. [Accessed: 01-May-2020].

[4] All .txt documents are extracted from news from Google queries

6 Appendix

See file *SVD NLP.ipynb*

See file *SVD NLP (answer).ipynb*

All other text documents are obtained from the internet (by Google search)

7 Solution

Problem 1

a): These three documents are not related in terms of topic.

b): Each entry represents normalized term frequency + inverse document frequency of a given word. If an entry is 0, this means that this particular word does not appear in this document. Each row in this matrix represents the tf-idf values of a given word across all documents.

c):

(i)The first document corresponds to the third topic, the second one corresponds to the third topic, and the third one corresponds to the first topic.

(ii) According to the tf-idf matrix, the 3 documents have no overlapping words. This suggests that each document is completely related to its own topic and addresses nothing else. It is important to note that in real-world, with each document using overlapping words, one document will have multiple non-zero entries in the topic\_doc matrix. Each topic here does not necessarily refer to real-world topics like computers, politics, but instead simply refers to a pattern in their choice of words, suggesting a commonality.

Problem 2:

Notice that the two cosine similarities are very different: one is significantly larger than the other one. A larger value means that the two documents are more likely to belong to the same category.

Problem 3:

See file *SVD NLP (answer).ipynb*