International Journal of Engineering & Technology, 7 (3.34) (2018) 872-873

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Text Summarization for Psoriasis of Text Extracted from Online Health Forums using TextRank Algorithm

[1]Mamatha Balipa, [2]Dr. Balasubramani R, [3]Harolin Vaz, [4]Christina Shilpa Jathanna

[134]Department of MCA, [2]Department of I S & E, NMAM Institute of Technology, Nitte, India
mamathabalipa@nitte.edu.inbalasubramani.r@nitte.edu.in
**Corresponding author E-mail: harolinvaz0397@gmail.com, jathannashilpa0@gmail.com*

Abstract

Text summarization is a process by which a document with enormous text and information is condensed into a summary of fewer sentences. In other words, it's a technique employed to extract significant contents from a text document, so that it can be represented in a summary.In recent years, this technique has seen a tremendous importance in various fields including search engine, business analysis, market review, medical applications etc. In this paper we have attempted to implement automatic text summarization for summarizing information from online health care forums about the disease Psoriasis**.**

Keywords: *Text summarization, TextRank algorithm, summary*

1. Introduction

People search for health issues online. But the search results in a collection of links which the user has to wade through to get the necessary information. Also with the growth of internet, the tremendous availability of text information and documents is overwhelming the internet users. So, the reduction of amount of this text to shorter and rapt summaries has become very much significant. Because summaries reduce the reading time and makes the process of selecting documents easier. However, it is not possible to produce summaries of all the text manually, so there is a greater need for automatic methods which achieves this.It's not just enough to generate phrases or sentences that capture the essence of the source document. The summary should be accurate and should be read assuredly as a novel standalone document. Because, the intention of automatic text summarization is to generate a summary which are as reliable as those written by humans.

There are many applications of summary in day to day life. One of such applications is to create a summary of notes for students in the form of outlines that helps them to prepare for examinations. We can also see its other applications in reviews of a book or of a movie, bulletins of weather forecasts or stock market reports, minutes of any business meeting, etc., In this paper we summarize details of the disease Psoriasis from text containing symptoms, treatments, types of Psoriasis, etc, extracted from online health forums.

Methods of automatic text summarization are significantly in need to mark the increasing amount of data obtainable online to benefitdetermining of relevant information and to discoverrelated information faster.

There are two principalmethods available to automatic text summarization; they are:

1. Extractive Methods
2. Abstractive Methods

Extractive text summarization focuses on extracting the important phrases and sentences from the original text document to obtain the summary. It involves the technique of ranking the relevance of phrases from the text so as to choose onlythose which are most pertinent to the gist of the source.

In abstractive method of summarization whollynovel phrases as well as sentences are generated in order to bring out the meaning of the source document. In other words, this method examines the text using advanced techniques available in natural language to generate a summary.

Extractive methods are the most successful approach as it is easy, but abstractive methods help to fetch the more general solutions to the problem.

2. Literature Survey

In this section we have tried to cite some of the literature works of research done in the past in the area of automatic text summarization.

In 2004, Rada Mihalcea et al [1] introduced a ranking model based on graph known as TextRank for text processing and illustrated how this model can be effectively used in natural language applications. The authors made use of graph-based method which involves addition of a vertex for each sentence by creating links for related sentences.

M. S. Patil et al [2] proposed a summarization method which is based on many extractive text summarization methodologies, and on the SVM (Support-Vector Machine). This system tries to increase the value and performance of the summary produced by the clustering technique by cascading it with SVM.

Rasim et al [3] suggested a way for automatic text summarization using the extractive method with the help of an evolutionary algorithm. In their study, they put forward an unsupervised document summarization technique that generates the summary by extracting and clustering sentences from the source document.

Naresh Kumar Nagwani et al [4] used a three-level scheme for

summarizing which is centred on semantic similarity and recurrent terms divided into 3 parts- an input document, a summarizing algorithm, and final summarised document

3. Textrank Algorithm

TextRank is one of the algorithms used in Extractive text summarization techniques using Python. The fundamentalprinciple of this algorithm is to provide a score for each sentence in a text. Then the top n sentences are sorted to form an automatic summary.

To summarize a text we have implemented TextRank algorithm that takes large body of text as input, summarizes it and produces the summary.Basically TextRank is an algorithm that summarizes the document in three simple steps:

Tokenizing the document into sentences which means breaking the text into different pieces.

Creating a graph where nodes are sentences and connect the sentences with each other by edges. The edge weight depends on how similar two sentences are.

Implementing PageRank algorithm on the graph and score the sentences. Select the sentences with a highest score as the important sentences.

4. Implementation

First the text from online forums is extracted using BeautifulSoup class available in urllib2 module.

The topic of the text is confirmed to be Psoriasis by using Latent Dirichlet Allocation (LDA) algorithm.

The below given is the code for implementation of TextRank Algorithm using Python library:

```
import networkxas nx
from nltk.tokenize import sent_tokenize
```

Importing nltk and networkxlibraries:

The function to break the text document into sentences is as follows:

```
defsentenceextract(document):
returnsent_tokenize(document)
```

The following function returns a list of edges connecting the sentences. Here based on their similarity, weights are assigned to the sentences:

```
defedgeextract(nodes):
return[(start,end,similartext(start,end))
for start in nodes
for end in nodes
if start is not end]

defsimilartext(s1,s2):
   return len(common_words(s1,s2)) /
(log(len(words(s1)))+log(len(words(s2))))
```

```
def ranking(nodes,edges):
graph=nx.DiGraph()
graph.add_nodes_from(nodes)
graph.add_weighted_edges_from(edges)
returnnx.pagerank(graph)
```

Next we define a function that computes and returns a ranking of the nodes. Finally we write a function that returns a summary of larger text.

```
def summarize(document,num_summaries):
nodes=sentenceextract (doc)
edges=edgeextract(nodes)
scores=ranking(nodes,edges)
returnsorted(scores,key=scores.get)
```

5. Output

The output of the above text summarization algorithm is shown below. The algorithm extracts the number of sentences specified that are most relevant.


*Common signs and symptoms include:

Red patches of skin covered with thick, silvery scales
Small scaling spots (commonly seen in children)
Dry, cracked skin that may bleed
Itching, burning or soreness
Thickened, pitted or ridged nails
Swollen and stiff joints
Psoriasis patches can range from a few spots of dandruff-like scaling to major eruptions that cover large areas.
* Overactive T cells also trigger increased production of healthy skin cells, more T cells and other white blood cells, especially neutrophils.
* The spectrum of disease ranges from mild with limited involvement of small areas of skin to large, thick plaques to red inflamed skin affecting the entire body surface.
* Skin cells build up in thick, scaly patches on the skin's surface, continuing until

5. Conclusion

In this paper, text pertaining to the disease Psoriasis is extracted from online health forums and summarized using text summarizing techniques. Text summarization has its importance in both commercial as well as research community. Even though it is related to Natural Language Processing, it doesn't have to be hard. Specially using Python, automatic text summarization has become possible to achieve with the availability of lot of libraries. As such, taking these attainments into consideration, there is still substantial amount of research left in the area of Text Summarization, as it is still problematic to produce a summary that is as meaningful as that produced by humans in all languages.

References

[1] Rada Mihalcea and Paul Tarau, "Text-rank: Bringing Order into Texts," Proceeding of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004.

[2] M. S. Patil, M. S. Bewoor, S. H. Patil "A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique", International Journal of Computer Science and Information Technologies, Vol. 5, IssueNo. 2, ISSN: 0975-9646, pp.1584-1586, 2014.

[3] RasimAlguliev, RamizAliguliyev, "Evolutionary Algorithm for Extractive Text Summarization." Intelligent Information Management, 1, pp. 128-138, November 2009.

[4] Naresh Kumar Nagwani, Dr. Shrish Verma Associate "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal ofComputer Applications(0975- 8887) Volume 17- No.2, March 2011.