International Journal for Modern Trends in Science and Technology, 6(12): 379-383, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061271
Available online at: http://www.ijmtst.com/vol6issue12.html

Natural Language Processing methods for Document Matching

Maitri Patel[1] | Dr Hemant D Vasava[2]

[1]Computer Engineering, Birla VishvakarmaMahavidyalaya
[2] Computer Engineering, Birla VishvakarmaMahavidyalaya

To Cite this Article
Maitri Patel and Dr Hemant D Vasava, "Natural Language Processing methods for Document Matching", *International Journal for Modern Trends in Science and Technology*, 6(12): 379-383, 2020.

Article Info
Received on 16-November-2020, Revised on 09-December-2020, Accepted on 12-December-2020, Published on 18-December-2020.

ABSTRACT

Data,Information or knoweldge,in this rapidly moving and growing world.we can find any kind of information on Internet.And this can be too useful,however for acedemic world too it is useful but along with it plagarism is highly in practice.Which makes orginality of work degrade and fraudly using someones original work and later not acknowleging them is becoming common.And some times teachers or professors could not identify the plagarised information provided.So higher educational systems nowadays use different types of tools to compare.Here we have an idea to match no of different documents like assignments of students to compare with each other to find out, did they copied each other's work?Also an idea to compare ideal answeer sheet of particular subject examination to similar test sheets of students.Idea is to compare and on similarity basis we can rank them.Both approach is one kind and that is to compare documents.To identify plagarism there are many methods used already.So we could compare and develop them if needed.

KEYWORDS:*Document Matching, Plagiarism, NLP, Data Mining, Python.*

I. INTRODUCTION

In this fast growing world with high knowledge and easyness of technology,backdrop comes side by side.Plagarism is one of the common problem in academic world.Students don't do much research and they don't use their intelligence to do their acedemic work.Along with it they even do it smartly and don't cite the original work.And this is common and big issue.To detect plaigairsm is the effective way to avoid it. There are many tools to detect plagarised texts or files. Some are paid and some are free.Usually they compare text or file to information available on internet.Different open/free tools available are Google Search Engine, Dustball, Dupli Checker, Plagiarisma, Academic Plaigiarism, The plaigiarism Checker, Plagiserve.Couple of paid or commercial are EVE2, Plag Aware, Plag Scan, Check for Plagiarism, Plagiarism Detection, Write Check, Turnitin, Ithenticate. Assignments and practicals are one of the common medium where students copy from each other or from internet.Also we can compare and rank answer sheets of students.So we here would like to compare student's assignments between them and not on internet.This is where we had reserched and developed methods in NLP using python.Other side to this is Q/A.We here secondly approach to compare Ideal answer sheet to student's answer sheets and would rank them according to similarity between ideal answer sheet and student's answer sheet.we will check and develop or use combination of different methods for

both apporoches.We would be using different assignments from students and/or text files written on topics by students and available online.[14].

II. Related Work

Different reserchers and developers had worked on it.Document Matching,Automatic assessment and Plagiarism Detection are topics of research.In paper [1] Methods for Identifying Versioned and Plagiarized Documents authors have used different ranking and fingerprinting methods where Cosine method worked very well.In paper [2]authors have used tri-grams sequence matching technique as similarity measure and then clustering using K means to improve latency. In paper [3] Efficiency Comparison of Document Matching Techniques authors Full TAAT,Turtle TAAT, Moffat TAAT, Turtle DAAT etc and they concluded that Turtle TAAT and Moffat TAAT were effective and precision was good compare to Full TAAT.

For Plagirism detection in paper [4] On Automatic Plagiarism Detection Based on n-Grams Comparison,authors have used n-gram(avg word per document) is been used as basic word grouping method and concluded that Bi-Grams and Tri-grams are best for comparision.In paper [5] Automatic Plagiarism Detection Using Similarity Analysis authors have used Similarity analysis methods like cosine, dice, jaccard, hellinger and harmonic to match similarities where cosine was best fitted and gave 90% average accurate result.In Paper [6] Plagiarism Detection Process using Data Mining Techniques authors used Tri-gram and clustering method after preprocessing the documents.In paper [7]Survey of Plagiarism Detection Methods authors surveyed different methods like Grammar-based method , Semantics-based method , Grammar semantics hybrid method which uses natural language text detection, index structure, and external plagiarism detection and clustering –based detection and author suggests use of Semantics-based method for cluster based method as it will achieve much better results.In paper [8]Approaches for Intrinsic and External Plagiarism Detection author used N-grams method and concluded that tri-grams method is perfect for similarity analysis. In paper [9]a study on plagiarism checking with appropriate algorithm in datamining authors used KDD Process in Data Mining,Text mining, naïve bayesian classifier.

For Automatic Assessment in paper [10]Automatic Evaluation of Question Answering System Based on BE Method authors intoduced Bes(Basic Elements) method.In paper [11]New Concepts of Automatic Answer Evaluation in Competence Based Learning authors used Competence based learning , E-learning environment ISC.

Table 1: Different Methods used for document matching

Methods	Dataset Taken	Results	Remarks
N-grams [4]	XML version of the METER corpus	Bigrams favour Recall while trigrams favour Precision	
Full TAAT,Turtle TAAT, Moffat TAAT, Turtle DAAT etc[3]	three different Web IR test collections, related to different domains and timescales	Turtle TAAT and Moffat TAAt were effective and precision were good compared to Full TAAt	
Similarity analysis methods like cosine, dice, jaccard, hellinger and harmonic[5]	Corpus collected from students of college on different topics	cosine measure was able to perform well then other methods	
KDD Process in Data Mining,Text mining, NAÏVE BAYESIAN CLASSIFIER [9]	A news topic is made up of a set of events and is discussed in a sequence of news stories	systems face is the collection of possible sources to compare the suspected documents with	ideal and real sources are not always available, limiting the potential of algorithms that compute similarity

			document-to-document
Competence based learning, E-learning environment ISC [11]		Not detailed result but ranks or grades can be given	algorithms are really complex to be created

III. Proposed Methodology

DATASET

Data set here taken is collection of text files.Where this corpus is of Text documents of assignment given to group of students. Which contains 5 original assignments and 95 assignments from different students. This Corpus is downloaded from [13]. We can even ask students to submit their assignments online and now a days sue to pandemic we see this happening. Assignments and answer sheets are submitted or received through online mediums.

PRE-PROCESSING

Pre-Processing of data consists of transferring taken assignments files and removing stop words and special characters to match them. Also pre-processing we need to define similar words and synonym to that words to detect same meaning words in different text files.

N-grams can also be defined as a method to pre-process text data before applying similarity analysis methods. As n-grams method creates bag of words or group of words to be matched. Word and sentence tokenization are used to create bag of words before we apply actual similarity analysis methods. Here both n-grams and tokenization are NLP concepts.

METHODS USED

Methods researched for development are:

Normalized method like cosine, dice, Jaccard, Hellinger, harmonic etc. Unsupervised topic modelling machine learning NLP library for python known as “Gensim”. For information retrieval (Ranking) and text mining “TF-IDF” is one of the most used method.

We have here by tested all these methods to check whether they do fit in our scenario for both approaches i.e. Approach 1: Assignment/Document Matching and Approach 2: Automatic Assessment

ALGORITHMS

Approach 1: Document Matching

Step1: Collecting dataset

We can ask students to provide assignment documents online.

Corpus available on internet.

Step2: Pre-Processing of Document

Removing Stop Words, stemming(removing suffixes).

N-grams and tokenization.

Step3: Similarity Analysis

Various normalized methods like Cosine, Dice, Jaccard, Hellinger & harmonic can be used to find similarity.

Doc2vec method using Gensim library package.

TF-IDF method.

USE and BERT methods.

Step4: Matrix representation

Pandas data frame to see word frequency.

Sparse matrix or scikit-learn library.

Approach 2: Automatic Assessment

Step1: Collecting dataset

Create Model Answer paper and students answer sheets.

Step2: Matching Techniques

Various normalized methods like Cosine, Dice, Jaccard, Hellinger & harmonic can be used to find similarity.

Doc2vec method using Gensim library package.

TF-IDF method.

USE and BERT methods.

Step3: Ranking/Marking/Grading

Different classification method can be used on basis of matching students answer to model answer.

TOOLS AND TECHNOLOGY

Jupyter Notebook Python Programming editor is used for completing the work at various stages of research and development. Different similarity analysis libraries and methods were used such as sklearn, cosine_similarity, collections, tensorflow, sentence_transformers etc also different Natural Language Processing Libraries were used for programming like nltk were tokenization and genism were developed.

RESULT

However, we need to be more accurate for some measures like similar word meanings, synonyms, tenses sentence are written, graphs, equations, Figures which are mostly used for both types of data sets i.e. Assignments and Answer sheets. But for normal and basic type of assignments and answer sheets as taken into consideration by us, we found cosine from normalized method, Gensim library from unsupervised machine learning method and TF-IDF method from information retrieval gives more accurate result then other methods.

Table 2: **Result of different methods used for matching**

Method	Time Taken to run	Accuracy
Cosine Similarity Analysis	15 min	89-90%
Jaccard Similarity Analysis	13 min	90%
Doc2Vec using gensim	43 min	94-96%
TF-IDF	2 min	90%
USE	62 min	79-80%
BERT	More than 24 hrs	

IV. CONCLUSION

Basic view of this project is to get more accurate result for documents matching and finding similarity between them. And making it available for free and easily accessible to common peoples, students and professors at university and school level. This could be more useful for further research and development process. Here we found few most likely developed and used methods for document matching are Normalized methods like cosine and Jaccard, TF-IDF, Doc2vec which used genism library, USE- Universal Sentence Encoder and BERT- Bidirectional Encoder Representations from Transformers. Where best performance for both approach was by TF-IDF method and in least time. Whereas USE and BERT performed well but took more than 1 hour to run. Cosine and Jaccard gave accuracy but this method was too old and TF-IDF gave similar accuracy.

REFERENCES

[1] Timothy C. Hoad and Justin Zobel.Methods for Identifying Versioned and Plagiarized Documents.journal of the american society for information science and technology, 54(3):203–215, 2003

[2] MAC Jiffriya, MAC Akmal Jahan, Roshan G Ragel and Sampath Deegalla. AntiPlag: Plagiarism Detection on Electronic Submissions of Text Based Assignments. 2013 IEEE 8th International Conference on Industrial and Information Systems, ICIIS 2013, Aug. 18-20, 2013, Sri Lanka

[3] Patrice Lacour, Craig Macdonald, and IadhOunis. Efficiency Comparison of Document Matching Techniques. Patrice Lacour, Craig Macdonald, and IadhOunis. Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK. {lacourp,craigm,ounis}@dcs.gla.ac.uk

[4] Alberto Barr´on-Cede˜no and Paolo Rosso. On Automatic Plagiarism Detection Based on n-Grams Comparison. Alberto Barr´on-Cede˜no and Paolo Rosso. Natural Language Engineering Lab. Dpto. SistemasInform´aticos y Computaci´on, Universidad Polit´ecnica de Valencia, Spain {lbarron,prosso}@dsic.upv.es

[5] Shanmugasundaram Hariharan. Automatic Plagiarism Detection Using Similarity Analysis. The International Arab Journal of Information Technology, Vol. 9, No. 4, July 2012

[6] Mahwish Abid, Muhammad Usman, Muhammad Waleed Ashraf .Plagiarism Detection Process using Data Mining Techniques. https://doi.org/10.3991/ijes.v5i4.7869 Mahwish Abid!!", Muhammad Usman, Muhammad Waleed Ashraf Riphah International University Faisalabad, Pakistan. mahwish.abid15@gmail.com

[7] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel. Survey of Plagiarism Detection Methods. 2011 Fifth Asia Modelling Symposium

[8] Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. Approaches for Intrinsic and External Plagiarism Detection. Notebook for PAN at CLEF2011

[9] Hemalatha A.M, Ms. M. Subha. a study on plagiarism checking with appropriate algorithm in datamining.international journal of research in computer applications and robotics www.ijrcar.com vol.2 issue.11, pg.: 50-58 november 2014

[10] Akiko Yamamoto and Junichi Fukumoto. Automatic Evaluation of Question Answering System Based on BE Method. The 23rd International Technical Conference on Circuits/ Systems, Computers and communications (ITC-CSCC 2008)

[11] Kadri Umbleja, Vello Kukk, Martin Jaanus, Andres Udal. New Concepts of Automatic Answer Evaluation in Competence

Based Learning. 2014 IEEE Global Engineering Education Conference (EDUCON)
[12]https://en.wikipedia.org/wiki/Natural_language_processing
[13]https://ir.shef.ac.uk/cloughie/resources/plagiarism_corpus.html
[14] Mahmoud Nadim Nahas. Survey and Comparison between Plagiarism Detection Tools. American Journal of Data Mining and Knowledge Discovery. Vol. 2, No. 2, 2017, pp. 50-53. doi: 10.11648/j.ajdmkd.20170202.12