International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT20639

Candidate Feature Extraction and Categorization for Unstructured Text Document

Prof. Prajakta P Shelke, Aditya A Pardeshi
Department of Computer Science and Engineering, Government College of Engineering, Amravati, India

ABSTRACT

In the phrases words contains crucial information which helps in feature extraction process. The established techniques for such has huge problem and has limitations in feature extraction process and also it ignores the grammatical structure for the phrases. So results as poor features get extracted. So to overcome this problem a system is proposed which is based on generation of parse tree for the input sentence and cut down into sub-tree subsequently. The branches of the tree are extracted using part-of-speech (POS) labelling intended for candidate phrase. To stay away from redundant phrases filtering is recommended. Finally machine learning is used for the Feature categorization progression. The result illustrates the effectiveness of the approach.

Keywords : Key Phrase Mining, Candidate Feature Mining, Feature Selection, Feature Classification, Natural Language Processing.

I. INTRODUCTION

In the today's world every person uses the electronic devices and internet on it. So the data produced by such gradually requires very faster updates. In lots of aspects such as phrase mining, document categorization the automatic key-phrase identification is an issue. The various methods are projected within the literature section to tackle such issue which is mainly classified within firstly supervised and secondly unsupervised learning method. The candidate feature choice persist an essential job in each projected result, and consequently effectiveness in addition to strength as concerns the method mainly based on the fundamental candidate feature mining method.

The n gram and noun phrase based are some approaches which are used in the candidate feature mining. This method has some difficulty regarding the catalogue of candidate phrases. The difficulty in the n gram based technique is of range concerning to n gram which is limited up to certain range and the other is the candidate phrases are likely to be grammatically inaccurate, also they do not all the times grasps complete information [1].Similarly in the noun phrase based technique the noun phrase may able to be a singular noun either the cluster regarding words this may outline noun and include complexity such no more each noun can be key phrase and equally, there may be phrases other than noun that are ultimately key phrases or it may be a part of an key phrase.

The whole scenario for input expression is shown while developing the parse tree for the input expression. This depicts in broad composition as well as association with different units of the input expression. In [2] the example of an input expression is specified, which shows the POS tag of words and furthermore inter linking among different parts of an expression. So therefore, after analysing the expression parse-tree and then mining the most considerable candidate phrase which is been based upon the parsing method and which is more capable as compare to the POS tagging technique. Moreover, when assessment regarding absolute sentence format is being carried out, inspection can be done that the input sentence may tag as simple sentence, compound sentence or else complex sentences. The simple sentences which have single clause; this clause inside the language comprise the two segments; one is noun phrase beside with verb phrase. The compound sentences in an English language which consists of the two or more-clause along with the coordinating conjunctions. The complex sentence in an English language which contains a particular main clause along with one either supplementary adverbial clause tied with auxiliary conjunction [3]. Therefore, clause composition is the elementary unit of the sentence which offer the total information regarding the key concept. Generally clauses may consist noun phrases, verb phrases and prepositional phrases. So, in the favour of capable candidate feature mining the three types about phrases must planned consecutively rather than a noun phrase only as well as then best work be able to attain by analysing an input sentence in the composition relating to the parse-tree.

II. RELATED WORK

In almost all the approaches for the key concept identification the candidate feature extraction is a common task. For such most of them depends on the two conventional techniques n gram and noun phrase based approaches. The techniques which is based on the n grams [4]-[9] aims at reaching maximum recall for candidate feature extraction. As we have discussed in above section that is of length of the n gram is restricted to an extend and also they are mostly grammatically incorrect as they do not always capture the complete information [1]. So for alternative to this problem the solution is to use POS tags or noun phrases [1], [10]-[16]. This follows the Hulth observation that mostly the key phrases are to be a noun phrases [17]. The pre-processing is the first step in this algorithm which depends on POS tags. These tags are for each word in an input sentence and therefore makes easy for fetching the noun phrases as a candidate feature. Most of the techniques are based on these dual approaches in favour of candidate feature mining.. The [18] employ lots of sentence constraint method which is located over the phrase network [19], [20], intended for obtaining the preliminary candidate feature, as, s some call it as compression candidate. The directed word graph for given wording are constructed using this approach. Where the node in a graph depicts the sole word and edges depicts composition concerning genuine lexemes which is phrase arrangement. Compression candidates are retrieved using common paths which is present in a graph. Another approach is word expansion for candidate feature extraction [21]. This approach firstly generates the set of nucleus words that helps in finding the logical location for the potential key phrases. So after this the method which is depending over the core lexeme extension tree is to be applicable and which construct the candidate features relative to such position. This approach entitles the highest concerning the dual candidates can be produced as of every occurrence of this core lexeme. A few may use heuristic approach [22], [23].

In [24] and in [25] they utilize narrative knowledge for key phrase mining and choose candidate phrase which is respected to the Wikipedia by means of the Wikipedia Miner to map among an specified input text as well as the Wikipedia entities. The [26] illustrate an approach that depends on noun-phrase technique with an order concerning POS tag's linguistic pattern. This simply picks the phrases which contains nil otherwise additional adjectives as well as coming with single or supplementary noun. In [27] Wu et al. Propose a probabilistic technique for the keyword mining which then motivated with visual consideration method. The probability-based scores are given for each keyword which is measured like candidate features. The [28] acclimatised the circulated depiction concerning the lexemes for mining the phrases used in favour of conspicuous categorization. Here the lexemes which may use like a candidate feature is assign class which depends over a cosine similarity concerning that candidate lexeme by a centroid lexeme of a specified document.

III. PROPOSED METHODOLOGY

Following are the steps of system model as shown in Figure. 1 :

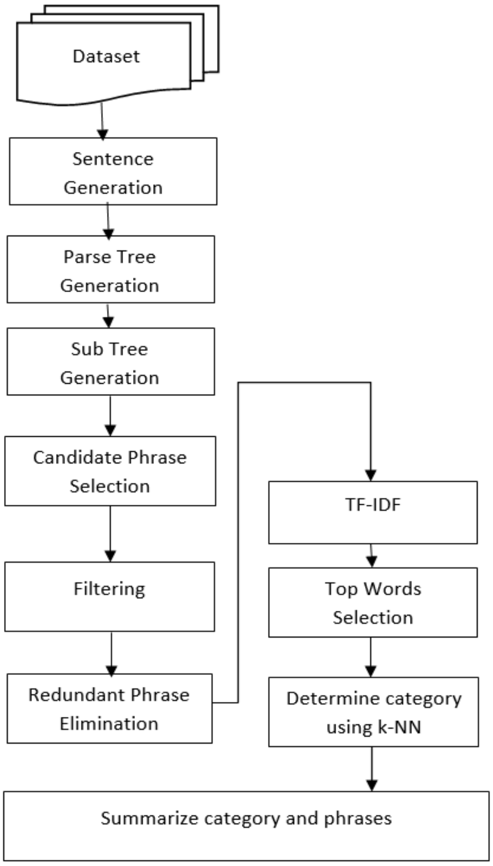
Dataset
Sentence Generation
Parse Tree Generation
Sub Tree Generation
Candidate Phrase Selection
Filtering
Redundant Phrase Elimination
TF-IDF
Top Words Selection
Determine category using k-NN
Summarize category and phrases

Figure. 1. Proposed System.

3.1. DATASET

The dataset will be the input to our technique. Our technique will take data set which will be unstructured text document. The typical example of a dataset which is unstructured is given in the [29] which consists of news stories and respective sources of them and also another Inspec database which consists of 2000 abstracts is used in [30].

3.2. SENTENCE PARSE TREE AND SUB TREE GENERATION

The unstructured text document which is our input is firstly divided into the catalogue of sentences. Then the divided catalogue of sentences is given to the sentence parser, then for each and every sentence it generates the parse tree and a catalogue of a corresponding parse tree is returned. For this purpose we make use of the Stanford Parser [31]. The catalogue of this tree structure that we have got in this step is then given to the next step which is of candidate phrase selection for the further analysis to get more meaningful features.

3.3. CANDIDATE PHRASE SELECTION

After parsing the document the next step is of generating the comprehensive catalogue of a candidate phrases. In this the significant list of candidate features is returned in the course of mining element in a sub-tree which will be meaningful. For this the sub tree is branched like it is noun, verb or prepositional phrases and then the leaves of sub tree are coupled and naming to it is given either as Noun Phrase (NP) or Verb Phrase (VP) to gain the candidate phrase. For this purpose the catalogue of POS tag are too extracted commencing the sub tree which is used in the subsequent step. The POS tag-sets are been used comprehensively for mining of the features from the sub-tree branches. In the stage of the POS assignment the POS tags are allocated to the branches using the cascade of stochastic and rule-driven taggers and accordingly the output is automatically tokenized. Within [32] a range of POS tags or the catalogue is shown for the POS tags on which the tagging of the branches of sub-tree is based like for Coordinating conjunction CC is used, for Cardinal number CD is used, for Interjection UH is used respectively which will be used during our work. The limited number of POS tag-sets are been used just to eliminate redundancy by taking into account of lexical as well as syntactic information. To keep away from redundancy tagging similar to for do--the base form (DO), the past tense (DOD), and the third person singular present (DOZ) is done. The redundancy can be removed by using such type of tags for other like be, was etc. For obtaining our final catalogue of features filtering process is carried out in the next succeeding step.

3.4. FILTERING

This process has huge contribution in the performance, so this is very important step. In this step the catalogue of the candidate features which are achieved in the previous step are filtered by means of various heuristic regulations. Some of the regulations are of type that it should not finish with stop word or it should not have any punctuation mark in it. By concerning such rules, the catalogue of desired candidate features can be gained.

3.5. REDUNDANT PHRASE ELIMINATION

Our filtered list which is obtained from the filtering step may contain some redundant phrases as one phrase may be the part of another phrase. So, in this step our focus is on removing such redundant phrases to obtain our final catalogue of candidate feature. In this the elimination of the phrases which are overlapped, or part of other phrases is done. Such as if one Verb phrase it is been part of another verb phrase in same sentence then remove it. Similarly, the phrase which is noun is eliminated in a similar manner. So by performing this redundancy can be eliminated.

3.6. FEATURE SELECTION AND FEATURE CATEGORIZATION

As we get the catalogue of the candidate features the top words need to be selected. So in the Feature selection process the top words are been selected. So for this we extensively use the Term frequency inverse document frequency (Tf-Idf) algorithm. This Tf-Idf is combination of the two different words that is term frequency (TF) and the inverse document frequency (IDF). TF(w) = (Number of times term w appears in a document) / (Total number of terms in the document) Consider a document containing 100 words where in the word 'Cauvery' appears 3 times.

TF = 3/100 = 0.03. (1)

IDF(w) = log_e(Total number of documents / Number of documents with term w in it)

Now, assume we have 10 million documents and the word 'Cauvery' appears in 1000 of these.

IDF = log_e (10,000,000/1000) = 4. (2)

Thus, the Tf-Idf weight is the product of these quantities TF-IDF = 0.03 * 4 = 0.12. (3)

So by calculating such score the top words are been selected. After the top word gets selected then there is need to categorize relevant to the class accordingly. So finally feature categorization process is carried out for text classification which requires training a model which is based on previously classified document. Then the text classifier is able to build and used in favour of predicting the class label of fresh document as per their content. So for this the k-Nearest Neighbour (k-NN) algorithm is used. K-Nearest Neighbour is used mainly when all the attributes are continues .Simple K-Nearest Neighbour algorithm have following steps:

Steps 1) find the K training instances which are closest to unknown instance.

Steps2) pick the most commonly occurring classification for these K instances.

There are various ways of measuring the similarity between two instances with n attribute values. Every measure has the following three requirements. Let Dist(A,B) be the distance between two points A,B then,

1) Dist(A,B)≥0 and dist(A,B)=0 if A=B
2) Dist(A,B)= Dist(B,A)
3) Dist(A,C)≤ Dist(A,B)+ Dist(B,C)

Property 3 is called as “Triangle in equality”.
Thus the top word get categorised and thus the final desired candidate feature is obtained.

IV. RESULT

The result shows the overall performance of the system for the candidate feature extraction and the categorization process. The graph for the parameters Precision, Recall, F-Measure and accuracy are shown below.

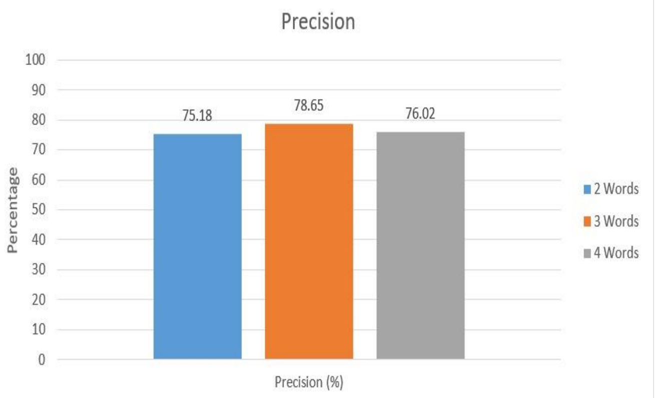

Figure. 2. Precision.

The 2 Words, 3 Words and 4 Words are the extracted candidate phrases. The above figure shows that the 3 Words have the highest Precision of 78.65 than the 2 Words and 4 Words.

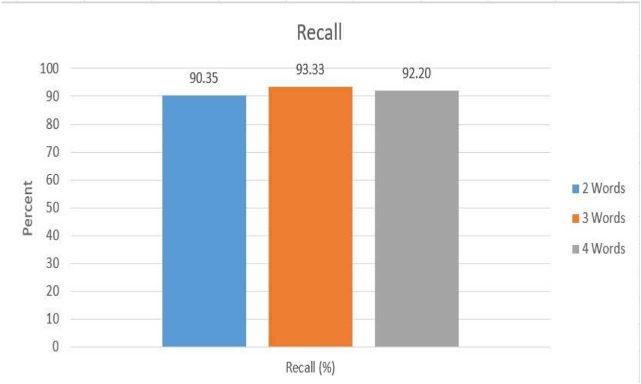

Figure. 3. Recall.

The 2 Words, 3 Words and 4 Words are the extracted candidate phrases. The above figure shows that the 3 Words have the highest Recall of 93.33 than the 2 Words and 4 Words.

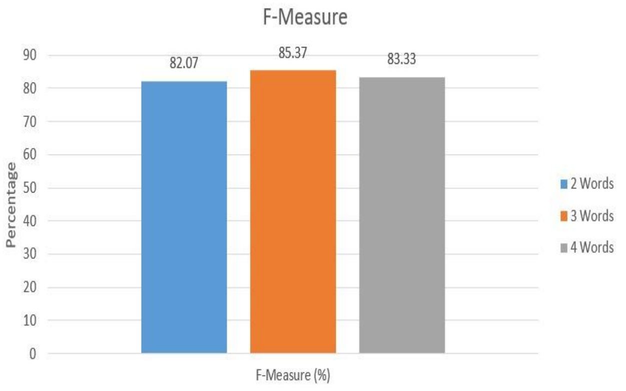

Figure. 4. F-Measure.

The 2 Words, 3 Words and 4 Words are the extracted candidate phrases. The above figure shows that the 3 Words have the highest F-Measure of 85.37 than the 2 Words and 4 Words.

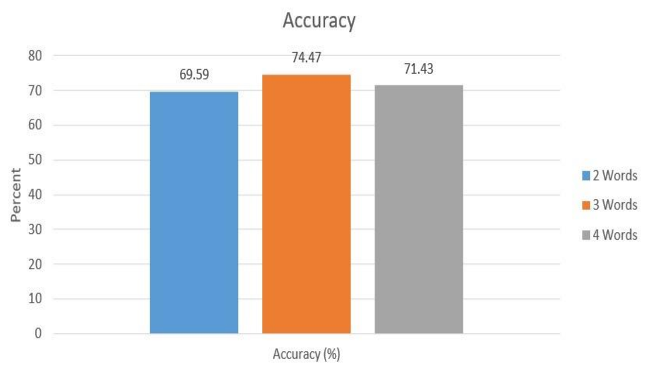

Figure. 5. Accuracy.

The 2 Words, 3 Words and 4 Words are the extracted candidate phrases. The above figure shows that the 3 Words have the highest Accuracy of 74.47 than the 2 Words and 4 Words.

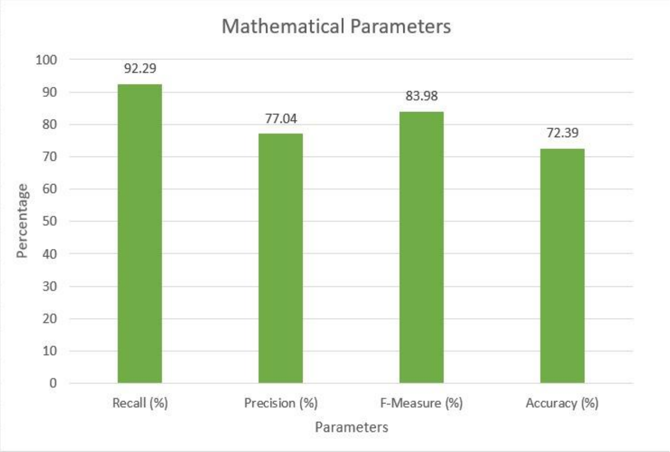

Figure. 6. Mathematical Parameters.

The above figure shows the overall percentages of the mathematical parameters, Recall, Precision, F-Measure and Accuracy.

V. CONCLUSION

This paper presents an extensive method to mine the candidate phrases from input unstructured document text. The propose technique makes the use of the parse tree and sub tree generation for analysing the sentence structure for feature mining. This sub-tree is branched using the POS tags and using this POS tags the wide-ranging catalogue of the candidate features are extracted. Then the top words are been selected using the Tf-Idf algorithm for further classification of the features. The k-NN methodology is used for the feature categorization process which categorizes the top words relevant to the class. So, the result shows the performance of the system for feature selection and categorization process.

VI. REFERENCES

[1]. A. Bougouin, F. Boudin, and B. Daille, "TopicRank: Graph based topic ranking for keyphrase extraction," in Proc. Int. Joint Conf. Natural Lang. Process. (IJCNLP), 2013, pp. 543–551.

[2]. (2017). The Stanford Parser. Accessed: May 2, 2017. Online. Available: https://nlp.stanford.edu/software/lex-parser.html

[3]. British Council. (2017). Learn English. Accessed: Dec. 30, 2017.Online.Available:https://learnenglish.britishcouncil.org/en/english-grammar/clause-phrase-and sentence/sente-nce-structure

[4]. M.-S. Paukkeri, I. T. Nieminen, M. Pöllä, and T. Honkela, "A language-independent approach to keyphrase extraction and evaluation," in Proc. Coling Companion, 2008, pp. 83–86.

[5]. S.R.El-BeltagyandA.Rafea,"KP-Miner Akeyphraseextrac-tionsystem for English and Arabic documents," Inf. Syst., vol. 34, no. 1, pp. 132–144, 2009.

[6]. O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in Proc. Conf. Empirical Methods Natural Lang. Process., vol. 3, 2009, pp. 1318–1327.

[7]. K. S. Nam, M. Olena, K. Min-Yen, and B. Timothy, "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles," in Proc. 5th Int. Workshop Semantic Eval., 2010, pp. 21–26.

[8]. S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "Automatic keyphraseextractionfromscientificarticles,"L-ang.Resour.Eval.,vol.47, no. 3, pp. 723–742, 2013.

[9]. S.Danesh,T.Sumner,andJ.H.Martin,"Sgrank:Combiningstatisticaandgraphicalmethodstoimprovethestateoftheartinunsupervisedkeyphrase extraction," in Proc. SEM NAACL-HLT, 2015, pp. 117–126.

[10]. F. Boudin, "A comparison of centrality measures for graph-based keyphrase extraction," in Proc. Int. Joint Conf. Natural Lang. Process. (IJCNLP), 2013, pp. 834–838.

[11]. Y.-B. Kang, P. D. Haghighi, and F. Burstein, "CFinder: An intelligent key concept finder from text for ontology development," Expert Syst. Appl., vol. 41, no. 9, pp. 4494–4504, 2014.

[12]. Z.Liu,W.Huang,Y.Zheng,andM.Sun,"Automatickeyphrase-extractionviatopicdecomposition,"inProc.Conf.EmpiricalMethodsNaturalLang. Process., 2010, pp. 366–376.

[13]. J. Martinez-Romo, L. Araujo, and A. D. Fernandez, "Semgraph: Extract- ing keyphrases following a novel semantic graph-based approach," J. Assoc. Inf. Sci. Technol., vol. 67, no. 1, pp. 71–82, 2016.

[14]. N. Teneva and W. Cheng, "Salience rank: Efficient keyphrase extraction with topic modeling," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 2, 2017, pp. 530–535.

[15]. C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2017, pp. 1105–1115.

[16]. J. Rafiei-Asl and A. Nickabadi, "TSAKE: A topical and structural auto- matic keyphrase extractor,"

Appl. Soft Comput., vol. 58, pp. 620–630, Sep. 2017.

[17]. A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in Proc. Conf. Empirical Methods Natural Lang. Process., 2003, pp. 216–223.

[18]. F. Boudin and E. Morin, "Keyphrase extraction for n-best reranking in multi-sentence compression," in Proc. North Amer. Chapter Assoc. Com- put. Linguistics (NAACL), 2013, pp. 1–9.

[19]. R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," Comput. Linguistics, vol. 31, no. 3, pp. 297–328, 2005.

[20]. K. Filippova and M. Strube, "Sentence fusion via dependency graph compression," in Proc. Conf. Empirical Methods Natural Lang. Process., 2008, pp. 177–185.

[21]. W. You, D. Fontaine, and J.-P. Barthés, "An automatic keyphrase extrac- tion system for scientific documents," Knowl. Inf. Syst., vol. 34, no. 3, pp. 691–724, 2013.

[22]. D. Newman, N. Koilada, J. H. Lau, and T. Baldwin, "Bayesian text segmentation for index term identification and keyphrase extraction," in Proc. COLING, 2012, pp. 2077–2092.

[23]. C. Huang, Y. Tian, Z. Zhou, C. X. Ling, and T. Huang, "Keyphrase extraction using semantic networks structure analysis," in Proc. 6th Int. Conf. Data Mining (ICDM), Dec. 2006, pp. 275–284.

[24]. F.Wang,Z.Wang,S.Wang,andZ.Li,"Exploiting description knowledge for keyphrase extraction," in Proc. Pacific Rim Int. Conf. Artif. Intell., 2014, pp. 130–142.

[25]. H.Zheng,Z.Li,S.Wang,Z.Yan,andJ.Zhou,"Aggregating -inter-sentence information to enhance relation extraction," in Proc. AAAI, 2016, pp. 3108–3115.

[26]. K. Bennani-Smires, C. Musat, M. Jaggi, A. Hossmann, and M. Baeriswyl. (2018). "EmbedRank: Unsupervised keyphrase extraction using sentence embeddings." Online. Available: https://arxiv.org/abs/1801.04470

[27]. X. Wu, Z. Du, and Y. Guo, "A visual attention-based keyword extraction for document classification," Multimedia Tools Appl., vol. 77, no. 19, pp. 25355–25367, 2018.

[28]. J.Hu,S.Li,Y.Yao,L.Yu,G.Yang,andJ.Hu,"Patent keyword extraction algorithm based on distributed representation for patent classification," Entropy, vol. 20, no. 2, p. 104, 2018.

[29]. L. Marujo, A. Gershman, J. Carbonell, R. Frederking, and J. P. Neto. (2013). "Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization." Online. Available: https://arxiv.org/abs/1306.4886

[30]. R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in Proc. Conf. Empirical Methods Natural Lang. Process., 2004, pp. 1–8.

[31]. D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in Proc. 41st Annu. Meeting Assoc. Comput. Linguistics, 2003, pp. 423–430.

[32]. M.P.MarcusandM.A.Marcinkiewicz,andB.Santorini,"Bui-lding a large annotated corpus of English: The penn treebank," Comput. Linguistics, vol. 19, no. 2, pp. 313–330, 1993.

Cite this article as :

Prof. Prajakta P Shelke, Aditya A Pardeshi, "Candidate Feature Extraction and Categorization for Unstructured Text Document", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 3, pp. 81-87, May-June 2020. Available at doi : https://doi.org/10.32628/CSEIT20639
Journal URL : http://ijsrcseit.com/CSEIT20639