Analysis of Parts of Speech Tagging on Text Clustering

Y. Sri Lalitha, J Sirisha Devi, L. Sukanya, N.V. Ganapathi Raju,

Abstract: Clustering is a machine intelligence which aimed at grouping a set of objects into Subsets or clusters. Clustering text documents into various classifications is a vital advance in indexing, recovery, administration and removal of abundant text data on the Web. In research and development to prove that a new clustering algorithm is efficient, one needs to compare the existing algorithm with the new technique, for which the standard datasets are required. In this paper we have pre-processed the datasets to a standardized format, with an expansion of houses appropriate for a wide range of clustering and related experiments. Our objective is to set up a benchmark document datasets and extract the parts of speech such as verbs, nouns, adverbs, adjectives and etc from the documents of a given dataset and analyze the impact of parts of speech in clustering process.

Index Terms*: Text Preprocessing, *POS Tagging, Vocabulary, Clustering

I. INTRODUCTION

Cluster evaluation seeks to divide a fixed of objects right into a small number of rather homogeneous corporations on the idea of their similarity over N variables.Cluster analysis may be considered both as a means of summarizing a data set or as a way of constructing a topology. Patterns within a legitimate collection are extra homogeneous to each aside from to a pattern belonging to a different cluster.

Clustering is beneficial in several prefatory pattern-analyses, selection-making, records mining, document retrieval, picture segmentation and pattern category. The end result of a clustering may be useful in many particular instances. It is able to be used as an intermediate step in a larger gadget or as a tool in its very own proper; to facilitate exploration of includes searching for engine or for any text set. The stop result of clustering algorithms is counting on a definition of a (dis)similarity most of the devices. For text clustering the similarity is generally described through an example of the texts the use of a few or all of the terms/tokens that appear in them. Texts are typically described as similar in the event that they use the equal words. Which terms/tokens this is used and the way they are preprocessed should have a awesome impact at the result. Lemmatization or stemming permits us to treat numerous associated tokens because of the fact the same, essential to an elevated similarity among texts, the usage of the outstanding styles of a phrase.

Element-of-Speech (EoS) tagging can be used to obtain the other; separate homographs in order that texts are not described comparable while they'll be the usage of the exclusive meanings of a token..

II. RELATED WORK

This work has two major steps before applying text clustering, they being Preprocessing and POS tagging. The related works on these two methods are elaborated in this section. Firstly, many attempts have been done on preprocessing the documents so far. Preprocessing of any dataset typically includes the techniques that are mentioned below.

A. Preprocessing Techniques

Preprocessing takes plain textual content record as input and outputs a fixed of tokens (which may be single terms or n-grams) for use in vector space representations of textual content files. getting rid of special characters and punctuation that are not concept to keep any discriminative strength below the vector space model is named as Filtering. That is greater reproving within the case of formatted documents, consisting of net pages, where formatting tags can either be remove or identified and their factor phrases attributed exceptional weights. Splitting sentences into character tokens is known as tokenization. Extra sophisticated methods, drawn from the field of NLP, parse the grammatical structure of the text to pick out excellent terms or chunks (sequences of words), such as noun terms[1]. The procedure of decrease words to their base shape, or stem is known as stemming. For instance, the words “givinggg,” “given”, “gives” are all decreased to the stem “give”. A forestall phrase is defined as a term which isn't notion to bring any meaning as a size in the vector area (i.e. without context). An ordinary technique to get rid of stop phrase is to compare every term with a compilation of recognized stop words. Some other method is to first follow a part-of-speech computing after which reject all tokens that is not nouns, verbs, or adjectives. Doing away with of words that appears in low frequency within the dataset is termed as Pruning. The underlying assumption is that those phrases, even though they had, are least insightful, could shape too small clusters to be beneficial. A pre-specified threshold is usually used, e.g. a small fraction of the quantity of words inside the corpus. On occasion words which occur too frequently (e.g. in 40% or extra of the documents) are also eliminated as they've very less discriminating feature and won't be helpful in clustering[1]. In addition to

Revised Manuscript Received on June13, 2019.

Dr. Y. Sri Lalitha,Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

Dr. J Sirisha Devi, Department of Computer Science and Engineering, Institute of Aeronautical Engineering, India.

L. Sukanya, Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

Dr. N. Ganapathi Raju, Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

the common steps above, a lexical database, WordNet, postaggers have been used to conquer phrases that specify common subject matter but use one-of-a-kind terminology (synonymy, hypernym) standards and introduce a more standard standards[6].

B. Parts of Speech Tagging

Part of speech tagging is the manner of redecorating or "tagging" phrases in an exit with every word's corresponding part of speech. A part of speech tagging is based totally each on the means of the word and its positional courting with adjacent words. Components of speech tagging is an utility in natural Language processing (NLP) that enables us to perceive the part of speech of each phrase a good way to get rid of the non-applicable data within the clustering system. In clustering methods, Parts of speech tagging are in main use as the first step of pre-processing of a document. It is a fact that, the text shape and morphological variations help us to decide the appropriate part-of-speech. For this reason, if it's miles required, POS tagging is step one to be achieved. After this, forestall-word removal is achieved, followed by way of stemming. This order is selected to reduce the amount of phrases to be stemmed. Those stemmed phrases similarly are used within the clustering technique.

C. Classification of POS Taggers

i. *Supervised Models :* The supervised POS Tagging fashions require a reannotated corpus which is used for education to learn statistics about the tagset, word-tag frequencies, and rule sets, and so on. The overall performance of the fashions mechanically rises with boom in the size of the corpus[2].

ii. *Unsupervised Models:* The unsupervised POS Tagging fashions do not want a pre-annotated corpus. As a substitute, they use advanced computational techniques like the Baum Welch algorithm to robotically result in targets, transformation policies, and so forth. Primarily based on these records, they either calculate the probabilistic data needed by way of the stochastic taggers or result in the contextual policies wanted by means of rule primarily based systems or transformation based totally structures [6, 7]. Each the supervised and unsupervised fashions may be further categorized into the subsequent classes[3].

iii. *Rule Based and Transformation Based Models :* The rule of thumb based POS Tagging methodology follow a set of handwritten rules and use contextual data to allocate POS tags to words. Those guidelines also are called context frame guidelines. Alternatively, the transformation based totally strategies use a predefined set of handmade regulations in addition to mechanically-triggered guidelines which might be generated all through training. a few models also utilize morphological rules, capitalization and punctuation, and many others.

iv. *Stochastic Models:* The stochastic fashions consist of frequency, chance or facts. They may be based on special techniques such as n-grams, most-chance estimation (MLE) or Hidden Markov fashions (HMM)[5]. HMM-based method calls for assessment of the argmax method, which may be very exorbitant as all viable tag series ought to be checked, that allows you to discover the collection that multiply the chance. So a dynamic programming approach referred to as the Viterbi algorithm is used to find the most advantageous tag collection. There are several research which making use of unsupervised gaining knowledge of method for schooling a HMM for POS Tagging. The maximum idely regarded is the Baum-Welch set of rules, which may be used to educate a HMM from un-annotated statistics. And in the end, both supervised and unsupervised POS Tagging models can be primarily based on neural networks[5].

v. *Stanford's Part Of Speech Tagger :* Stanford's Pos Tagger uses maximum entropy-based totally a part of speech approach, which achieves advanced overall performance basically via enriching the statistics sources used for tagging. specially, They were given progressed outcomes via incorporating these features:

1. More significant remedy of capitalization for new features
2. Features for the disambiguation of the tense form of verbs;
3. Features for disambiguating particles from prepositions and adverbs.

This assigns a probability for each tag t in the set T of possible tags given a phrase and its context h, which is normally defined as the series of many phrases and tags preceding the word. This model may be used for estimating the chance of a tag series $t_1…t_n$ given a P($t_1....t_n$ | $w_1...w_n$)[5][3].

vi. *QTag :* QTag is a language unbiased, natural probabilistic element-of-speech tagger.[4] The part of speech tagger QTag is primarily based on a Hidden Markov model and is freely to be had for individual research and private non-commercial purposes. QTAG is a probabilistic components-of-speech tagger. meaning it's a software that reads text and for each token within the text returns the component-of-speech (eg noun, verb, punctuation, and so forth). It works the usage of statistical techniques, for this reason the `probabilistic'. As a result it does make errors (as does each POS tagger), but it is reasonably robust and (from casual assessment) tags texts with properly accuracy.

This section has three steps, preprocessing, pos tagging and clustering. In this section we discuss the steps considered in preprocessing a dataset, the process of extracting different parts of speech from the documents, and the application of clustering algorithms.

A. Preprocessing

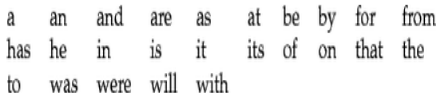
i. Labelling to the files: To determine the documents of high similarity are grouped together we labelled the documents with appropriate names. This enables us to determine the clustering accuracy with a simple entropy measure.

ii. Filtering: It is an activity of separating special characters and punctuation that are not notion to maintain any discriminative power. That is important in the case of formatted files, inclusive of net pages, where formatting tags can both be discarded or identified and removed.

iii. Computing Average File Size: All the files in a dataset need not be of same size. Some files may contain only one word and some may be empty. These files are not useful when clustering the documents. So we need to remove such files from dataset. To determine the file size the average file size of the dataset is considered and discarded the files less than this threshold.

iv. Tokenisation: Tokenization is the system of breaking text into terms or different meaningful factors called tokens. The unique tokens derived from the collection of datasets are used further processing consisting of parsing or textual content mining. Words are split apart in a process called tokenization.

v. Dropping Common Terms (Stop Words): Occasionally, some extremely common words which would seem like of little cost in assisting pick out documents matching a consumer need are excluded from the vocabulary entirely. These phrases are called prevent phrases or stop words. The overall approach for figuring out a prevent listing is to sort the phrases via series frequency (the full variety of times every time period seems inside the file collection), after which to take the maximum recurrent phrases, regularly hand-filtered for their semantic content comparable to the area of the documents being listed, as a prevent listing. Below is an example for the stop words list. A more comprehensive methodology uses a list of stop words from an external file and removes each word in the collection



a an and are as at be by for from
has he in is it its of on that the
to was were will with

Figure 1 Stop Words

vi. Stemming: Stemming is a method for the depletion of words into their root. Many phrases within the English language can be decreased to their base form or stem. The stem is still beneficial, due to the fact all other inflections of the foundation are transformed into the equal stem. Case sensitive systems ought to have issues while creating assessment among a word in capital letters and any other with the same that means in lower case.

vii. Finding Relative Hardness: Sometimes one file can belong to many categories. By finding the overlapping factor between the categories we will know how much percentage of the files is belonging to different categories. The quantitative measure of closeness of the documents is typically referred as dissimilarity, distance or similarity, with a preferred term being proximity. Files are ‘near’ whilst their dissimilarity or distance is small or their similarity huge. A similarity measure “sim” quantifies the relatedness of documents. Let R is a mapping such that d1, d2 ϵ D, where D is dataset and sim(d1, d2) quantifies how comparable the documents d1 and d2 are. On the way to set up the Relative Hardness (RH Relative hardness is the expression to determine the overlapping factor in the dataset.) of a given corpus, we have examined the vocabulary overlapping a number of the texts of the corpus. We have considered the poplar Jaccard coefficient for calculating the overlapping. We examined all the feasible combinations of more than categories from the corpus and for each of them we calculated its RH. As an instance, for a given corpus of n classes, 2n – (n + 1) feasible subcorpora can be received: e.g. for the R8 (eight categories) we acquired 247 subsets. We calculated their RHs as follows: given a corpus C_i made from n categories (CAT), the RH of C_i = CAT_1, CAT_2, ..., CAT_n is[7]:

$$RH(C_i) = \frac{1}{n(n-1)/2} \times \sum_{j,k=1;j<k}^{n} Similarity(CAT_j, CAT_k)$$

In which the comparison among categories is received with the aid of using the Jaccard coefficient to be able to find their overlapping. However, extra state-of-the-art measures also may be used, along with the one offered within the plagiarism diploma calculation framework.

$$Similarity\ (CAT_j, CAT_k) = \frac{|CAT_j \cap CAT_k|}{|CAT_j \cup CAT_k|}$$

In the above method we've taken into consideration every class j as the “document” acquired through concatenating all the files belonging to the category j[7].

viii.Bag of words : Clustering process is performed on words of the documents in the dataset. The approach after the application of above steps is a collection of words, from which we determined the unique words and stored in a text file for clustering process.

B. PoS Tagging Approach

After getting a Bag of words from the documents, apply pos tagger on the dataset to extract different parts of speech like verbs, nouns, nouns & verbs and with different combinations of parts of speech and etc. The process of obtaining the parts of speech tags for the documents begins with after separating unique words from the documents, each word is compared with the corpus and based on the rules of the corpus the word will be tagged. In this work we have built a PoS tagger and also used a second a popular bench mark Stanford’s tagger. The results of tagged words are accumulated in the respective directories. The Stanford pos tagger is the implemented in java by using Net Beans IDE. To get the pos tags from the documents it needs a jar file creation. So, we run the Stanford tagger in net beans only to get the proper pos tags and remaining work is done in the normal edit plus IDE.

Table 1 : Datasets Characteristics

Dataset	Size of Dataset	Files Considered	Avg. file size in bytes	Avg. Relative Hardness	Categories in Data set
20 news-groups	19,997	4,665	1,268	17%	6
Dt	200	200	837	14%	4
500AB	442	437	941	12%	5
Classic	7,095	800	905	12%	4
Reuters-21578	21,578	1,914	709	15%	4

C. Clustering

The unique parts of speech derived by applying the PoS Taggers on different documents of the dataset, these Unique terms are employed to generate Term Frequency Matrix, called Vector Space Model. This VSM thus derived are different for different PoS, these VSMs are used in k-Means clustering algorithm to categorize the document into sub-groups of related documents. There by derived different clusters based on PoS considered. The aim of this work is to analyze the impact of these parts of speech tagging over text document clustering using cluster evaluation measure called Entropy.

III. RESULTS ANALYSIS

The results of our experiments are tabulated below. The datasets and its characteristics are described. There are in total 5 datasets in the experiment. 20 newsgroups, Classic and Reuters are the most widely used Datasets in Text Mining Research. 500AB and Dt are two datasets prepared by ourselves for testing our hypothesis. These are the collection of Abstracts from various Computer Science Research with a varying data sizes. Table:1 represents dataset's original size and the quantity of files that were considered in experimenting, Average File size of the Datasets considered, average Relative Hardness and no. of categories of the dataset.

We experimented with Stanford's maximum entropy parts of speech tagger on classic dataset for evaluating the cluster quality and we compare the results with our-own created tagger in terms of Entropy. After getting parts of speech from maximum entropy tagger we applied on k-means for evaluating the cluster. The cluster qualities are computed using Entropy measure. The following table summarizes the results by comparing both the PoS methods Our implemented tagger and Maximum Entropy Based PoS Tagger Method using Jaccard Similarity and we got the entropy values for different parts of speech are like this. Table 2 depicts the PoS and the Clustering Accuracy obtained using our PoS tagger. Figure 1: shows the graphical Representation.

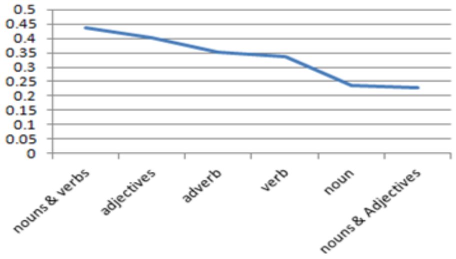


Figure 1: PoS wise Entropy curve Our-own Tagger

Table 2 : Entropy values with our-own Tagger

Parts of speech	Entropy Value
Noun	0.2372724312
Verb	0.3354664877
Adjective	0.4036973326
Adverb	0.3518217469
Noun & Adjectives	0.229513656
Noun & Adverb	0.437259154

Figure 2 : Entropy curve for Maximum Entropy

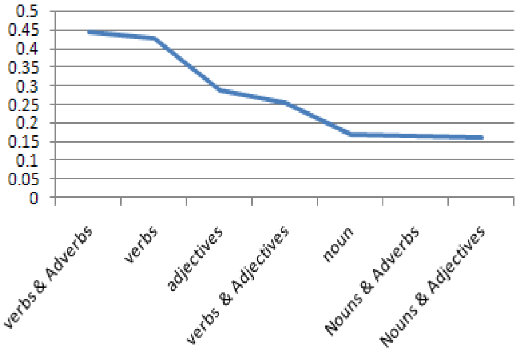


Table 3 : Entropy values with Maximum Entropy Tagger

Parts of speech	**Entropy Value**
Noun	0.1684927006
Verb	0.426183159
Adjective	0.287982780
Adverb	NaN
Noun & Adjectives	0.1611250777
Noun & Adverb	0.1644759240
Verbs & Adjectives	0.2522852755

In this case, we got the best results for the combination of nouns & adjectives in terms of entropy. i.e. if we select the noun & adjectives for the clustering process or if we give any query in the combination of noun & adjective it can fetch effective documents in IRS system.

Table 3 represents the results of Maximum Entropy PoS tagger and Figure 2 depicts it graphically.

From the graph we can observe that the combination of Nouns & Adjectives is giving the best entropy value.

In comparison to our-own pos tagger and Stanford's pos tagger there is a lot difference in the values of Entropy and formation of clusters. Anyhow, to represent texts using best nouns and right names offers a smaller depiction without exacerbate results.

IV. CONCLUSIONS

Document clustering using preprocessing of documents and parts of speech tagging can give the effective clustering results. Parts of speech taggers can be implemented using different techniques. Whether two documents are similar or otherwise is or should be determined on the basis of the words which are identical in the weighted matrix. When this concept combined with pos tagging is guaranteed to be more efficient as this tagging will help determine document similarity based on the word and the context in which it is used.

For future work, Hidden Markov Model (HMM) will be implemented on our-own pos tagger in order to extract the parts of speech by knowing whether the word is noun or verb based on its value of probability.

REFERENCES

1. Keno Buss, Literature Review on Preprocessing for Text Mining. De Montfort University, Institute Of Creative Technologies, Software Technology Research Laboratory, UK.
2. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas: Data Preprocessing for Supervised Leaning. International journal of computer science volume 1 number 2
3. Mark P. Sinka, David W. Corne: A Large Benchmark Dataset for Web Document Clustering. University of Reading, Reading, RG6 6AY, UK.
4. Carlberger and V. Kann. 1999. Implementing an efficient part-of-speech tagger. *Softw. Pract. Exper.*,29(9):815–832.
5. Y. Li, C. Luo, S.M. Chung, Text clustering with feature selection by using statistical data, IEEE Transactions on Knowledge and Data Engineering 20 (5)(2008) 641–652.
6. Y. Sri Lalitha, Dr. A. Govardhan "Semantic Framework for Text Clustering with Neighbors ", S.C. Satapathy et al. (eds.), *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI - Volume II*, Advances in Intelligent Systems and Computing 249, DOI: 10.1007/978-3-319-03095-1_29, © Springer International Publishing Switzerland 2014 pp.261-271.
7. Y. Sri Lalitha, Dr. A. Govardhan, "Analysis of Heuristic Measures for cluster Split in Bisecting K-means", CiiT International Journal of Data Mining and Knowledge Engineering, Vol 5, No 12, 2013 pp 438-443.

AUTHORS PROFILE

Dr. Y. Sri Lalitha, Professor, Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, JNTU(H), India. Ph.D from ANU, Guntur, M.Tech. (CS) from JNTU(H) has around 15 publications iasn renowned Journals. Her Areas of Research Interests includes Machine Learning, NLP, Data Science and Big Data Analytics.

Dr. J Sirisha Devi was awarded B. Tech. in Computer Science and Engineering from Acharya Nagarjuna University -2003. She was awarded M. Tech. in Computer Science and Engineering from GITAM University, Visakhapatnam - 2010. She was awarded doctorate in the year 2016. Her research interests include Human Computer Interaction and Natural Language Processing.

Ms Sukanya Ledalla, Bachelor of Technology from JNTUH, 2007, M.Tech. from JNTUH, 20011, and pursuing Ph.D. from GITAM University. Assistant Professor Department of IT, Gokaraju Rangaraju Institute of Engineering & Technology since 2012. She is a member of IAENG. Her main research work focuses on Sentiment Analysis, Data Minnig, Big Data Analytics. She has 6 years of teaching experience and 5 years of Industry Experience.

N. V. Ganapathi Raju is working as a Professor in I.T. Department, GRIET, Ph.D from JNTUK, Kakinada. M.Tech (C.S.T.) from Andhra University. He has 18 Years of Teaching Experience and 7 Years of Research and his Areas of research interests include IRS and NLP, Data Science and Machine Learning. He got UGC minor project grant MRP-4590/14 (SERO/UGC) in March 2014He published research articles in various International journals and Conferences.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication