Chapter 4: Latent Semantic Analysis

In this chapter, the detailed concept of Latent Semantic Analysis (LSA) has been discussed. The working of Singular Value Decomposition on documents is explained followed by different approaches of selecting important sentences from a given document.

4.1. Introduction

Latent Semantic Analysis (LSA) is one of the statistical techniques to analyze relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. It is assumed that words having similar meaning will occur in similar pieces of text i.e. the distributional hypothesis. Distributional semantics is a research area that develops and studies theories and methods for quantifying and categorizing semantic similarities between linguistic items based on their distributional properties in large samples of language data. The basic idea of distributional semantics can be summed up in the so-called Distributional hypothesis: linguistic items with similar distributions have similar meanings.

It is possible to demonstrate the words meaning, and the sentences meaning at the same time by adopting this method. This is accomplished by extracting the meaning of the sentence by considering the words it covers and deciding the meaning of the words using the sentences, which contain the word. An algebraic technique called Singular Value Decomposition (SVD) is applied to determine the interrelations between sentences and words. The accuracy of the extracted information is improved by adopting the SVD approach since it has the capacity to reduce noise reduction besides having the ability to model the relationship between words and sentences.

Latent Semantic Analysis represents the meaning of sentences and words, and to understand how exactly it works an example is given below (Makbule Gulcin Ozsoy and Ferda Nur Alpaslan 2011).

Assume the following three sentences as input:

Sentence 1 (S1): - “The man walked the dog”

Sentence 2 (S2): - “The man took the dog to the park”

Sentence 3 (S3): - “The dog went to the park”
When LSA is applied on these three sentences, what we get is the connection between the words of the sentences as follows in Figure 4.1. The squares are the sentences and the diamonds are the words.

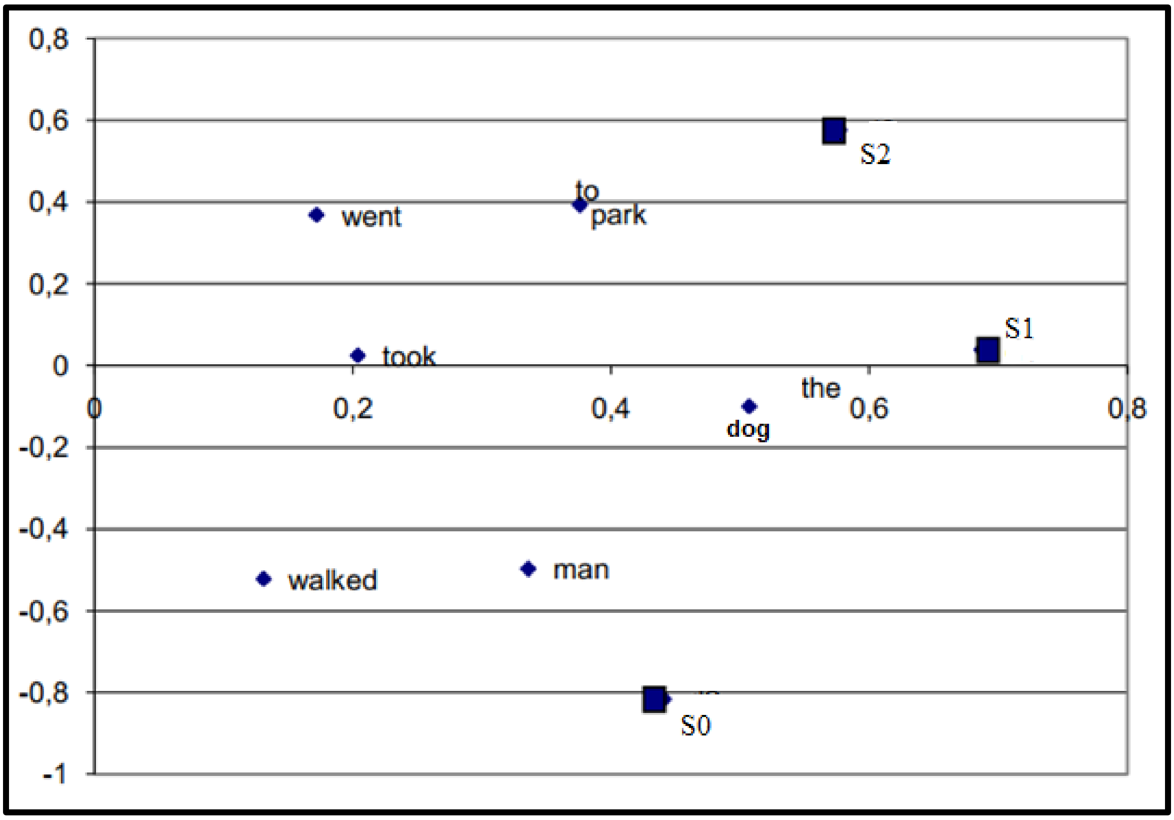

Figure 4-1. Meaning of sentences and words

From the above Figure 4.1, we can see that S2 is more related or close to S3 as compared to S1. Moreover, the word “walked” is not much related to “park” whereas it has more relation with the word “man”. Without any external knowledge, such information or analysis can be retrieved with the help of LSA.

The LSA has three steps for summarizing text from documents which have been briefly described below and in detail in later part of this chapter.

1. Input Matrix Generation: - an input document needs to be represented in a way that enables a computer to understand and perform calculations on it. This representation is usually a matrix representation where columns are sentences and rows are words/phrases. The cells are used to represent the importance of words in sentences.

2. Singular Value Decomposition: - SVD is an algebraic method that can model relationships among words/phrases and sentences. In this method, the given input matrix is decomposed into three new matrices.
3. Sentence Selection: - After performing SVD, important sentences are selected to generate the summary. There are various approaches and methods to find the important sentences which are discussed in detail later.

4.2. Input Matrix Generation

This is the first step for implementing LSA and a very important one as the next two steps depend on the output of this one. The matrix is to be prepared and presented in such a way that computers can learn and perform the calculation as necessary. For that, the representation can be done via a matrix where the columns are represented as documents/paragraphs and rows are represented as unique words/terms, which appear in the documents. Various methods can be used for filling cell values i.e. frequency of the words, a binary representation of words mentioning the presence of absence of the word, the TF-IDF scores etc.

Before generating the matrix it is important to pre-process the document. The pre-processing steps have been already discussed in detail in the previous chapter. They are briefly explained below.

4.2.1. Tokenization

Tokenization is a method, which splits sentences into small chunks and tokens. Sentences are divided into words after performing tokenization. Tokenization is known as text segmentation or lexical analysis. Sentences are tokenized using punctuation (full stops, commas, etc.) and also by space separators.

4.2.2. Stop Words Removal

As a part of pre-processing, stop words removal from the bag of words that we have after tokenization is very important. The advantage of the removing stop word is to reduce the size of the document by almost 30% to 40%. Stop words normally occur very frequently in documents and they are not important for Text Mining in general. Some of the common stop words that appear are articles such as a, an, the etc.

Stopword is different as per requirement and usage. It applies significantly to the application, which means for Text Summarization, stop words are different than that for Information Retrieval and Topic Modeling. The words that are not associated with noun and verb parts should be considered in the stop word corpus. A, an, the, special symbols, double quotes, single quotes, etc. are included in the stop word list to increase the performance of the model in the summarization process.

4.2.3. Stemming

Stemming is a next step after removing after stop words from the document. Stemming is also known as truncation. Stemming is a method that sanctions us to search for different word endings and spellings at the same time. It confirms that words having the same stem are considered together. In other words, terms like run, running, and runner – all have the same root word which is – run. Thus they all should be considered together as the same word.

There are many stemming algorithms already developed, each with their own set of advantages and disadvantages. One of the most popular stemmer is Porter's (Porter 1997) algorithm which has proven to be very popular for its effectiveness for stemming. With the help of the stemming algorithm, the size of the corpus is reduced, resulting in faster and accurate execution of the Text Summarization algorithm.

As discussed earlier, many different types of stemming algorithms are available like the popular ones - Lovings stemmer (Lovins 1968) and Paice/Husk stemmer (Paice 1990a). In the work done in this research, Porter's stemming algorithm has been implemented.

4.2.4. Term Frequency

The next step after stop words removal and stemming is finding the term frequencies to determine the importance of a word in a sentence or in the given document.

The term frequency (tf) method has widely used in the information retrieval system and Text Summarization due to its simplicity and flexibility. Term frequency basically means counting the number of times each word occurs in the sentence / document. However, in the case where the length of documents varies greatly, adjustments are often made to normalize the term frequency. This has also been discussed in the previous chapter.

4.2.5. Weighting Cell Values

The result of Singular Value Decomposition may change as per the cell values. The value of cell demonstrates words in sentences. Many types of techniques are available to fill the cell values. Some popular techniques are given below.

1. Word Frequency: - The number of times the word occurs in the sentence is the number filled with the cell.
2. Binary Representation: - The cell is filled with either 0 or 1 as a binary number as per the absence or presence of a word respectively in the sentence.
3. TF-IDF (Term Frequency-Inverse Document Frequency): - As briefly explained earlier, the cell values are filled with TF-IDF scores. If the TF-IDF score is high, it means it is a rare word and a common word has a low score. Thus a higher TF-IDF indicates that, that word is most illustrative and significant.
4. Log Entropy: - The cell value is occupied with log entropy that provides details on word is how useful in the sentence. It is calculated as follows:
5. Root Type: - The root type is a technique, which is same as the binary demonstration. Here, the cell value is filled with 1or 0. If word type is a noun than value is occupied with 1 else 0.
6. Modified TF-IDF: - With this approach, noise values are removed from the input matrix. First, all cell values are filled with the TF-IDF scores. If cell values have fewer score compared to average TF-IDF score then it is set to 0.

4.3. Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is a way of extracting features from documents. SVD is a statistical method that maps relation between sentences and words. In SVD, the provided matrix A is decomposed into three different matrices.

Yihong Gong et al (2002) presented the idea of Text Summarization, by use of Latent Semantic Analysis (LSA). In LSA based Text Summarization, we need to apply SVD on it. In the first step, the process starts with generating a matrix, the matrix is formed using sentences matrix A= [A1, A2, A3….An], where Ai represents a column vector and value of it filled any one method as explained previously. At a later stage, m X m matrix is ready for a document, where m is total terms and n sentences in the document. The matrix A may be sparse because every word does not occur or appear in each sentence or documents.

Given a m X n matrix A, the Singular Value Decomposition calculates as follows.

$$A = U\Sigma V^T$$

A: Input matrix (m x n)

U: Words x Extracted Concepts (m x n) (4.1)

∑: Scaling values, diagonal descending matrix (n x n)

V: Sentences x Extracted Concepts (n x n)

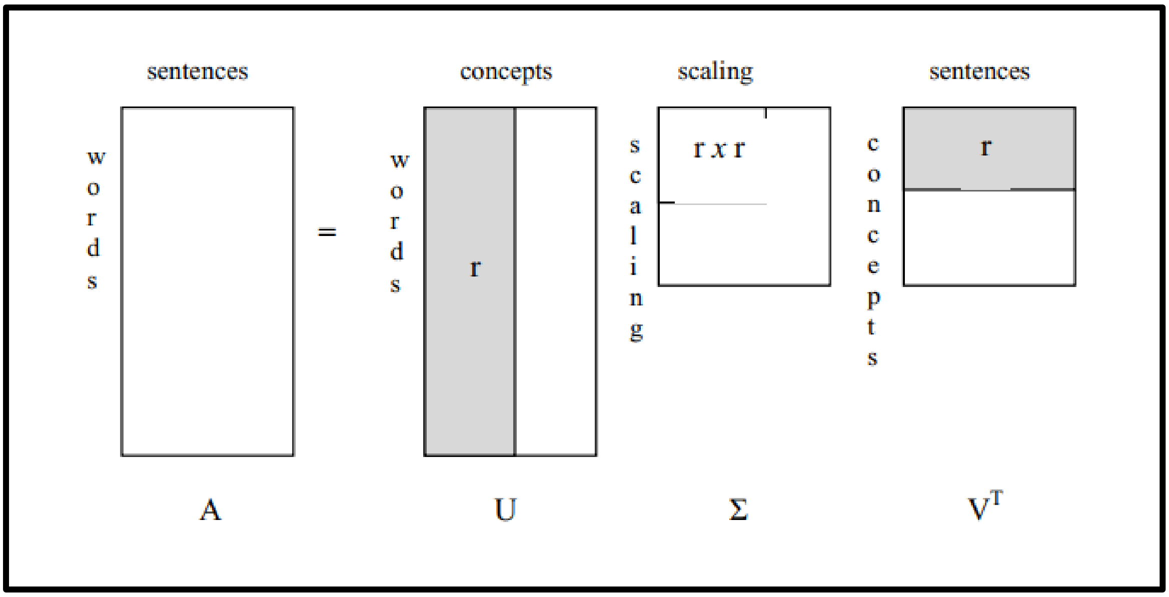

Figure 4-2 Singular Value Decomposition

In the Figure 4.2, ∑ describes the relative strengths of the features, U describes the relationship between terms (rows) and features (columns) and V^T describes the relationship between features (rows) and documents (columns). Here U = [Uij] is the orthonormal matrix whose columns are left singular vectors. ∑ is n x n diagonal matrix, whose value of diagonal elements is positive singular values and sorted in descendant orders.

SVD is able to convert to r-dimensional singular vector space from the m-dimensional term vector space. With help of mapping, it acknowledges the input document's latent semantic structure. The topic of the document is represented by the linearly independent vectors *r*, the value of r is equal or less than min (m, n).

The value of SVD is based occurrence of words. If words are frequently occurring in the document, then it considers them to be related to each other. The magnitude of vector provides information of the concept that is significant. The concept refers to sentences, which contains the connected words are estimated with singular vectors. The sentence that holds the highest singular values is considered as one of the most representative sentences in the document.

The dimension reduction of *r* is very important because it affects directly in the performance of calculation. As explained by (Deerwester et al., 1990), the input data should be filled with a value of *r*. Moreover, it does not have any noisy data i.e. unimportant information.

One of the major drawbacks of SVD is it is too time consuming. When different terms/sentences are added into the matrix then SVD calculation is to be performed again to update the values. Another additional drawback is that of polysemy. Polysemy means that a word has multiple meaning. For example, the word bank refers for financial matters, and it refers to the area near a river. Such words may not be explicitly well separated through SVD.

4.4. The approach to Sentence Selection

In Text summarization, the most important is selecting the sentences, which have the highest ranking, or importance as far as generating a summary is concerned. There are a number of sentence selection algorithms. Next section explains the sentence selection algorithms that are popular currently.

4.4.1. Gong and Liu Approach to Sentence selection (2011)

The algorithm of Gong and Liu is one of the main studies conducted in LSA-based text summarization. As explained above, the V^T matrix which stores the extracted concepts for sentences is used to find the important and significant sentences. In the V^T matrix, the row order indicates the importance of the concepts in such a way that the first row represents the most important concept extracted. The cell values of this matrix show the relation between the sentence and the concept. A higher cell value indicates that the sentence is more related to the concept.

In the Gong and Liu approach, one sentence is chosen from the most important concept, and then a second sentence is chosen from the second most important concept until a predefined number of sentences are collected. The number of sentences to be collected is given as a parameter. In Example 1, three sentences were given, and the SVD calculations were performed accordingly. The resulting V^T matrix having rank set to two is given in Figure 4.2.

Table 4-1 V^T matrix from each row

V^T matrix (r=2)			
	Sentence 1	**Sentence 2**	**Sentence 3**
Concept 0	**0,457**	**<u>0,728</u>**	**0,510**
Concept 1	**-0,770**	**0,037**	**0,637**

In the above Table 4-1, V^T matrix calculations as per three sentences are provided. As per strategy of the algorithm, it initially selects Concept 0, in which the Sentence 2 has the highest cell value and hence it gets selected for summarization.

The disadvantages of Gong and Liu approach have been addressed by Steinberger and Jezek (2004). The disadvantages can be enumerated as follows:

1. The number of sentences to be collected remains the same with the reduced dimension. This means if the initial number is large there is every possibility that sentences from trivial concepts get selected too.
2. There is every possibility that some important concepts have more than one sentence which should get selected but since we are selecting only one sentence from each concept, sentences which are more important but not top ranking and not having the highest cell value do not get selected at all.
3. It is assumed that all the chosen concepts have the same importance level which is generally not true.

4.4.2. Steinberger and Jezek (2004)

In this approach, few steps remain the same as the previous approach i.e. Gong and Liu. Initially, the input matrix is generated, and after that, SVD is performed on the input matrix. The Steinberger and Jezek technique uses both matrices i.e. V and ∑ for selection of sentence.

In this approach, the length of each sentence vector, represented by the row of V matrix, is used for sentence selection. The length of the sentence i is calculated using the concepts whose indexes are less than or equal to the given dimension. ∑ matrix is used as a multiplication parameter in order to give more emphasis to the most important concepts. The length of sentence i is calculated as per Equation 4.2.

$$Length = \sqrt{\sum_{j=1}^{n} V_{ij} * \sum jj} \qquad (4.2)$$

The sentence with the highest length value is chosen to be a part of the resulting summary.The dimension size is two for this example. Since the sentence sentence1 has the highest length, it is extracted first as a part of the summary. The main purpose of this algorithm is to create a better summary by getting rid of the disadvantages of the Gong and Liu summarization algorithm. In the Steinberger and Jezek approach, selecting more than one sentence from a concept which is more significant is allowed and selecting at least one from each important concept is also done.

By using the formula given in Equation 4.2, we calculate the length of sentence, which is shown in Table 4-3. We have used two dimensions for this example, as per dimension Sentence 2 has a maximum length. Therefore, it may be extracted for the summary.

Table 4-2 Length Scores

Length Scores	
Sentence 1	**1,043**
Sentence 2	**1,929**
Sentence 3	**1,889**

4.4.3. Murray, Renals, and Carletta (2005)

In this methodology, the first two steps remain same as per previous two techniques i.e. creation on input matrix and performing SVD. The authors, for selection of sentences, have used two matrices i.e. V^T and ∑.

In this methodology, from the uppermost concept, more than one sentence can be selected, which are located in the first row of V^T matrix. After that the decision of selecting more number of sentences from each concept is done using the ∑ matrix. The value is decided by getting the percentage of the related singular value over the sum of all singular values, for each concept.

Table 4-3 Each row V^T matrix

V^T matrix (r=2)			
	Sentence 1	**Sentence 2**	**Sentence 3**
Concept 0	**0,457**	**0,728**	**0,510**
Concept 1	**-0,770**	**0,037**	**0,637**

In the above Table 4-3, the V^T matrix is calculated as per example that is discussed in Gong and Liu method. As we can see that from Concept 0, two sentences are selected - Sentence 2 and Sentence 3.

This method (Murray, Rentals, and Carletta) resolves the problem of selecting more than one sentence from the same concept, which existed with Gong and Liu's method. By using this approach, even if a cell does not have maximum value for the same concept it can select more than one sentence from the same concept. Moreover, the reduced dimension does not have to be the same as the number of sentences in the resulting summary.

4.4.4. Cross Method

Cross method is proposed by Ozsoy et al, which is an advancement of Steinberger and Jezek(2004) approach. In the cross method, the creation of the input matrix and calculation of SVD is the same as discussed before. Here V^T matrix is used for choosing of the sentence. A pre-processing step is performed in between sentence selection and calculation of SVD.

The primary purpose of pre-processing is that; eliminate the effect of the sentence, which are related to a concept in some way. In addition, it has to make sure that it does not eliminate the sentence, which is the most important one for a concept. For the concept, the average sentence value is computed in V^T matrix. After computing the average score, the cell score, which is less than average score, is to set zero. This process of setting the cell values to zero whose score is less than average removes less related sentences while keeping more related ones for that concept.

In cross approach, after performing pre-processing the V^T column matrix demonstrates the length of the vector for each sentence. The number of concepts to be selected can be given as input by the user. In case we do not take this input, all concepts are taken into account. For summary, the longest vectors are chosen.

Table 4-4 After Pre-processing V^T

V^T matrix (r=2)				
	Sentence 1	**Sentence 2**	**Sentence 3**	**Average**
Concept 0	~~0,457~~	0,728	~~0,510~~	0,565
Concept 1	~~-0,770~~	0,037	0,637	-0,021
Length	0	0,765	0.637	

As per given Table 4-4, after performing pre-processing, V^T matrix is prepared. For the same, initially the average score is computed and if the cell has less than average score, then cell value is set to zero.

Table 4-5 Matrix and length values

V^T matrix (r=2)			
	Sentence 1	**Sentence 2**	**Sentence 3**
Concept 0	0	0,728	0
Concept 1	0	0,037	0,637
Length	0	0,765	0,637

As per above Table 4-5, Sentence 2 has the highest length. Therefore, it is included in the summary.

4.4.5. Topic Method

The topic method is introduced by Ozey et al. The topic method is similar to another method that we have discussed earlier, which was also based on Latent Semantic Analysis. In this approach, initially, the input matrix is created, then after Singular Value Decomposition is processed on the input matrix. In the next stage, pre-processing is performed before extracting sentences for the summary. Here V^T matrix is used for pre-processing and selection of sentences.

This approach identifies the main-concept and sub-concept from the words of the sentences. After performing SVD, the concepts or the topics are extracted from the document. However, these topics can be sub-topics of other main topics as well. After finding the main topics, subtopics are merged, and then all sentences are selected for a summary with main topics.

The topic method follows the same steps as the Cross method. First, pre-processing is performed on SVD generated matrices, next, with the use of V^T matrix of the row, average score calculation is performed. After computing the average score, the cell scores which are less than average score are set to zero. If the sentence is not highly related to the concept then that sentence would be removed. Moreover, it makes sure that only significant sentences remain in the concept.

Table 4-6After pre-processing V^T

V^T matrix (r=2)				
	Sentence 1	**Sentence 2**	**Sentence 3**	**Average**
Concept 0	~~**0,457**~~	**0,728**	~~**0,510**~~	**0,565**
Concept 1	~~**-0,770**~~	**0,037**	**0,637**	**-0,021**

In the next step after pre-processing, the main topic is derived. A concept X concept is generated in the next step. For this step, a concept × concept matrix is created by finding out the concepts that have common sentences. The common sentences are the ones that have cell values other than zero in both concepts that are considered. Then the new cell values of the concept × concept matrix are set to the total of common sentence scores.

Table 4-7New Concept X Concept matrix

	Concept 0	**Concept 1**	**Strength**
Concept 0	**1,456**	**0,765**	**2.221**
Concept 1	**0,765**	**1,348**	**2.113**

In Table 4-7, Concept X Concept matrix is derived on V^T matrix. After the generation of the matrix, the strength of each concept is computed. For each concept, the strength value is computed by getting the cumulative cell values for each row of the concept × concept matrix. The concept with the highest strength value is chosen as the main topic of the input document. A higher strength value indicates that the concept is much more related to the other concepts, and it is one of the main topics of the input text. In Table 4-8, calculated strength values can be seen. Since Concept0 has the highest strength value, it is chosen to be the main topic.

Table 4-8 Strength values

	Strength
Concept 0	**2,221**
Concept 1	**2,113**

Table 4-9 After pre-processing V^T

V^T matrix (r=2)			
	Sentence 1	**Sentence 2**	**Sentence 3**
Concept 0	**0**	**0,728**	**0**
Concept 1	**0**	**0,037**	**0,637**

After performing the above steps, with help of Gong and Liu approach, pre-processing and V^T matrix, sentenced are selected. From each concept, one sentence is selected until the desired numbers of sentences are selected for summarization. In the topic approach, the selected main concept is used for sentence selection. As per Table 4-9, Sentence 2 is selected from concept 0 because it has a maximum score.

4.5. Limitations of LSA

LSA has several limitations. The first one is that it does not use the information about word order, syntactic relations, and morphologies. This kind of information can be necessary for finding out the meaning of words and texts. The second limitation is that it uses no world knowledge, but just the information that exists in input document. The third limitation is related to the performance of the algorithm. With larger and more inhomogeneous data the performance decreases sharply. The decrease in performance is caused by SVD, which is a very complex algorithm.

SUMMARY

In spite of the limitations, LSA has proven to be very effective in Text Summarization as would be seen in Chapter 5 and Chapter 6 wherein the actual work on Text Summarization has been done keeping LSA as the baseline in conjunction with other approaches.