Smart Conceptual Appraisal Using Natural Language Processing

Shweta D. Pardeshi[1], Pravin B. Mane[2], Nazneen Inamdar[3], Shubham Nanekar[4],Mahesh S. Shinde[5]

[1,2,3,4,5]*Modern Education Society's College Of Engineering, Dept. of Computer Engineering, Pune, Maharashtra, India*

---***---

Abstract - *In today's teaching Environment, examining every student's level of understanding for a specific topic and determining the approach of teaching takes a lot of time,a system is required which can analyse the level of understanding without human intervention. By using Doc2vec an NLP tool which converts document into numeric form along with cosine similarity to calculate the similarity between two documents we designed a smart conceptual appraisal system that evaluates students' understanding by matching student and teacher synopsis. Our system has 80 to 90% of correctness, which is helpful for students' evaluation depending on certain topics.*

Key Words: Doc2vec, Cosine similarity, Natural Language Processing, Continuous bag of words(CBOW), Skipgram, Tokenization of sentences, Word embeddings, Vector Space.

1. INTRODUCTION

Every teacher has a different style of teaching which can not be understood by all students. Also, it may happen that teachers' teaching style is proper but students are not paying attention. In today's education system colleges tend to have good reports and students want to get good marks, so analyzing every student's understanding level is difficult with human involvement. Because of these problems so many students are interested in online or self learning,they do study from different websites, videos and even they join extra tuitions[1].

Universities also know that they have to improve their teaching techniques but somehow they fail. Universities like Cornerstone find the methods they are using are outdated and start learner-centered criteria for examination and for the success of students. They take individual responses in the form of clicker technique where students can directly take part and immediate action is taken based on this some peer-led education is also started which is an alternative to classical student learning technique. Some teachers with old teaching techniques get frustrated while teaching new techniques and find them difficult. In some cases universities grant permanent leave to the teachers. There have been many attempts to improve the education system like taking feedback from students, teaching online, taking tests on chapters etc.[2]

Teachers are trained by using Bloom's taxonomy where Bloom's taxonomy is a toolbox where teachers or students can easily classify learning objectives based on popular domains and assume that the teaching should be structured from easy to difficult level. Teachers improve their teaching techniques by using a toolbox which is very helpful for students as well as teachers[3].

We designed a system which uses doc2vec and cosine similarity for getting synopsis from students and teacher and finds similarity in percentage form. First the data is taken from the teacher and students in the form of synopsis. Then from the synopsis the important data is extracted and stopwords are removed by using NLTK. NLP plays a very important role because when we write we omit some words or miss punctuation but NLP processes this data and extracts only important data by nominating the error matic data. This useful data is then processed by using doc2vec algorithm for comparing both synopsis by calculating the result. Doc2vec algorithm generates vector space for the synopsis the document/word is distributed in the vector space then after distributing this vector is used to find the similarity between two documents[4]. This vector is given to the cosine similarity algorithm which calculates the similarity and gives the result in the form of percentage[5].

The goals we expect from this study is that it finds similarity between synopsis by using the Doc2vec algorithm to extract meaningful data by eliminating the stopwords and cosine similarity which finds the similarity between two synopsis to give accurate results for betterment of students. In this paper,we mentioned related work in section II, we mentioned the concepts and technologies we used in this paper in section III, we mentioned approach for thighs study in section IV, we mentioned model evaluation and comparison result of synopsis in section V, at last we mentioned the conclusion and future work of this paper.

2. RELATED WORK

The main approach for finding the similarity between two synopsis is cosine similarity[5].

2.1 Automated Essay Scoring with Ontology based on Text Mining and NLTK tools

The authors applied the initial approach for automatically generating the result ontology in essays using OntoGen and applied natural language processing algorithms using NLTK (Natural Language ToolKit) that enhanced the teachers essay grading technique. Authors also analyzed the student essay and gave the similarity result between two documents. Authors have used Support Vector Machine algorithms for extracting important keywords. Cosine similarity is used to find sameness between two vectors. By examining essays from the trained model authors have demonstrated their hypothesis that the number of words the students can disclose contribute to the score of the essay[1].

2.2 A Doc2Vec-Based Assessment of Comments and Its Application to Change-Prone Method Analysis

Author provided a program implementation that can be helpful in antiquity for program comprehension. While there are so many comments which are helpful for the program but also there are many comments which are not that useful. Firstly, A simple document and its composition is converted into a vector. Then, two vectors relating to two different methods are developed, the original vector and the comment-removed vector. Followed by the similarity computation between these two vectors. If the deleted comments provided meaningful information for the main program, the correlated vector would have a big change by removal of the comment. Analysis of the relationship between the worth of comments in a method and the change-liability, using the data collected from five popular open source software projects is done. It can be concluded from the results that a method having less important data which is not informative likely to be changed, i. e. , such a method could not stand unharmed after release[6].

2.3 Sentence similarity measuring by vector space model

Authors provided better methods and techniques for measuring the sentence similarity in a better structural format. Instead of focusing on sentence similarities of two corpus, this paper provides required prior information about WordNet lexical databases and the way we can use it in a context. Authors have provided detailed knowledge about the required basic information on word similarity, which is the basic for the sentence similarity. To get the word similarity capabilities authors have used WS4J library[7].

2.4 Sentence Similarity Based on Semantic Nets and Corpus Statistics

Research shows that semantic similarity apprehends common human understanding as well as it adapts to the domain using the text data i.e. corpus specific to that domain. The given method by authors reviews the impression of order of words on sentence meaning. The obtained word order similarity calculates the count of individual words as well as the number of word pairs in an individual order. The final sentence similarity is defined as a combination of word order similarity and semantic similarity. Author less considered the way that word order plays a role for getting sentence meaning, author reviews word order similarity less important in contributing the general sentence similarity[8].

2.5 Multilingual Inappropriate Text Content Detection System Based on Doc2vec

An inappropriate multilingual text content detection method is proposed based on Neural Machine Translation - Doc2Vec (NMT-D2V) is proposed. NMT-D2V has three features, i. e. , it extends an existing detection system (e. g. based on English) to another language using a multilingual scheme, it improves detection system transparency based on similar text presentation, and it does not required text translation into the original language for each input. An experimental comparison using a Japanese dataset demonstrated that the evaluation index area under the curve of the Receiver Operating Characteristic of NMT-D2V (0. 717) is better than that of afn existing translation method(0. 703)[9].

3. PRELIMINARY CONCEPTS

3.1 Synopsis

Synopsis is nothing but summary. Rather than reading each and every line of the chapter students were assigned to read over many days, it will be more helpful for students if they have been provided a synopsis of what happened.

The meaning of synopsis in ancient greek word is “General view”. In Synonyms document reduction is done, as in a short book and brief, which may be a legal word, sketch, a quick outline of a story. Synonyms can be abstract, compendium, digest.

If a teacher wants to find if a student understands the topic which he/she has taught in lecture he/she can find it by taking synopsi from the student. Synopsis is a short summary of main points of any argument or theory.

3.2 Doc2Vec

Doc2Vec represents documents in a vector space where the semantic data of the documents are highly observed in this model. it is an updated model of word2vec, If documents vectors are close means they are semantically similar documents. The paragraph vector is an alternative to the bag of words model because a bag of words somewhere lacks to consider semantic of the words[11].

Doc2Vec can be implemented in two ways.

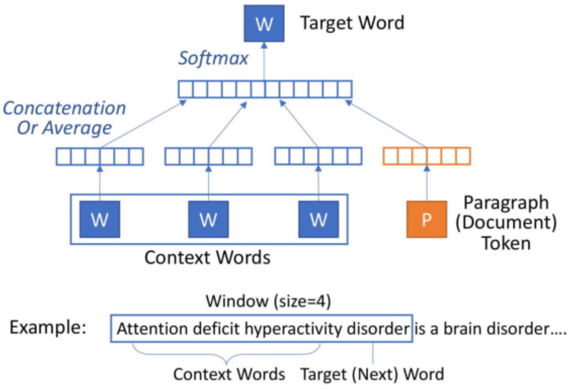
1) Distributed memory model vector

2) Distributed bag of words version of vector

Let's discuss this terms:

1) Distributed memory model vector

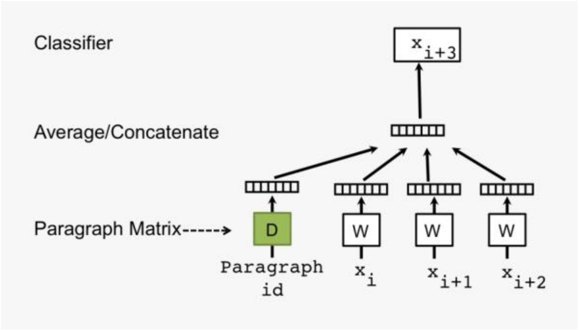
Distributed Memory Model of Paragraph Vectors (PV-DM) is the first technique in determining the Doc2Vec vector representation. In this technique the document's context words are formed and trained in such a way that it will predict the targeted word. Sliding window derives targeted words and context words by using sliding windows of specific length thich pass over start to end. The targeted word is surrounded by context words. Additionally, PV-DM is an addition to the continuous bag of words(CBOW)[11].



The paragraph/document is trained in a neural network model with three layers: the input layer, hidden layer, and output layer. The context words and document/paragraph are given as input; they are at the input layer while the target words are output;they are at the output layer. The hidden layer is the average or concatenation of the context words and document/paragraph together with its certain value. The softmax classifier calculates the probability of the target word from hidden layer and value between hidden layer and output layer, a model that will give us the probability of the target word given the context words and paragraph/document. The goal is to maximize the probability of predicting the target word so back propagation is performed in order to update the value. Stochastic gradient descent is used and the gradient, calculated from the back propagation, is used to update the values between input layer which contains the context words and document and hidden layer which have targeted word, also the weights between the hidden layer and output layer[9][11].

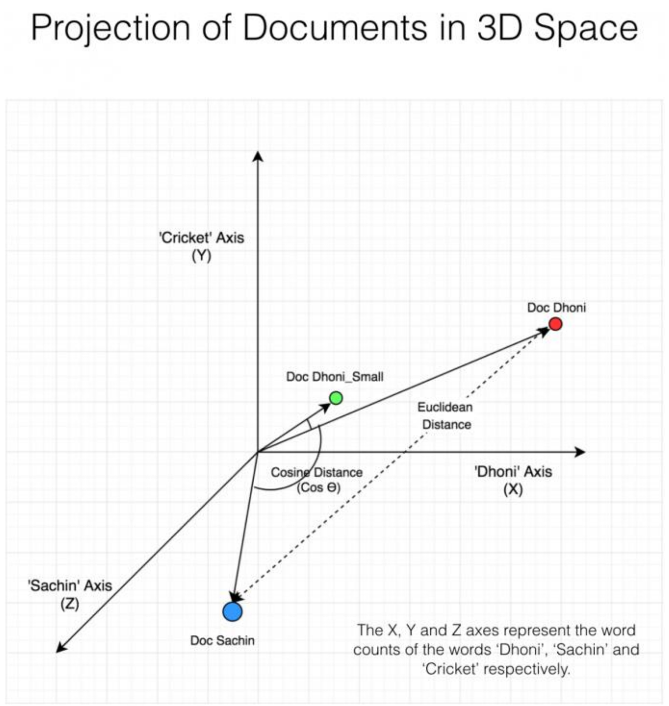
2) Distributed bag of words version of vector

Distributed Bag of Words version of Paragraph Vector is the second technique in developing Doc2Vec which is derived from the skip-gram model of Word2Vec. Skip-gram model works in a reverse way of the CBOW model, target word is predicted in this way. In the neural network model, the document vector is trained in such a way that it can predict the words in the sliding window. At the end the weight is calculated between the hidden layer and the output layer [11].

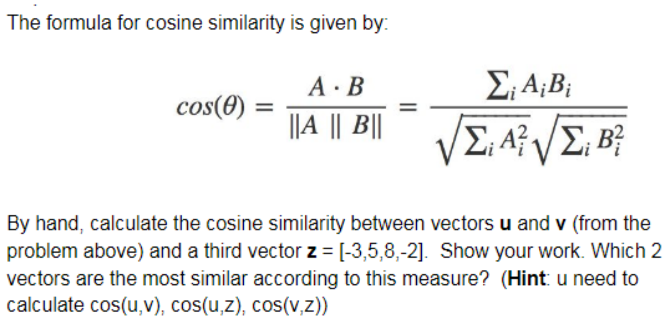


3.3 Cosine Similarity

Cosine similarity is a method which is used to find how two documents are similar irrespective to the document size. It finds the similarity between two vectors from vector space and represents the result numerically. In a multi-dimensional vector space cosine similarity calculates the cosine distance between two vectors. In this case doc2vec converts documents in vectors and uses cosine similarity to determine the similarity between two document vectors.


Projection of Documents in 3D Space
'Cricket' Axis
(Y)
Doc Dhoni
Doc Dhoni_Small
Euclidean
Distance
Cosine Distance
(Cos θ)
'Dhoni' Axis
(X)
'Sachin' Axis
(Z)
Doc Sachin
The X, Y and Z axes represent the word counts of the words 'Dhoni', 'Sachin' and 'Cricket' respectively.

In the above figure, vectors Doc Dhoni and Doc sachin are placed in vector space. These two vectors are far away from each other but looking closer in this 3D space. Similarity between these vectors are calculated by using cosine similarity.



The formula for cosine similarity is given by:

$$cos(\theta) = \frac{A \cdot B}{||A\ ||\ B||} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

By hand, calculate the cosine similarity between vectors **u** and **v** (from the problem above) and a third vector **z** = [-3,5,8,-2]. Show your work. Which 2 vectors are the most similar according to this measure? (**Hint**: u need to calculate cos(u,v), cos(u,z), cos(v,z))

Closer the angle between document vectors, higher the cosine similarity of that document. In this study the similarity scores are greater than 50%.

4. METHODOLOGY

1) Data collection

AI subject dataset is used in this system, where data is present in the form of the paragraph. Web scraping is used to collect data from different sources. Currently the system is designed for the processing of AI subject data.

2) Data preprocessing

There are different NLP techniques which are used for preprocessing of data sets. All the preprocessing techniques used in this study are mentioned below [11].

1. Tokenizing

Tokenization is used for dividing up a sequence of text into parts such as keywords, phrases, symbols, words and other elements that all are called as tokens[7]. It is the process where a given text document is divided into words and punctuation marks. Each unit is known as a token[3][4]. This step is important because an array is the common data structure for documents in NLP since it is more appropriate to count words in this form.

2. Lowercasing

Lowercasing is a process in which all uppercase letters are transformed into lowercase form. This process is very important because the computer treats two words with different cases as different terms[4].

3. Removing Non-alphabetical symbols

Punctuation symbols are removed because they do not contribute any specific meaning to the synopsis' content. In our case,we keep A-Z, a-z, 1-0 other than this every symbol is eliminated from the synopsis[4][11].

4. Removal of Stop words

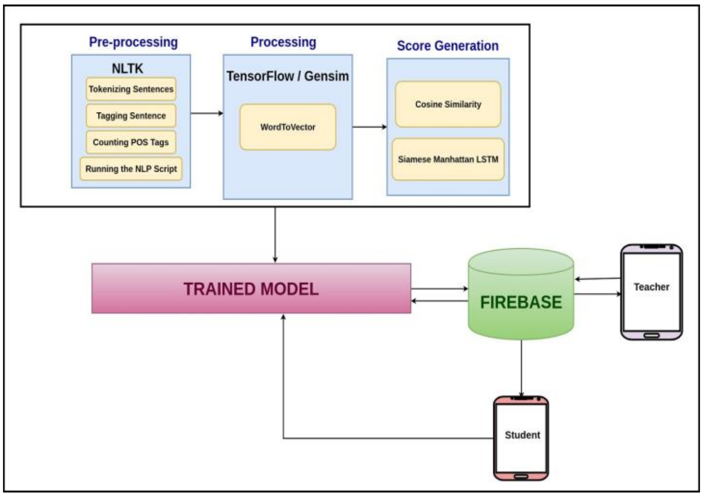
Stop words are a set of commonly used words that naturally appear in every text document but do not provide any notable meaning in the content matching process. This process will reduce the vocabulary size of the corpus leaving only relevant words which provide meaning of sentence[3][4]. Linking verbs, conjunctions, and articulate are some examples of stop words. We can import a collection of English stop words from NLTK and GitHub[10].

5. Stemming

Stemming is a technique for the reduction of words into their root or word stem. many words are deducted into its base form. In stemming words are reduced into their simplest form this is important in Natural Language Processing(NLP) e.g. ran, run, running, and runnable belong to one word run [4][5][6].

3) Doc2Vec Model Training

Model training is done using Gensim Doc2Vec algorithm's implementation.[11] PV-DM and average these two are the default configurations of the model which are also used in this study. A tagged corpus is created by mapping each synopsis with its particular id. The tagged corpus is then imported into this Doc2Vec model and trained for some iterations with given vector size.



5. CONCLUSION AND FUTURE WORK

The similarity check is done between two synopsis using doc2vec which converts the document into vector and cosine similarity finds this similarity in the form of percentage.

Checking every student's level of understanding for a particular topic and deciding the approach of teaching consumes a lot of time. There is a need of assessment for level of understanding without manual evaluation for every student, a system is required which can analyse the level of understanding without human intervention. Smart Conceptual Appraisal Using Natural Language Processing, provides this platform where evaluation of individual students can be done according to the synopsis it provides to the model. The model is trained using different Natural Language Processing tools. and algorithms like Natural Language Toolkit,Doc2Vec, Cosine Similarity which l Using Natural Language Processing is beneficial to students as well, if they are not understanding the teaching method they can indirectly ask the teacher to change it through this system.

This concept can be widely used in different areas of education as assessment for level of understanding is gaining importance.

REFERENCES

[1] Jennifer O. Contreras, Shadi Hilles, Zainab Binti Abubakar, "Automated Essay Scoring with Ontology based on Text Mining and NLTK tools", 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), July 2018.

[2] Jason Allaire, Cornerstone University, "Five Issues Facing Higher Education", https://www.cornerstone.edu/blogs/lifelong-learning-matters/post/five-issues-facing-higher-education-in-2018, January 15 2018.

[3] Nancy E. Adams, "Bloom's Taxonomy Of Cognitive Learning Objectives", Journal of the Medical Library Association, 2015 Jul; 103(3): 152–153.

[4] Lorenz Timothy Barco Ranera, Geoffrey A. Solano, Nathaniel Oco, "Retrieval of Semantically Similar Philippine Supreme Court Case Decisions using Doc2Vec", 2019 International Symposium on Multimedia and Communication Technology (ISMAC), IEEE, January 2020.

[5] Dewi Soyusiawaty, Yahya Zakaria, "Book Data Content Similarity Detector With Cosine Similarity", 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA), IEEE, May 2019.

[6] Hirohisa Aman, Sousuke Amasaki, Tomoyuki Yokogawa, Minoru Kawahara, "A Doc2Vec-Based Assessment of Comments and Its Application to Change-Prone Method Analysis", The 25th Asia-Pacific Software Engineering Conference (APSEC 2018), At Nara, Japan, December 2018.

[7] U. L. D. N. Gunasinghe, W. A. M. De Silva, N. H. N. D. de Silva, A. S. Perera, "Sentence Similarity Measuring by Vector Space Model", 2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer), December 2014.

[8] Y. Li, D. McLean, Z.A. Bandar, J.D. O'Shea, K. Crockett, "Sentence Similarity Based On Semantic Nets and Corpus Statistics", IEEE Transactions on Knowledge and Data Engineering, Volume: 18, Issue: 8 , Aug. 2006 .

[9] Kazuki Aikawa, Shin Kawai, Hajime Nobuhara, "Multilingual Inappropriate Text Content Detection System Based on Doc2vec", 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), Oct 2019.

[10] Sean Bleier, "NLTK's List of English Stopwords", https://gist.github.com/sebleier/554280, Aug 2010.

[11] Petros Karvelis, Dimitris Gavrilis, George Georgoulas, Chrysostomos Stylios, "Topic Recommendation using Doc2Vec", 2018 International Joint Conference on Neural Networks (IJCNN), July 2018.