Text Image Extraction and Summarization

Neha Joshi

Vishwakarma Institute of Technology,Pune.

Abstract— **with the huge amount of increase in data these days data processing has become very important. It filters a large amount of data. The text mining tool finds relation between the words in text content and analyzes the results as well. Deriving quality information from text forms the crux of data analysis or text mining. This paper focuses on a text mining application. Text information present in images is recognized and is summarized according to requirement, i.e. number of lines that text needs to summarize is dependent on user. Text mining thus is used to save time of the user, increase the data efficiency. It is used to make computation on data that a human would definitely fail to do, that is for analytics of large volumes of data. Hence text image extraction and summarization is a necessity in the current scenario. If the efficiency of the proposed model is optimized, this model of Text Image Extraction and Summarization can be very beneficial. It can used as a ready to go , click image and get summary application in a variety of situations. In this proposed model, even the number of lines the content has to be specified, thus ensuring that the extent of summarization is completely user controlled.**

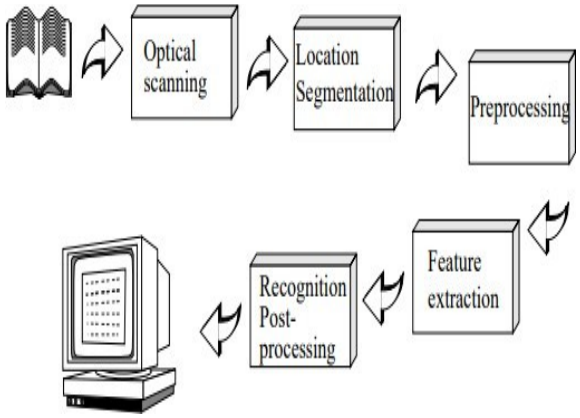
Keywords— *Text mining, Natural Language Processing, Optical Character Recognition, Summarization, Image Processing*

Introduction

Data Processing and Analysis have become the need of hour with the ever increasing amount of data that is being produced. This paper and the proposed model basically focuses on the textual data mining and analysis. This paper proposes an application that is basically dependent on two major domains Optical Character Recognition and Natural Language Processing. So the proposed model gives a basic solution in text mining: finding relevant information in textual images. This can be applied to various applications and used as quick summarizer of bulk data. Text Analytics, also known as text mining, is the

process of examining large collections of written or documented (typed) resources to generate new information. It is used to transform the unstructured text into structured data for use in further analysis and applications. Text mining identifies relationships and assertions that would otherwise have lost in the mass of textual big data. This big data is extracted and turned into structured data, for analysis, visualization.

Optical Character Recognition is primarily used to convert the human readable characters into machine readable codes like ASCII. Character to be recognized can be printed or handwritten.



I. LITERATURE SURVEY

[1] Optical Character Recognition - Tesseract is Open source OCR engine. It was initially developed between1984 to1994 at HP. In 1995, it was sent to UNLV for Annual Test of Optical Character Recognition Accuracy after the joint project between HP Labs Bristol and HP's Scanner Division in Colorado. Finally in 2005, Tesseract was released as open source by HP an available at Tesseract OCR website.

[2] Natural Language Processing with Python: Analyzing text with Natural Language Toolkit.

[3] Information extraction and text summarization using linguistic knowledge acquisition.- The lack of extensive linguistic coverage is the major barrier to extracting useful information from large bodies of text. Current natural language processing (NLP) systems do not have rich enough lexicons to cover all the important words and phrases in extended texts that is all basically all of the spoken language.

II. PROPOSED METHOD

The proposed model is divided in two main parts:

3.1 *Image Processing*- . Preprocessing is the primary step of OCR function. At this stage, bound operations are performed on the scanned image i.e. de-skew, changing a picture Text Extraction from pictures, from color to black and white, identifies columns, paragraphs, captions as completely

Natural language processing is the ability of a program to understand human language as it is spoken. NLP is a component of artificial intelligence and Machine Learning that is used in Data Analysis for meaningful data extraction in our model. Summarizer is built with the help of Natural Language Processing.

Fig 1-Examples of Text Images

different blocks and normalization. We've used the Optical Character Recognition Module from Python – Tesseract for preprocessing the text in the image. Tesseract works excellent when there is clean segmentation of the foreground text from the background. In practice, it can be extremely challenging to ensure these types of

segmentations. Hence we usually train domain-specific image classifiers. This extracted text is stored in a text file.

Fig 2- OCR System

2.1.1 Optical Scanning - when performing Optical Character Recognition, it is a common practice to convert the multilevel image into a bi-level image i.e., black and white. This process is known as Thresholding. The thresholding process is important because the result of the OCR is totally dependent of the quality of the bi- level image.

2.1.2 Location Segmentation - Segmentation is a process that determines the constituents of bi- level image. It is used to locate the regions of the image where data is present and distinguish it from background, graphics and other unnecessary figures It gives isolation to the character of words. Then these characters are recognized individually .Segmentation is performed by connecting separating all the connected black areas.

2.1.3 Pre-Processing – Scanned Image contains a lot of noise. Smoothing and Normalization are two of the most commonly used image preprocessing techniques.

2.1.4 Feature Extraction – It is basically extracting the most essential characteristics of recognized character. It is of the most difficult things in Pattern Recognition. Usually this is done with the help of Structural analysis and transformations. Template Matching and Correlation techniques are used in this step. Image matrix of input image containing text character is directly matched with a set of prototype characters representing each possible class. The Euclidian distance between the pattern and each prototype is calculated, and the class of the prototype giving the best match is assigned to the pattern and feature matching is completed.

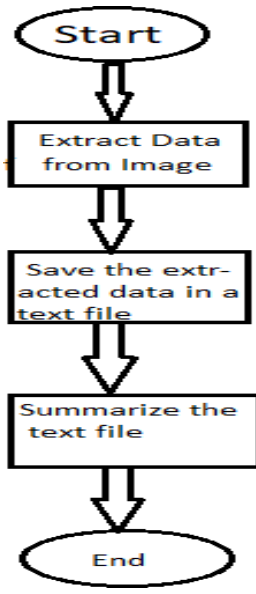
2.1.5 Classification – Matching type of classification covers the groups of techniques based on similarity measures where the distance between the feature vector, describing the extracted character and the description of each class is computed. Different measures may be used, but is Euclidean distance is most commonly used. This minimum distance classifier works quite well when the classes are separated but sufficient amount. When the entire character is used as input to the classification, and no features are extracted, an alternative - correlation approach is used. Here the distance between the character image and prototype images representing each character class is calculated.

3.2 Text Analysis – Summarization

Natural Language Processing is an intersection field of Machine Learning and Data Science that deals with the pre- processing, processing and analytics like forming summary. It is used to produce meaningful and quality information from the vast amount of data. In this proposed model, text analysis occurs in following steps-

1. Data Cleanup - Data is preprocessed. The objective of this data sanitization is to replace any extra whitespace characters besides the one whitespace that is after ending the punctuation. Sanitize input command of nltk library of Python is used to convert different white spaces and new line characters removes them and used String_translate to do so. [4]
2. Tokenizing – Once the data is processed, then the input is tokenized into words and sentences. Before this stopwords are eliminated. Stopwords are the words like punctuations,articles,or basically the words that do not convey much sense.Stopwords are hence prevented from being _scored‘ in final step ,so that they are not part of the final summary. Word_tokensize() and sent_tokenize() functions of nltk corpus are used to tokenize data. [5]
3. Scoring – Frequency of each word is scored and sentences are graded.FreqDist() is another function imported from nltk library of Python - It accepts a list of tokens, that is filtered word list in this case, and returns a structure where each key is the word and each value is the number of times that word occurred. Then the structure defaultdict is initialized, iterated over sentences and the score is increased based on the frequency of the particular word. The value of ranking will then contain key values of the sentence‘s numeric position, and hence their score.
4. Selection – N highest scoring sentences, where N is the desired length that we can be specified by the user.

f. Selection—Choose the top N sentences based on their scores and summarize.



V.ALGORITHM

ge data using Python and write e in a text file.[6] t file—Read the om the source, a exercise that is ract extracted text

c. Perform Preprocessing—Remove the white spaces, punctuation, stopwords – articles or words that do not convey much information from further computing.[7]

d. Tokenize text—Take the input and break it up into its individual words using parsed input data and tokenizing them.[8]

e. Scoring-Count Frequency of each word in data and assign a score to each one of them.

Fig 2- Flowchart of Model

III. RESULTS

Input Image-

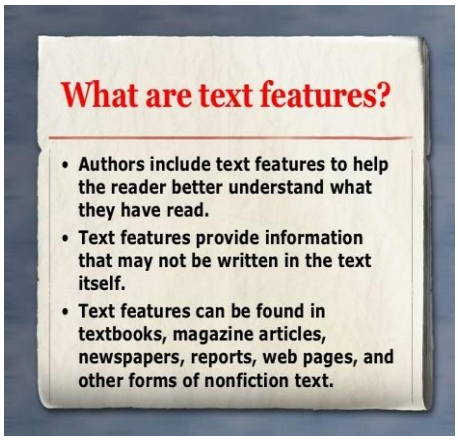


Fig 3 –Text Image

Text File Created –

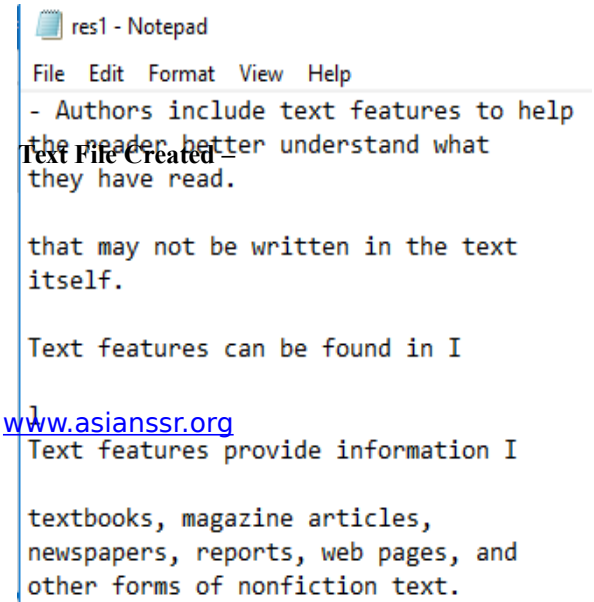


Fig 4 Text File

Final Output-
The output is the text extracted from the given text image.The text extracted is the converted to this output file.
Output -
–Authors include text features to help the reader better understand what they have read. Text features can be found in I l Text features provide information I textbooks, magazine art, newspapers, reports, web pages, and other forms of nonfiction text.‖

IV. CONCLUSION

Hence from the proposed algorithm text image extraction is done using Optical Character Recognition using Tesseract using Python. This extracted text is stored in text file and using Natural Language Processing Toolkit of Python[6], this text file is summarized according to user requirement of number of lines asked for summarization. It is a Text Mining Application and with the increase in accuracy, this proposed model can be very effective as data produced is increasing per day and data analytics is the need of hour

V. REFERRENCES

[1] Ravina Mithe, Supriya Indalkar, Nilam Divekar , _Optical Character Recognition‘ in International Journal of Recent Technology and Engineering, Vol 2 Issue 1 march 2013.

[2] The Tesseract open source OCR engine http://code.google.com/p/tesseract-ocr

[3] Lisa F Rau,Paul S Jacobs,Uri Zernik , _Information Extraction and Text Summarisation using linguistic knowledge acquisition‘ in Information Processing and Management, Volume 25,Issue 4 Page No -419-428.

[4] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, _Data Preprocessing for Supervised Leaning‘, International Journal Of Computer Science Volume 1 Number 1 2006 ISSN 1306-4428

[5] Jonathan Webster ,Chunya Kit, –Tokenization as Initial phase in NLP‖,City Polytechnic of Hong Kong,in proccedings of 14th Conference on Computational Linguistics,Vol 4 page-1106-1110

[6]A Mitthal,P Kumarguru –,Optical Character Recognition tool”,**IIIT D**

Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya ,‘ Preprocessing Techniques for Text Mining - An Overview‘ in International Journal of Computer Science & Communication Networks,Vol 5(1),7-16.

[7] Meyer, David and Hornik, Kurt and Feinerer, Ingo (2008) Text Mining Infrastructure in R. *Journal of Statistical Software*, 25 (5). pp. 1-54.

[8] Steven Bird,Edward Loper, _ NLTK : Natural Language Toolkit _,in proceedings of
Proceedings of the ACL 2004 on Interactive poster and demonstration sessions,Article no 31

[9] R. Smith. –An overview of the Tesseract OCR Engine.‖ Proc 9th Int. Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil, Sep 2007, pp629-633.

[10] The Tesseract open source OCR engine, http://code.google.com/p/tesseract-ocr.

[11] R.W. Smith, The Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987.

[12] Heuristic-Based OCR Post-Correction for Smart Phone Applications the university of North Carolina at chapel hill department of computer science honors thesis Author: Wing-Soon Wilson Lian 2009.

[13] Implementing Optical Character Recognition on the Android Operating System for Business Cards By Sonia Bhaskar, Nicholas Lavassar, Scott Green EE 368 Digital Image Processing.

[14] Hybrid Page Layout Analysis via Tab-Stop Detection

Ray Smith Google Inc. 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA. theraysmith@gmail.com, 2009.

[15] Optical Character Recognition Line Eikvil December 1993.

[16] NLP Applications of Sinhala: TTS & OCR Ruvan Weerasinghe, Asanka Wasala, Dulip Herath and Viraj Welgama Language Technology Research Laboratory, University of Colombo School of Computing, 35, Reid Avenue, Colombo 00700, Sri Lanka.

[17] Text To Speech: A Simple Tutorial D.Sasirekha, E.Chandra, March 2012.

.