Twitter: Concept Identification using Latent Semantic Analysis LING-L645-Advanced Natural Linguistic Processing

Akshay Reddy
Shridivya Sharma

December 14, 2017

Contents

1 Abstract

It is a challenge task to discover major topics from text, which provide a better understanding of the whole corpus and can be regarded as a text categorization problem. The goal of this paper is to apply latent semantic analysis (LSA) approach and attempt to extract common factors that representing concepts hidden in a large group of text. LSA involves three steps: the rst step is to set up a term-document matrix;the second step is to transform the term frequencies into a term-document matrix using various weighting schemes; the third step performs singular value decomposition(SVD) on the matrix to reduce the dimensionality.

We begin the experiment by focusing on the tweets posted during Las Vegas Shooting (Oct 1, 2017) and Thanksgiving week (2017). Later we also uses the tweets posted by the top ten new channels in the world over a span of 10 days to find out the major concepts contend in the text. We apply the latent semantic analysis approach on the twitter text. The result is able to identify not only principle categories, but also major themes contained in the text.

Keywords: Latent Semantic Analysis, Topic Extraction, Text Mining, Information Retrieval

2 Introduction

Many multidisciplinary elds, such as data mining, bioinformatics, biochemistry, and neuroscience, emerged in the past several decades. Since multidisciplinary elds involve theories,methods, and techniques from multiple disciplines, it is not easy to comprehend all the research eorts in these elds. Text categorization, which organizes documents into groups based on their underlying structures, can help capturing the large amount of activities and diversity of a multidisciplinary eld.

The goal of this paper is to apply latent semantic analysis (LSA) approach on the tweets posted by popular new channels and find out the most trending breaking news over a specific span of time. LSA is an automatic mathematical and statistical technique for uncovering common factors that representing concepts hidden in Text [1],[2],[3],[4].Previous investigations in psychology and computer science have proved that LSA resembles the way the human brain distills meaning from text and is capable of inferring much deeper relations in the text data[3],[5]

The rest of the paper is organized as follows. Section 2 describes the basic concepts of LSA and overview on twitter. Section 3 presents the experimental study that was used to identify the core research areas. Section 4 discusses the results of this analysis. Section 5 summarizer the paper.

3 Research Method

3.1 Latent semantic analysis (LSA)

This is a theory of knowledge acquisition, induction and representation[2]. It was rst introduced as an information retrieval (IR) technique by [1] and [6].It is an automatic mathematical learning technique for analyzing the relationships and similarity structures among documents and terms, relying on no human experiences, prior theoretic models,semantic dictionaries, or knowledge bases[3].

Similar to factor analysis, principal components analysis, and linear neural networks, the main purpose of LSA is dimension reduction, which is realized through a matrix operation called singular value decomposition (SVD). SVD is a means of decomposing a matrix into a product of three simpler matrices. By retaining the k largest singular values, the resulting reduced-order SVD provides the best k-dimensional approximation to the original matrix, in the least square error sense[7]. In the results of SVD, two sets of

factor loadings, one for the words and one for the documents, are generated. Each term and document is represented as a k-dimensional vector in the same latent semantic space derived by the SVD. Thus each latent semantic factor is now associated with a collection of high-loading terms and high-loading documents[5].

High-loading terms and documents are used to interpret and label the corresponding factor. The number of actors is an input parameter that needs to be provided before SVD computation. As the number of factors changes, LSA groups key terms or documents into various levels of aggregation. When it is applied to identify important topics of a certain discipline using a collection of representative papers, a higher level of aggregation (e.g., 2 factors) indicates key research areas and a lower level of aggregation (e.g., 100 factors) represents general research themes[5]. The LSA analysis can be summarized in three main steps.

The rst step is to set up a term-document matrix in which each row stands for a keyword or term and each column stands for a document or context in which the key word appears. An entry in the matrix is the frequency of a key word in the corresponding document.
The second step is to transform the term frequencies in a term-document matrix using various weighting schemes.

The third step is to perform SVD on the matrix to reduce the dimensionality, which is the key feature of the LSA method. In this step only the k largest singular values are retained.
The reduced-order SVD is the best k-dimensional approximation to the original matrix[7]. Extensive experiments have demonstrated that the classication performance of LSA is robust[8] and it is capable of inferring relations in the text [3,5]. It can be used in information retrieval (IR), search optimization, classication, clustering, ltering and other IR-related applications[7].

3.2 Twitter:

Twitter is an online news and social networking service where users post and interact with messages, called "tweets".

3.2.1 What is twitter data?

Twitter data is the information collected by either the user, the access point, what's in the post and how users view or use your post. While this might sound somewhat vague, it's largely due to the massive amount of data that can be collected from a single Tweet.

With this information, you can know demographics, total clicks on your profile or how many people saw your Tweet. This is just the tip of the iceberg, but understanding the data allows you to know how it's used and the patterns of your content.[8]

3.2.2 Twitter data applications:

Twitter's popularity as an information source has led to the development of applications and research in various domains. Humanitarian Assistance and Disaster Relief is one domain where information from Twitter is used to provide situational awareness to a crisis situation.

Researchers have used Twitter to predict the occurrence of earthquakes [10] and identify relevant users to follow to obtain disaster related information [9]. It has also been used to predict the change in the stock market rates.

4 Experimental Study

4.1 Inspecting the Dataset:

GetOldTweets, is a python module which was used to retrieve the old tweets. The module works as a web scraping tool. The strings/username that is need to be searched is passed to the twitter page along with the specific date. The resulting information obtained in the form of HTML/ XML is processed and stored in the desired file format.

4.1.1 Las Vegas Shooting:

We used the string "LasVegas" to fetch all the tweets containing the term "LasVegas" from September 26, 2017 to October 3, 2017.With an average of 3000 tweets were fetched for each day.

4.1.2 Thanksgiving:

We used the string "Thanksgiving" to fetch all the tweets containing the term "Thanksgiving" from November 21, 2017 to November 26, 2017. With an average of 3000 tweets were fetched for each day.

4.1.3 Breaking News:

Below is the list of popular news channels that we considered to fetch the tweets.

1. Breaking News - @BreakingNews
2. BBC Breaking News - @BBCBreaking
3. CNN Breaking News - @cnnbrk
4. WSJ Breaking News - @WSJbreakingnews
5. Reuters Live - @ReutersLive
6. CBS Top News - @CBSTopNews
7. AJE Live - @AJELive
8. Sky Newsdesk - @SkyNewsBreak
9. ABC News Live - @ABCNewsLive
10. TWC Breaking - @TWCBreaking

We used the string respective username to fetch all the tweets from December 01, 2017 to December 10, 2017. Maximum of 100 tweets were retrieved per day, totaling an approximate of 1000 tweets with respect to each of the news channels over a span of 10 days..Modification was done to the module to retrieve tweets only in English language. The tweets are stored in .txt format.

4.2 Text Preprocessing:

The initial step of LSA analysis is to represent the text as a term-document matrix in which each row stands for a term and each column stands for a document. In order to set up such a matrix, this study started the analysis with text preprocessing procedures that are popular in the information retrieval and text mining.

4.2.1 Cleaning:

The tweets that we obtained contained a lot of unnecessary entities like URLs, Hashtags, Mentions, Reserved, words(RT, FAV), Emojis, Smileys. It was thus necessary to get rid of these entities. One approach is to directly remove them by the use of specific regular expressions. Another approach is to use appropriate packages and modules. We have used a text cleaning python module called as tweet-processor.

4.2.2 Removal of Stopwords:

Some words in the English language, while necessary, don't contribute much to the meaning of a phrase. These words, such as "when", "had", "those" or "before", are called stop words and should be filtered out. The Natural Language Toolkit (NLTK), a popular Python library for NLP, provides common stop words. We have made use of the list of English stopword present in NLTK corpora. We have also added few of the stopwords as per our analysis.

4.3 Tokenization (N-gram model):

Now that we've enriched the corpus with meaningful terms, we're ready to construct features. We began by breaking apart the corpus into a vocabulary of unique terms, a process called tokenization. However, there are several ways to approach this step.

N-gram model, this is one of the tokenization method where divide the string in a sequence of N term, called as N-gram. The N-gram model preserves word order and can potentially capture more information than the bag of words model. After many experimental analysis we have decide to use a Bigram model.

4.4 Implementation of tf-idf:

Having selected a tokenization strategy, the next step was to compute the n-gram's frequency using using some statistic. We decide to use tf- idf.

The term frequency (tf) tallies the occurrences of each n-gram for document. However, some n-grams will undoubtedly show up often in any document, while others rarely appear in the overall corpus but show up frequently in certain subsets of documents. Therefore, to emphasize the latter, more interesting set of n-grams, we downweighted the term frequency with inverse document frequency (idf), which is calculated by logarithmically scaling the inverse of the fraction of training examples that contain a given term. Combining these two statistics yields the tf-idf statistic

We used TfidfVectorizer tool provided by Scikit-learn that performs n-gram tokenization and also computes the tf-idf statistic.

Finally, we're equipped to transform a corpus of text data into a matrix of numbers with one row per training example and one column per n-gram.

4.5 Latent Semantic Analysis (LSA):

Input: X is our matrix where m is number of documents and n is the number of terms.

We are going to decompose X into three matrices U, S and T. After the decomposition we determine k, that is the number of concepts that we are going to find.

$$X = U\ S\ V \wedge t$$

1. U will be m x k matrix. The rows will be documents and the columns will be the concepts.
2. S will be k x k diagonal matrix.The elements will be the amount of variations captured from the concepts.
3. V will be m x k (Transpose) matrix. The rows will be terms and the columns will be the concepts.

In this section we perform dimensionality reduction using truncated SVD (aka LSA). The truncated SVD works on term count/tf-idf matrices as returned by the vectorizers in sklearn.feature_extraction.text. In that context, it is known as latent semantic analysis (LSA).

The truncated SVD function accepts two parameters.

1. Number of concepts
2. Number of iterations to perform

In our experiments we will be finding 5 concepts and performing 100 iteration.

5 Result and Discussion

5.1 Thanksgiving (2017):

	Important Terms	Inference
Concept 1	Happy, family, day, hope, great, weekend, dinner, break, holiday, time	Theme of Thanksgiving
Concept 2	break, weekend, leftovers, hope, back, great, christmas mlbb, giveaway, via, mlbb	Mobile Legend Bang Bang (online gaming) and Gifts giveaways.
Concept 3	break, week, america vote america vote yes, bless america vote, concern god concern god bless, discover rescue, discover rescue plan, ebook discover	American elections and Hope
Concept 4	harvey weinstein accusers, accusers spend, accusers spend holiday holiday together, spend holiday together, weinstein accusers spend climb, climb record climb record high, online sales climb	Celebrity Fight and Increase in the online sales
Concept 5	day, leftovers, yesterday, mlbb giveaway, mlbb, hope, giveaway, man charged, friday, class cookbook	Thanksgiving leftovers

Table 1

Table 1 describe the concepts that we tried to extract during the Thanksgiving week.

Concept 1: Looking at the terms we can see that this concept signifies the theme of thanksgiving. With the terms like "family", "dinner","weekend", "happy".

Concept 2 [**11**]: Mobile Legend Bang Bang is a online gaming website where people were allowed to enjoy the feast by playing a match, exchange gifts with friends. The terms "mlbb", "giveaway" strongly signifies this concept.

Concept 3: This concept more highlights terms like "God", "America", "Vote". Which is more indicated about the american elections and hope.

Concept 4[**12**]: Nov 23, 2017 "Uma Thurman blasts Weinstein in Thanksgiving message". This concept highlights the celebrity fight with terms like "harvey weinstein accusers". It also the increase in the online sales with terms like "online sales climb"

Concept 5: This was more of a weaker concept, which had a lots of idea which were repeated. The new concept that hilights is the thanksgiving leftovers.

5.2 Las Vegas Shooting (2017)

	Important Terms	Inference
Concept 1	Shooting, new, video, like, go, see, via, get, going, one	Start of the attack, video sharing
Concept 2	Shooting, victims, massacre, prayers, shooter, mass shooting, tragedy, mass, gun people	Extent of tragedy
Concept 3	Robbery, nine years, prison, simpson, years prison, years nine years prison, years robbery, botched robbery, simpson served	O. J. Simpson robbery case decided
Concept 4	Edc, going, tickets, go, kuchar, er shooting police, jhonattan, time, like, scott	ashton kutcher's movie las vegas release

Table 2: Describe the concepts that we tried to extract during the Las Vegas attack.

Concept 1: This concepts highlights the start of the unfortunate attack. Terms like "Shooting" , "Video", "going" tell us the about videos being shared of the live situation.

Concept 2: This concept highlights the mass tragic death of people with the terms like "mass", "massacre", "gun", "tragedy".

Concept 3: October 3, 2017 O. J. Simpson robbery case cased was decided by the court. Terms like "simpson", "robbery", "nine years" highlights this topic.

Concept 4: This concept shows an amalgamation of several topics and has tried to move away from the shooting tragedy. It highlights release of ashton kutcher's movie las vegas. The terms like "kucher", "tickets", "go" signifies this concept.

5.3 Breaking News:

	Important Terms	Inference
Concept 1	Army, first, admit, admit army,admit army invaded, army invaded, army invaded kanu, army must, defence minister, defence minister admit	A civil war situation in Nigeria, Army tried to kill leader Mazi Nnamdi Kanu
Concept 2	Rt, news, 8u, trump, reports, says, house, media, say, let	President Donald Trump (**Ambiguous**)
Concept 3	Wakey, wakey ignore,wakey wakey, wakey wakey ignore, aide might, aide might story, army attacks, army attacks villages, arrested, arrested exception	Army attacks villages (**Ambiguous**)
Concept 4	Reports, tillerson, says, replaced, secretary, cia, gaddafi, briefing, cia director cia director pompeo	CIA director pompeo, a briefing on gaddafi death

Table 3: Describe the concepts that we tried to extract during a span of 10 days.

6 Discussion

Based on the strength of the terms we have changed the number of concepts. LSA was able to determine most of the concepts occuring during a time frame. The table 3 has a couple of ambiguous concepts, we believe the reason to be a lager span of time (10 days). Since the result from table 1 and table 2 considers a span of 6 days.
It was very important to update the stop words with the term used to filter the tweets. If not the concept of all the filter terms will be formed.

7 Conclusion

This paper attempted to identify the major concepts and themes from the current trends by examining the twitter data using latent semantic analysis. In the experimental study, which involved data spanning over 6-10 days was processed and analyzed using several natural language processing techniques. The results were successful in inferring much deeper relations in the text data and determining the principle concepts.

References

[1] Deerwester, S.; Dumais, S.; Furnas, G.; et al. (1990). Indexing by Latent Semantic Analysis,Journal of the American Society for Information Science, 41(6): 391-407.

[2] Landauer, T.; Dumais, S. T. (1997). A solution to Platos problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, Psychological Review, 104: 211-240.

[3] Landauer, T.; Foltz, P.; Laham, D. (1998). Introduction to Latent Semantic Analysis, Discourse Processes, 25: 259-284.

[4] Kou, G.; Lou, C. (2012). Multiple Factor Hierarchical Clustering Algorithm for Large Scale Web Page and Search Engine Clickstream Data, Annals of Operations Research, 197(1)25:123-134.

[5] Sidorova, A.; Evangelopoulos, N.; Valacich, J. S.; et al. (2008). Uncovering the intellectual core of the information systems discipline, MIS Quarterly, 32(3): 467-482.

[6] Dumais, S. T.; Furnas, G. W.; Landauer, T. K.;et al (1988). Using latent semantic analysis to improve information retrieval, Proceedings of CHI88 Conference on Human Factors in Computing Systems, 281-285.

[7] Dumais, S. T. (2004). Latent Semantic Analysis, Annual Review of Information Science and Technology, 38: 189-230

[8] Twitter data
https://sproutsocial.com/insights/twitter-data/

[9] S. Kumar, F. Morstatter, R. Zafarani, and H. Liu. Whom Should I Follow? Identifying Relevant Users During Crises. In Proc

[10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: RealTime Event Detection by Social Sensors. In Proceedings of the 19th international conference on World wide web, pages 851860. ACM, 2010.

[11] Reddit
https://www.reddit.com/r/mobilelegends/comments/7e0ma2/mlbb_thanksgiving_feast_is_right_around_the_corner/

[12] Uma Thurman blasts Weinstein in Thanksgiving message — Page Six