International Journal of Research in Engineering, Science and Management
Volume 4, Issue 7, July 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Automatic Quiz Generator

Vaibhav Gupta[1], Hemlata Pant[2], Arjun Chaurasia[3*], Astha Dwivedi[4], Shubham Singh[5], Pragati Singh[6]
[1,3,4,5,6]*Student, Department of Computer Science and Engineering , Babu Banarasi Das Institute of Technology & Management, Lucknow, India*
[2]*Assistant Professor, Department of Computer Science and Engineering , Babu Banarasi Das Institute of Technology & Management, Lucknow, India*

***Abstract*: Multiple Choice Questions for any text image, either from long boring books or just some random hand written notes of yours. These questions are not from only that text image but also from internet, so that you can also encounter new questions i.e. extra facts, and who denies extra knowledge. Further it helps to revise and keep you up to date with your learning interests. In this paper we have aggregated different natural language processing technologies (NLP) and implemented them by their effective use for our project Automatic Quiz Generator in which text image is to be uploaded on API we provide so that the user gets the Multiple Choice Questions (MCQs) by searching online. We shall achieve that with the help of OCR, Keyword Extraction, Web Scraping, Django and different NLTK (Natural Language Tool Kit) tool.**

***Keywords*: MCQ generation, Natural Language Processing, Deep Learning, Online text, Text image.**

1. Introduction

In the fast growing world everyone is focusing on enhancing their skills by gaining more knowledge than they do possess. So that they could effectively use their learnings to make something creative but what is difficulty? It is to retain that knowledge for long time. That is why we keep on practicing to remember facts, but still reading long paragraphs is a difficult task which also contains some irrelevant information that wastes our lots of time. When we get some pragmatic problems there is nothing we can see from the phrases we read. To improve performance and grab lots of information in a less amount of time is tough, we need to search lots of information on sites, and yet we forgot them in fractions of seconds. If we somehow revise them not by just reading whole paragraph again but by taking a quiz which analyze our performance and tells how much we need to improve, this would definitely make our tedious work much easier. There are some sites which takes text and summarize them but they don't provide quizzes and information out of those texts but we have tried to make every work easier. We will make a deep learning model where the user just have to upload a text image and we process it to make MCQs (Multiple Choice Questions) some of which questions are not even in the text, as we also search different sites on behalf of user to make good Quizzes.

We shall achieve that with the help of Optical Character Recognition (OCR) which is an electronic conversion of the typed, handwritten or printed text images into machine-encoded text. [5], [6], Keyword Extraction is the automated process of extracting the most relevant words and expressions from text [1], [2], Web scraping is the process of processing a web document and extracting information out of it. [7], [8], Django and different NLTK (Natural Language Tool Kit) tool.

2. Proposed Work

In this paper we have done the following process:

A. Data Collection

We have collected an image or a pdf file from user and then we stored it for further processing. We pull together text images by photographs from camera that could be handwritten text, newspaper, magazines or circulars, booklets etc., screen shots of pdfs, docs file, online forms, websites, phrases, e-books, scanned data or any other documents.

B. OCR

Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine- encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

We used OCR in our project to extract the text content from the image uploaded by the user. We have used PyTesseract which is an engine.

C. Keyword Extraction

This is tasked with the automatic identification of terms that best describe the subject of a document Keyword extraction (also known as keyword detection or keyword analysis) is a text analysis technique that automatically extracts the most important words and expressions from a text which are termed as keywords, key segments or phrases.

Among different keywords extraction algorithms, we have implemented Text Rank algorithms, which is a graph based algorithms uses Google's famous page rank algorithm, where words are nodes and scores are edges. This helps us in extracting the important keywords from bulk data.

*Corresponding author: arjun2000ishu@gmail.com

D. Web Scraping

Web Scraping is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet)format.

We have used beautiful soup for this process which helps us in extracting the contents using <html> and <xml> tags for paragraphs from the web about the keywords extracted from the last phase.

E. Quiz Generation

In this phase, we generated quiz consisting of multiple choice questions (MCQ's) from the content we got from the last phase. For this we used Natural Language Processing Technique for information extraction and deep learning techniques to train model for finding most similar options of our answer. For this we used Spacy, Glove Model.

F. Modulation/Aggregation

In this phase we have aggregated all the working modules such as Keyword Extraction, Web Scraping and Quiz Generation as a single entity. Then we modified these modules to provide input and output for each other for better processing and parallelism.

G. Testing

Software testing is the process of verifying a system with the purpose of identifying any errors, gaps or missing requirement versus the actual requirement. Software testing is broadly categorized into two types - functional testing and non-functional testing.

Software testing involves the execution of a software component or system component to evaluate one or more properties of interest. We have tested our project with the various handwritten, typed images of topic and also test it with screenshots of the pdfs.

H. Deployment

Software deployment is all of the activities that make a software system available for use. The general deployment process consists of several interrelated activities with possible transitions between them. These activities can occur at the producer side or at the consumer side or both. Because every software system is unique, the precise processes or procedures within each activity can hardly be defined. Therefore, "deployment" should be interpreted as a general process that has to be customized according to specific requirements or characteristics

The complexity and variability of software products have fostered the emergence of specialized roles for coordinating and engineering the deployment process. For desktop systems, end-users frequently also become the "software deployers" when they install a software package on their machine. The deployment of enterprise software involves many more roles, and those roles typically change as the application progresses from the test (pre-production) to production environments.

3. Experimental Results

In our Application some certain experiments were taken to check the proper functioning of our project and some of them are:

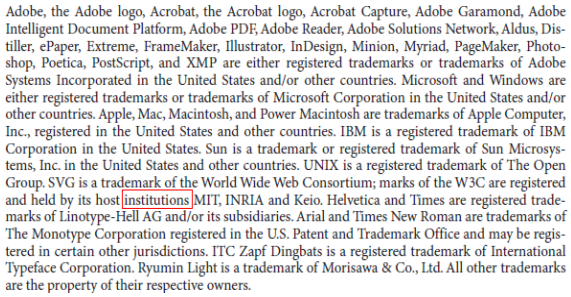
Adobe, the Adobe logo, Acrobat, the Acrobat logo, Acrobat Capture, Adobe Garamond, Adobe Intelligent Document Platform, Adobe PDF, Adobe Reader, Adobe Solutions Network, Aldus, Distiller, ePaper, Extreme, FrameMaker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photoshop, Poetica, PostScript, and XMP are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Microsoft and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Apple, Mac, Macintosh, and Power Macintosh are trademarks of Apple Computer, Inc., registered in the United States and other countries. IBM is a registered trademark of IBM Corporation in the United States. Sun is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries. UNIX is a registered trademark of The Open Group. SVG is a tradeмark of the World Wide Web Consortium; marks of the W3C are registered and held by its host institutions MIT, INRIA and Keio. Helvetica and Times are registered trademarks of Linotype-Hell AG and/or its subsidiaries. Arial and Times New Roman are trademarks of The Monotype Corporation registered in the U.S. Patent and Trademark Office and may be registered in certain other jurisdictions. ITC Zapf Dingbats is a registered trademark of International Typeface Corporation. Ryumin Light is a trademark of Morisawa & Co., Ltd. All other trademarks are the property of their respective owners.

Fig. 1. adobe.png, this text image is taken for the MCQ conversion

A. OCR

We have passed the image adobe.png fig. 1 to OCR Module for this experiment. And it gives the following output. Yield's initial not many lines are.

Adobe, the Adobe logo, Acrobat, the Acrobat logo, Acrobat Capture, Adobe Garamond, Adobe Intelligent Document Platform, Adobe PDF, Adobe Reader, Adobe Solutions Network, Aldus, Distiller, ePaper, Extreme, Frame Maker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photoshop, Poetica, PostScript, and XMP are either registered trademarks or trademarks of Adobe 'Systems Incorporated in the United States and/or other countries.

B. Keyword Extraction

Now the text from the previous module is keyed to create important keywords. First few lines of output are.

['registered trademark', ' trademarks', 'registered trademark adobe solutions network', ' registered trademark adobe reader', 'registered trademark adobe pdf', 'registered trademark adobe', 'registered trademark adobe garamond', ' registered trademark other countries']

C. Web Crawl

After important keywords extracted, we submit them to crawl module to get new paragraphs from the internet. Beginning not many lines of yield.

'Adobe Inc. (/əˈdoʊbiː/ ə-DOH-bee) is an American multinational computer software company. Incorporated in Delaware ', 'and headquartered in San Jose, California, it has historically specialized in software for the creation and publication of a wide range of content, including graphics, photography, illustration, animation, multimedia/video, motion pictures and print. The company has expanded into digital marketing management software. Adobe has millions of users worldwide. Flagship products include: Photoshop image editing software, Adobe Illustrator vector-based illustration software, Adobe Acrobat Reader and the Portable Document Format (PDF), plus a host of tools primarily for audio-visual content creation, editing and publishing.

D. Quiz Generation

We get content from the web, which we used to create tests.

A portion of the MCQs fig. 2 and its final assessment score fig. 3, from the application.

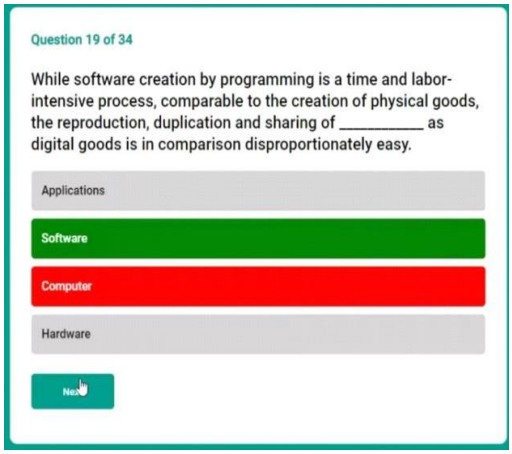

Fig. 2. MCQ 19, this quiz is one of the 34 MCQs which is generated from the input text image adobe.png, where green colour marks the correct answer, while red one shows the wrong answer for the quiz

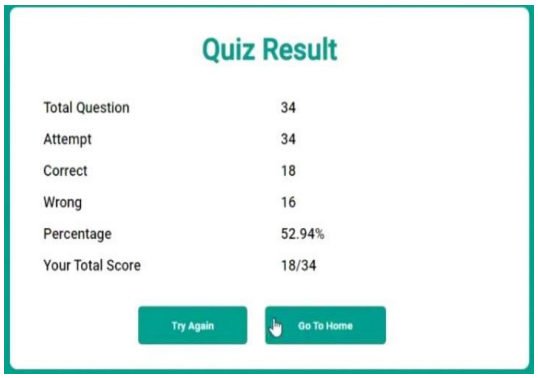

Fig. 3. Final Result, gives full analysis of questions attempts, percentage and scores

4. Conclusion and Future Work

As our application is based on E-learning. Hence, it has great scope on present as well as future scenario as we know due to this pandemic everything is turning online and hence the interaction between teachers and students getting less due to which students cannot check their knowledge on particular subject by giving test and cannot prepare themselves for competition. Hence, our application helps them to prepare for competition. They can check their knowledge by giving test and its results give the details of their performance. They can give these tests as many times as they want.

Our application could generate a quiz within a minute on any subjects whatever a user wants. Hence user don't have to search on Google for preparing the preparing the particular subject. At present we are going to give users these features but later we are going to enhance its more features i.e. the project can be scaled up to handle inputs from many other languages, text to speech facility etc.

References

[1] Mihalcea, R., Tarau, P., & Figa, E. (2004). PageRank on semantic networks, with application to word sense disambiguation. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, pp. 1126-1132.

[2] Mihalcea R, (2004, July). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL interactive poster and demonstration sessions, pp. 170-173.

[3] Naili, M., Chaibi A. H, & Ghezala H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. Procedia computer science, 112, 340-349.

[4] Mandelbaum, A., & Shalev, A. (2016). Word embedding's and their use in sentence classification tasks.

[5] Hamad, K. A., & Mehmet, K. A. Y. A. (2016). A detailed analysis of optical character recognition technology. International Journal of Applied Mathematics Electronics and Computers, (Special Issue-1), 244-249.

[6] Karthick, K., Ravindrakumar, K. B., Francis, R., & Ilankannan, S. (2019). Steps Involved in Text Recognition and Recent Research in OCR; A Study. International Journal of Recent Technology and Engineering, 8(1), 2277-3878.

[7] Zheng, C., He, G., & Peng, Z. (2015). A Study of Web Information Extraction Technology Based on Beautiful Soup. JCP, 10(6), 381-387.

[8] Ashiwal, P., Tandan, S. R., Tripathi, P., & Miri, R. (2016). Web Information Retrieval Using Python and Beautiful Soup. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 4(6).

[9] Rahiman, M. A., & Rajasree, M. S. (2009, October). A detailed study and analysis of ocr research in south Indian scripts. In 2009 International Conference on Advances in Recent Technologies in Communication and Computing, pp. 31-38.

[10] Shrivastava, V. (2018), A methodical study of web crawler, Journal of Engineering Research and Application, 8(11), 1-8.