Journal of Computer Science and Information Technology
December 2017, Vol. 5, No. 2, pp. 26-39
ISSN: 2334-2366 (Print), 2334-2374 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jcsit.v5n2a4
URL: https://doi.org/10.15640/jcsit.v5n2a4

Design and Develop Semantic Textual Document Clustering Model

SK Ahammad Fahad[1] & Wael Mohamed Shaher Yafooz[2]

Abstract

The utilization of textual documents is spontaneously increasing over the internet, email, web pages, reports, journals, articles and they stored in the electronic database format. It is challenging to find and access these documents without proper classification mechanisms. To overcome such difficulties we proposed a semantic document clustering model and develop this model. The document pre-processing steps, semantic information from WordNet help us to be bioavailable the semantic relation from raw text. By reminding the limitation of traditional clustering algorithms on the natural language, we consider semantic clustering by COBWEB conceptual clustering. Clustering quality and high accuracy were one of the most important aims of our research, and we chose F-Measure evaluation for ensuring the purity of clustering. However, there still exist many challenges, like the word, high spatial property, extracting core linguistics from texts, and assignment adequate description for the generated clusters. By the help of Word Net database, we eliminate those issues. In this research paper, there have a proposed framework and describe our development evaluation with evaluation.

Keywords: Textual Document Clustering, WordNet, Conceptual Clustering, COBWEB, F-Measure.

Introduction

A substantial a part of the offered data persisted in Text databases that comprise vast collections of documents from varied sources. Text documents unit is overgrowing as a result of the increasing measure of information offered in electronic and digitized sort, like electronic publications, various styles of electronic records, e-mail, and also the World Wide Internet. Recently most of the information regarding government, industry, business, and the various establishments unit hold on electronically, inside the kind of text databases. Most of the text databases, sometimes they are semi-structured, and most of the time they are unstructured. Rarely those are structured. Document Pre-processing and agglomeration is very helpful gizmo in today's world where a good deal of documents and information area unit cap and retrieved electronically.

As text info area unit inherently unstructured, some researchers applied the different technique for document management. Researchers have presented the info discovery in writing system that uses the most effective knowledge extraction to induce fascinating experience and knowledge from unstructured text assortment. For illustration technique and efficient transformation, the word frequencies ought to be normalized regarding their relative frequencies that area unit gift during a document and over the entire assortment. Organize a document clustering, making later navigating; the document browsing becomes more comfortable, friendly and economical. Almost, it is not doable for the creature to scan through all the text documents and ascertain the about a selected topic and also the thanks to preparing a large document. To organize a significant amount of knowledge and keep throughout structured format-specific processing techniques is a unit able to use or extract the desired information from the unstructured document collections. The goal of our paper is; text mining is to structure record collections to spice up the flexibleness of users to retrieve and apply the info implicitly contained in those collections. Text mining yields through entirely different phases to complete the goal: pre-processing, using WordNet and term alternative approach.

[1] Faculty of Computer and Information Technology, Al-Madinah International University, Shah Alam, Malaysia
[2] Faculty of Computer and Information Technology, Al-Madinah International University, Shah Alam, Malaysia

Attributes and dimension reduction area unit the required limits in text mining. WordNet is that the merchandise of the associate inquiry project in Princeton (Miller, 1995) that has tried to model the lexical info of a verbalizing of English.

Documents get pre-processed by several steps: firstly, of all; Remove all the stop words, secondly is stemming. Stemming performed by victimization porter's steamer rule, thirdly, it is related to Wordnet. WordNet senses applied, distinctive world words and common word set gets generated by victimization feature alternative approaches. Traditional clustering ways do not seem to be valid on matter clustering. After oral cluster communication, it got to settle on an accomplished clustering technique that may create a real clustering on scientific communication. IT tends to select abstract clump and that we picked the COBWEB clustering algorithmic rule. We live our cluster accuracy by f-measure. It considers every the truth and also the recall of the check to reason the score: the fact is that the very of accurate positive results divided by the quantity of all positive results, and recall is that the very of exact positive results divided by the variety of positive results that need to return. The f-measure are going to understand as a weighted average of the truth and recall, where the associate f - measure reaches its best price at one and worst at zero.

Background and Related Work

Large Corpora area unit high-dimensional about words, documents area unit skinny, area unit of different length, and should contain terms (Aggarwal & Zhai, 2012; Zhai, & Massung, 2016). Several researchers have recognized that partitional cluster algorithms area unit like-minded for cluster large document data sets thanks to their relatively low method wants (Steinbach, Karypis & Kumar, 2000). The presence of logical structure clues inside the document, scientific criteria and math similarity measures are primarily accustomed figure thematically coherent, contiguous text blocks in unstructured documents (Lee, Han & Whang, 2007; Hung, Peng & Lee, 2015; MacQueen, 1967; Ferrari, & De Castro, 2015). Use PLSA to cipher word–topic distributions, fold in those distributions at the block level, and so choose segmentation points supported the similarity values of adjacent block pairs. (Sun, Li, Luo& Wu, 2008; Zhang, Kang, Qian& Huang, 2014; Rangel, Faria, Lima & Oliveira, 2016) use LDA on a corpus of segments, inter-segment cipher similarities via a Fisher kernel, and optimize segmentation via dynamic programming. (Misra, Yvon, Jose, & Cappe, 2009; Glavaš, Nanni & Ponzetto, 2016) use a document-level LDA model, treat sections as new documents and predict their LDA models, and so do segmentation via dynamic programming with probabilistic scores. It is together a challenge to look out the useful data from the large documents (Aggarwal & Zhai, 2012; Zhai, & Massung, 2016). The traditional document cluster unit high-dimensional about texts.(Misra et al., 2009; Glavaš, Nanni & Ponzetto, 2016). The presence of logical structure clues within the document, scientific criteria and applied math similarity measures chiefly accustomed figure thematically coherent, contiguous text blocks in unstructured documents (Sun et al., 2008; Zhang et al., 2014; Rangel et al., 2016). Recent segmentation techniques have taken advantage of advances in generative topic modeling algorithms, which were specifically designed to spot issues at intervals text to cipher word–topic distributions (Lee, Han &Whang, 2007; Hung, Peng& Lee, 2015).

Most of the techniques utilized in document agglomeration affect a document as a bag of words, whereas not considering the linguistics of each document (Fahad & Yafooz, 2017). A traditional formula primarily uses choices like words, phrases and sequences of the documents supported enumeration and frequency of the choices to perform agglomeration freelance of the context (Chim& Deng, 2008; Li & Chung, 2008; Fung, Wang & Ester, 2003). They ignore the linguistics of words in documents. Cluster methods have to be compelled to discover the connections between the documents, then supported these connections the documents area unit clustered (Fahad & Alam, 2016). Given large volumes of documents, a good document cluster methodology might organize those immense numbers of documents into pregnant groups, which modifies other browsing and navigation of this corpus be teeming easier. With an accurate text cluster methodology, a document corpus usually organized into a pregnant cluster hierarchy (Fahad & Yafooz, 2017). That facilitates Associate in Nursing economic browsing and navigation of the corpus or economic information retrieval by that focus on relevant subsets (clusters) rather than whole collections (McKeown et al., 2002; Liu & Croft, 2004). Partitional cluster tries to interrupt the given knowledge set into k disjoint categories such the info objects in an exceeding category are nearer to at least one other than the info objects in alternative categories. The foremost well-known and ordinarily used partitional cluster formula is K-Means (Hartigan, 1975), still as its variances Bisecting K-Means (Forgy, 1965) and K-Medoids (Kaufman & Rousseeuw, 2009). Regarding the distance/similarity live, a hierarchical cluster could use minimum distance (single-link) (Sneath & Sokal, 1973), most distance (complete-link) (King, 1967), distance, or average distance (Fahad & Alam, 2016).

Model-based cluster algorithms plan to optimize the work between the given data and some mathematical models beneath the thought that the information generated by a mix of the underlying probability distributions. Asian nation unit (Kohonen-2012) is one all told the foremost trendy model-based algorithms that use neural network methods for the cluster.

It represents all points in Associate in Nursing passing high-dimensional space by the points in Associate in Nursing passing low-dimensional (2-D or 3-D) spot, such the house and proximity relationship area unit preserved the most quantity as potential. Graph-based cluster algorithms apply graph theories to a cluster. A widely known graph-based discordant cluster formula (Zahn, 1971) depends on the event of rock-bottom spanning tree (MST) of the information then deleting the time edges with the first necessary lengths to urge clusters. Another trendy graph-based cluster formula is MCL (Markov Cluster formula (Van, 2001). It will be mentioned with plenty of details later throughout this section.

A different approach is conceptual clustering. These methods are incremental and build a hierarchy of probabilistic concepts. COBWEB and its successor CLASSIT are the most notable among them. Unlike traditional hierarchical methods (that use similarity measures) they use Category Utility as the cluster quality measure. Conceptual clustering is based on numerical taxonomy (Fisher & Langley, 1986) and was initially introduced (Michalski&Stepp, 1983). Gennari et al. (Gennari, 1989) described the problem of conceptual clustering. Despite variations in the illustration (Witten, Frank, Hall & Pal, 2016) and quality judgments, clustering systems judge general category quality by wanting to an outline or idea description of the category. Fisher and Langley (Fisher & Langley, 1985; Fisher & Langley, 1986) adapt the read of learning as a look to suit abstract clump. Clump and characterization dictate a two-tiered search, a look through an area of object clusters and a subordinate search through an area of ideas. One within the case of stratified techniques, this becomes a three-tiered search, with a ranking search through an area of hierarchies. A robust conceptual clustering algorithm that has been the basis for many other algorithms, for example, LABYRINTH (Thompson & Langley, 1991), ITERATE (Biswas, Weinberg, Yang &Koller, 1991), and COBWEB (Fisher, 1987). The cobweb is a conceptual clustering algorithm developed by Fisher (Fisher & Langley, 1986) for the analysis of categorical data that cannot order. The goal of Cobweb, like all conceptual clustering algorithms, is to build a model that can use for future predictions (Gennari, 1989). Biswas et al. (Biswas et al.,1996) use Cobweb for predicting missing values. Perkowitz & Etzioni discuss the suitability of Cobweb for data mining on the web, (Hurst, Marriott & Moulder, 2003) and Paliouras et al. use Cobweb on the internet, while Li et al. (Li, 2005) combine Cobweb with k-means (MacQueen, 1967) to present an algorithm for large-scale clustering. The algorithm is, also, part of some famous general purpose data mining tools. Two of these data mining tools are (i) Weka, which provides an implementation of Cobweb that applies to categorical and numeric data, and (ii) OIDM, which gives an implementation of Cobweb based on the original Fisher's paper.

WordNet® could also be an enormous on-line database of English. Nouns, verbs, adjectives and adverbs unit classified into sets of psychological feature synonyms (synsets), each expressing a particular construct. Synsets; unit interlinked by suggesting that of conceptual-semantic and lexical relations. Sir James Murray's Oxford English lexicon was compiled "on traditional principles", and no-one doubts the value of the Oxford English Dictionary in subsidence issues with word use or sense priority. Every linguist and psycholinguists have explored in intensive depth the factors deciding the up thus far (synchronic) structure of semantic information unremarkable, and lexical information specifically, Miller and Johnson-Laird have planned that analysis involved the lexical part of the language got to be called unfortunate linguistics. Definitions of common nouns typically give a subject term with characteristic features; that data provides the premise that organizes noun files in WordNet. Three types of individual choices unit discussed: attributes (modification), parts (metonymy), and functions (prediction). The semantic relation found between nouns; but, it is not a basic organizing principle for nouns (Miller-1990). WordNet divides adjectives into two broad classes: descriptive and relative. Descriptive adjectives assign to their head nouns values of the (typically) bipolar attribute and consequently, a unit organized regarding binary oppositions (antonym) and similarity of which means (synonym). Relative adjectives unit assumed to be rhetorical variants of modifying nouns, then unit cross-referenced to the noun files. Also, four variants of lexical illation unit distinguished, that acts in systematic ways in which with the linguistics relations. Finally, the lexical properties of the variant verb group unit created public. Some study efforts explored the use of WordNet as data to spice up document cluster by providing relations between vocabulary terms.

Thus the results area unit altogether entirely different. Where some studies prompt that the use of a WordNet is helpful for cluster methodology, whereas others have consistent with that the WordNet is not useful (Sedding & Kazakov, 2004; Moravec, Kolovrat & Snasel, 2004;Fodeh, Punch & Tan-2009; Recupero, 2007; Yoo, Hu & Song, 2006; Wang & Hodges, 2006; Termier, Sebag & Rousset, 2001).

Hotho et al. used WordNetsyn sets to bolster document vector, showed that enhancing the bag of words with Wordnetsynsets from the phrase among the text and their Hypernyms (up to an exact distance) can produce higher clusters than an understandable bag of words illustration. Recupero and Reforgiato (Recupero, 2007), Wang and Hodges (Wang & Hodges, 2006) used WordNet as data in a document cluster with altogether entirely different data sets; the results are a unit showed that the use of philosophy is helpful for the cluster. Jing, L., et al. used the same technique as Hotho et al. and enhances it by computing a word similarity active support what they call 'mutual information' over their cluster corpus. However, their technique did not prove any intensive improvement over Hotho et al.'s baseline. Passos and Wainer showed that plenty of similarity measures between words derived from Wordnet unit worse than the baseline for the wants of text cluster. Sedding and Kazakov (Sedding&Kazakov, 2004) showed synonyms and Hypernyms, disambiguated solely by Part-of-Speech tags do not seem to be thriving in up cluster effectiveness. Foden et al. (Fodeh, Punch& Tan-2009), Terrier, An et al. (Termier, Sebag& Rousset-2001) used WordNet with entirely different datasets; The results have according to that the abstract ideas adds no worth and impairs the performance of document clusters. Foden et al. self-addressed the problem of the impact of incorporating the ambiguous and synonymous into document cluster, that showed the ambiguous and synonymous nouns play a vital role in the cluster, albeit their clarification does not necessarily cause significant improvement in cluster purity. Moravec et al. (Moravecet al.,2004) showed different results once mistreatment two analysis measures.

The F-measure or F-score is one in all the foremost ordinarily used "single number" measures in Info Retrieval, scientific communication process and Machine Learning. F-measure, generally called F-score or (incorrectly) the F1 metric (the β=1 case of the additional general measure), maybe a weighted mean value of Recall & exactness (R & P). There square measure many motivations for this alternative of mean. Above all, the average value usually applied once averaging rates or frequencies. The first general type, F, permits the differential weight of Recall and exactness; however, ordinarily, they are given equal weight, giving rise to F-Measure. However, as a result of it is so gifting this area unit sometimes understood once about F-Measure. F-measure comes from data Retrieval (IR) wherever Recall is that the frequency with that relevant documents is retrieved or 'recalled' by a system, however, it is notable elsewhere as Sensitivity or True Positive Rate (TPR). The exactitude is that the frequency thereupon retrieved document or predictions unit connectedness or 'correct' and is properly a mode of Accuracy, collectively known as Positive prognostic worth (PPV) or True Positive Accuracy (TPA). F is meant to mix these into one live of search 'effectiveness'.

Proposed Method with Material:

In our study and research, we identified some steps for our whole proposed system. In our methodology, we found six basic steps to complete whole semantic clustering.

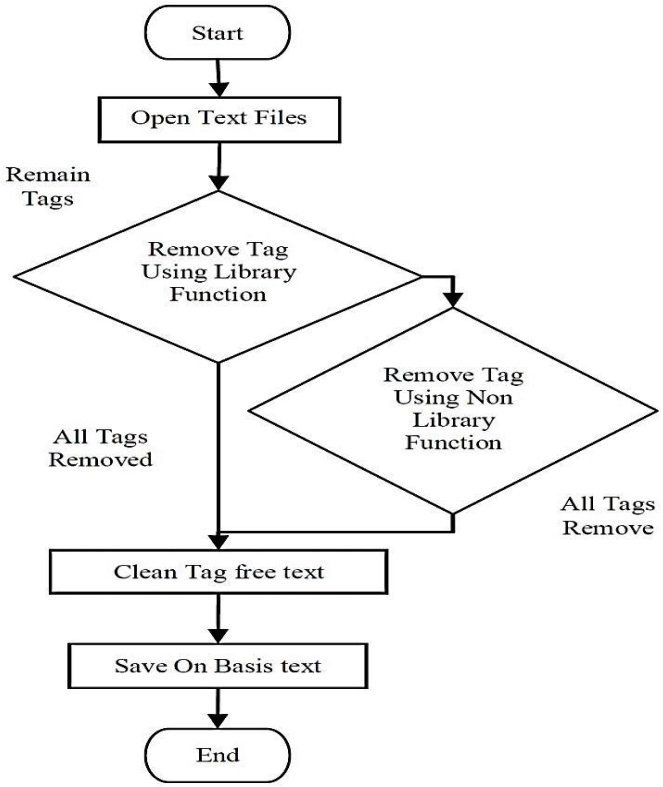
1. Remove Tags
2. Tokenizing Document
3. Remove Stop Words
4. Streaming & Lemmatization (WordNet)
5. Conceptual Clustering (COBWEB Algorithm)
6. Apply F-measure (Cluster)

In the first step, it was going to remove all tags of the text file. Tags are used to represent text, but it was not necessary for the raw text for textual document clustering. This proposed model needs tags or label removed clean text, and that is why there has a developed system to cleaning the sample text to remove the tags or labels. After removing the labels (tags), the proposed system going to split our text into the token. In this study, it needs words as a token. In some case, the researcher has divided the text into a sentence, and they got a token of the phrase. However, for this proposed clustering method, the system needs a token of the word. After completing the tokenized process, it focused on removing unwanted words from those token. By the guidance of the Oxford dictionary, this development includes four-handed and twenty-nine stops word lists. We remove the stop words from our token and get the pure token. When we get noise-free pure text; then we play streaming process.

In streaming, the process needs a tool or algorithm that can help us to get the similar semantic word with a synonym. For this case, in this proposed model, it is going to use WordNet. WordNet is a semantic lexical database can give back the meaning of the word and the type with synonyms.

It completes the streaming process and lemmatizing process and steps forward to clustering. Proposed Clustering will be Conceptual clustering method, and for this textual document clustering, in this development, it was going to apply COBWEB clustering algorithm to complete conceptual clustering of streamed data. It followed COBWEB algorithm steps. After clustering process had finished, it gets the clusters. Those are the clusters can tell us about the documents. Finally, we are going to apply the F - Measure technique to confirm the accuracy. After getting the clusters, it applies the f-measure technique to the clusters. F-measure will give the output of our clustering accuracy. For associate degree correct matter agglomeration, typically it would like sensible data; a radical cleansing of the information is a vital step to boost the standard of knowledge mining ways. Not solely the correctness; conjointly the consistency of values is essential. Pre-processing method for textual document clustering plays a paramount role in text clustering techniques and applications. It is the first step in the semantic text clustering process. There have some tags in a text file for representing the text file. Sometimes those tags make some space, new line, justification, left-oriented size, and more, but when we are going to apply clustering to that text there, those tags have nothing to do. Tags need to clean before for achieving a more accurate result on clustering process.
We remove all tags for the batter clustering process.

Figure 01: Remove tags from input text

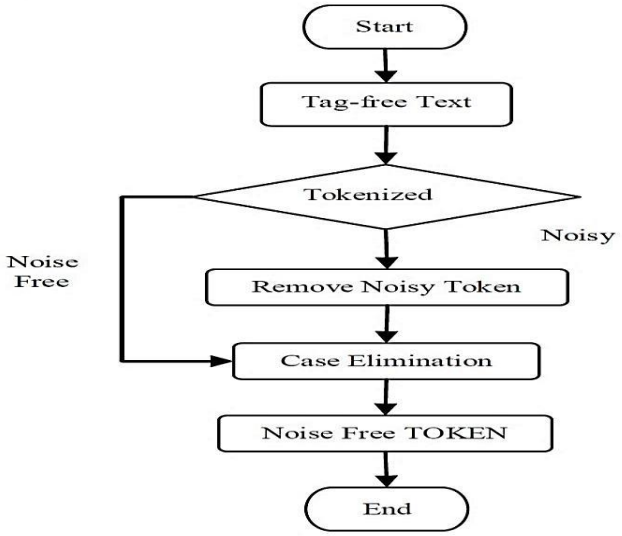


Our concern to work tokenization on lexical analysis for textual data normalization. In lexical analysis, tokenization is that the strategy of breaking a stream of text up into words, phrases, symbols, or various necessary components mentioned as tokens. The list of tokens becomes input for an additional method of parsing or text mining. From given document and these known words, numbers, and alternative characters square measure referred to as tokens (Qui& Tang, 2007; Karthikeyan & Aruna, 2012). In conjunction with token generation, this method additionally evaluates the frequency worth of these token gifts within the input documents. In this step of our proposed methodology, we are going to tokenized out texts. By the previous step, we have tag free text, and those saved as our given name and location. We Open those tags free texts and split the text to word token.

In our development phase, we use NLTK in python. Installing 'NLTK' on python, this semantic clustering process is much easier than previous traditional coding. In the case of tokenization some time some special character going to fact. Sometimes those are count as a token. As like "?" count as a token.

Yes, those are tokens, but in this case, it was going to semantic clustering. It is not necessary them at all. In this paper, it needs clear and burdens free token collection. We are going to eliminate those tokens. We are going to cluster textual database on semantic. So case sensitivity was not a face to get the semantic meaning of the word, but sometimes case sensitivity was a significant issue to get accurate clustering. For those, get appropriate noise free token, we are going to make all texts as lower case.

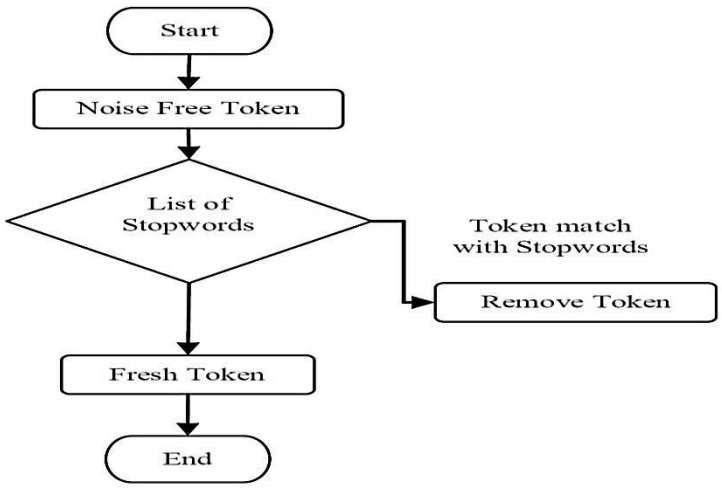
Figure 02: Removing unwanted Noise from Tokens.



There have three forms of stop words in English. Those are; Generic stop-words, Misspelling stop-words and Domain stop-words. Generic stop-words are often picked up once scanning the documents; the latter two have to be compelled to wait until all documents within the corpus read through, and applied math calculations applied. Generic stop words square measure non-information bearing words at intervals in the selected language. Stop words need to remove while not considering any domain information. They are in English; the extraordinarily common words like; "a"; "all"; "and"; "by" and so on.

Misspelling stop-words are not real words, however, misspelling words. Inevitably folks could search by mistake input some words that are not in the dictionaries, like orthography "world" as "wolrd". Of course, at intervals, a context, an individual's being could resolve this is often an orthography error and still be ready to get proper which means from it. However, it might be tough for a PC to confirm the proper spell.

Figure 03: Steps to remove stop words from tokens.



Domain stop-words, in general, aren't ubiquitous words however they transform stop-words solely underneath specific domain data or contents. As an example; in an exceedingly document corpus containing documents from classes animal, automobile, geography, economy, politics and PC, the word "computer" is not a stop-word. As a result, it is not common altogether alternative classes, and it helps to differentiate the PC relative documents from alternative documents like animal-relative or geography-relative ones. However, once considering a corpus at intervals that all documents square measure is discussing different aspects of computers like computer code, hardware, and PC applications, the words "computer" is too ordinary to be enclosed within the following process. In our development, we tend to square measure centered on Oxford English wordbook declared four hundred and twenty-nine stop words. Stop words square measure words that from non-linguistic read do not carry data. They are square measure words in English like pronouns, prepositions, and conjunctions that square measure is accustomed to giving structure to the language instead of content. These words, that square measure encountered soft and carried no helpful data regarding the content and so the class of documents, a square measure known as stop words. Removing stop words from the documents is extremely common in data retrieval. One significant property of stop-words is that they are ubiquitous words. The reason of the sentences still command when these stopwords removed. Stop-words square measure smitten by the tongue. Different languages have their stop-words list.

Stemming techniques unit needs to see the root/stem of the award. Stemming converts words to their stems, which has an honest deal of language-dependent scientific information. Behind stemming process, the hypothesis is that words with continuous steam or word root mainly describe same or relatively shut ideas within the text, then words square measure usually conflated by pattern stems (Gaigole, Patil & Chaudhari, 2013). In most languages there exist different syntactical forms (Vester & Martiny, 2005) of a word describe constant thought. In English, nouns have singular and plural forms; verbs have a gift, past and participial tenses. These different styles of the constant word may well be a drag for text knowledge analysis as a result of their needs, different spellings, however, share the similar that means. Though the steaming method could also be useful for the clump algorithms, it should conjointly contraryaffect them if over-stemming happens. Over-stemming implies that words square measure unsuccessfully steamed along as a result of their sufficiently entirely different in that means and that they must not sort along. Over-stemming introduces noise into the process and leads to poor clump performance. A good steamer needs to be ready to convert various syntactic kinds of a word into its normalized kind deflate variation of index terms, save memory and storage and may increase the performance of clump algorithms to some extent; within the in the meantime, it needs to attempt to avoid over-stemming.

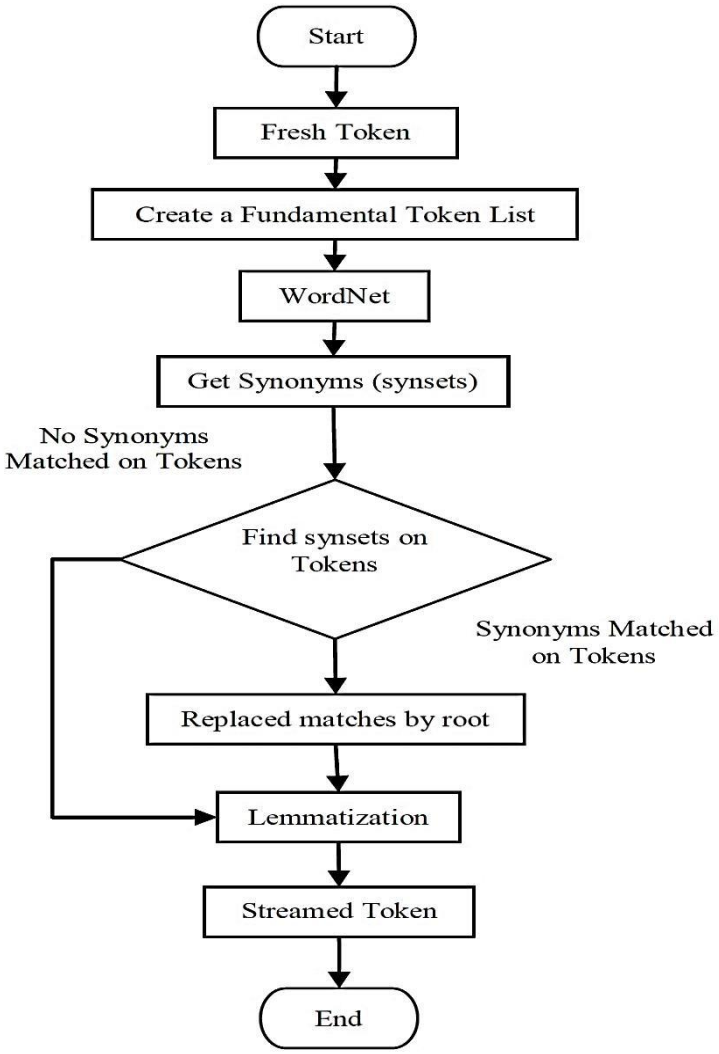
In our development, we are taking help from WordNet to get appropriate and most valuable streaming process. We prepare a first token list for an entire token; we have and send them one by one to WordNet. We query the synonyms to WordNet and get back synsets. When it gets any synonyms of any words, it called synset. After replacing to synsets and lemma by root word out data are more normalized and those are helpful to our clustering algorithms to get more appropriate clustering. In this stage, we are focused on lemmatization to finish the streaming process. Lemmatization is the process of the normalized word from its forms. We used to customize wordnet lemmatized tools to lemmatization. That will clear out lemmatization process. Lemmatization process completes means we have streamed token. That means lemmatization is the last things to do in the streaming process. Now we have outstreamed token, and we are going to apply the clustering process. In this stage, we are going to make clustering to our steamed token to get the clusters. To making clustering process, we are going to apply conceptual clustering, and in this case, we choose COBWEB algorithm to complete clustering process. Before clustering process, we have to take one more step for finally prepares data for clustering. For clustering, we do not need the data every time. In our development, we are using SQLite database for saving tokens. For each text document, we allocate an array and save those token to the array. Now we have a normalized streamed token in that array. We just prepare a table of token information according to our existing database SQLite. Name of input file name and its token with the number of frequencies. We need to measure the frequency for preparing the data table. For measure the frequency we need a first token list for each document. We have the original list; we prepare this list for the streaming process.

When we have the frequency of token in each text with the list, we are going to save them into one table. In thischart, we have the word (token) source name, name or word (token) and frequency of the word (token). Now send those data to COBWEB algorithm for hierarchical clustering. For our development project, we use Python platform. To complete the clustering process of those data to COBWEB algorithm, we use the tools, named "concept_formation 0.1.3" this is a complete package for COBWEB conceptual clustering.

The package has the same pure COBWEB clustering algorithm in code. Here we attach the general COBWEB algorithm. To use COBWEB algorithm to our development purpose.

When we apply some accuracy measure technique to clusters and get a satisfactory result, then we can assure our method is useful and convenient for textual document clustering. For measuring the accuracy, we are going to apply f-measure to cluster. There hastwo main measure component of f–measure; Precision and recall. The maximum value of the f - measure is 1, and the minimum value is 0. It means that the clusters 100% accurate when it is 1. Our development developed in Python. In case of Python, we are giving the general precision and recall value retrieve coding was given below;

Figure 04: Lemmatization and Streaming process Flow-Chart



When there have the precision and recall it is just a matter of time to know the accuracy of the cluster.$F = \frac{2PR}{P+R}$. Here, F means F-Measure. P Represents the precision and R means the recall. We know the precision and recall. The whole f-measure process can be completed by usingNLTK tolls. "[nltk.metrics.scores. f_measure ($reference, test, alpha = 0.5$)" this tool can do the same thing. Calculate the precision and recall and then we use those two values to calculate the accuracy of clusters.

Evaluation

For development, we use Python as our language. We use several Tolls for testing and developing our framework. Our whole development and testing accomplished in Windows environment.

Python is a brilliant and dynamic programming language. It used for general purpose, and it is a procedural language. It is an interpreted language, and that is the point that makes comfortable to find a mistake to the programmer. We develop in Python because there have a lot of useful tools for python that can make Natural language processing very faster and very easy to operate. Inside our Microsoft Windows 10 we use IDE for Python, and in Python, we used many tools for making sure the code is simple and code simple.

The testing and development phase we use Windows 10 ultimate edition IN HP Touch Smart 320 Desktop PC. It is technical; hardware specification is in below;

- Display: 50.80 cm (20 inch) Resolution: 1600 x 900 (16:9 aspect ratio)
- Motherboard: Angelino2-UB
- Processor: AMD A6-3600, 4 MB Cache
- Memory: 16 GB, PC3-10600 MB/Sec
- Hard Drive: 1 TB, 7200 RPM Rotational Speed

For testing our development, we need some sample data. We can test a massive amount of data. However, the experimental information in a limited data because in testing face we will try to add the data and static result and every detail. We get 20 papers abstract those are related to our study. So technically we have 20 sample data. We are going to apply start to end every step to this 20 sample data.

We have the development framework and sample data. In testing, we complete our process in two phases. First one is Pre-processing, and then it performs Clustering and accurate measure. By taking input our sample data, firstly we make an operation remove and tokenize to each text file. We have 20 text files. After removing tags and tokenized 20 files, token reports given in Table 1.

Table 1: Sample text file report after tokenize

Name of Source File	Number of Token in File
"Sample 1"	213
"Sample 2"	257
"Sample 3"	127
"Sample 4"	204
"Sample 5"	451
"Sample 6"	216
"Sample 7"	108
"Sample 8"	259
"Sample 9"	151
"Sample 10"	79
"Sample 11"	149
"Sample 12"	86
"Sample 13"	154
"Sample 14"	100
"Sample 15"	132
"Sample 16"	84
"Sample 17"	152
"Sample 18"	139
"Sample 19"	114
"Sample 20"	117

After tokenize, we are forward to remove stop words tokens from the token we have. We remove all the stop words among on our 20 sample files. List of stop words collected from Oxford dictionary service. Those are only 429 words but uses frequency of those words are very much higher. After removing the stop word our token report and remove token numbers mentioned in Table 2.

Table 2: Number of tokens removed and Number of tokenleft.

Name of Source File	Total Remove Token	Token after Remove Stop Word
"Sample 1"	86	127
"Sample 2"	105	152
"Sample 3"	71	56
"Sample 4"	86	118
"Sample 5"	213	218
"Sample 6"	103	113
"Sample 7"	61	47
"Sample 8"	128	131
"Sample 9"	62	89
"Sample 10"	34	45
"Sample 11"	60	89
"Sample 12"	49	37
"Sample 13"	74	80
"Sample 14"	47	53
"Sample 15"	66	66
"Sample 16"	37	47
"Sample 17"	59	93
"Sample 18"	60	79
"Sample 19"	52	62
"Sample 20"	71	46

In this step, we get help from Word Net. We get the synset (synonyms) of the word and replaced synonyms to the real word. The result of Word Netsynset superseded by the root word. After this, we are focused on lemmatization process, and by lemmatization, we are going to make out text more stable and more normalized for clustering. We are attaching tokens after lemmatized by our lemmatization process. After completing the lemmatization us going forward to our last step for data pre-processing. Get the frequency. In this case, here we give the frequency for all 20 sample files. Only more than two times means, minimum three times appear words, we are taking for clustering. So our list has only three times appeared world's list with the source name and frequency.

In this section, we perform the conceptual clustering, and we follow COBWEB algorithm. When we have the frequency of words for each input text, then we input the words in the frequency table input to our development algorithm, and it returns the clustering result.

Table 3: Clusters with the member With Accuracy

Cluster Name	Member of Cluster (Source file)	Accuracy by F-Measure
Algorithm	Sample 5, Sample 8, Sample 11, Sample 13, Sample 19	88.57 %
Approach	Sample 2, Sample18	92.87%
Citat	Sample 5	71.42%
Classif	Sample 8	85.71%
Cliqu	Sample 4	85.71%
Cluster	Sample 2, Sample 4, Sample 5, Sample 6, Sample 7, Sample 9, Sample 10, Sample 11, Sample 12, Sample 13, Sample 14	90.90%
Cobweb	Sample 8	100%
Concept	Sample 10, Sample 17	71.42%
Data	Sample 9, Sample 13, Sample 14	85.71%
Document	Sample 1, Sample 2, Sample 4, Sample 5	100%
f-measur	Sample 20	85.71%
Function	Sample 8	85.71%

Inform	Sample 2	71.42%
Insert	Sample 8	87.71%
Language	Sample 2	100%
Measure	Sample 5, Sample 15	71.42%
Model	Sample 5	85.71%
Multilingu	Sample 2	100%
Node	Sample 8	100%
Object	Sample 8	71.42%
Ontolog	Sample 5, Sample 17	78.57%
Oper	Sample 8	71.42%
Pass	Sample 19	88.57 %
Probabl	Sample 15	92.87%
Select	Sample 5	71.42%
Semant	Sample 4, Sample 5, Sample 17	85.71%
Separ	Sample 8	85.71%
Similar	Sample 17	90.90%
Singl	Sample 19	100%
Technique	Sample 17	71.42%
Term	Sample 1	85.71%
Tree	Sample 8	100%
Valu	Sample 8	85.71%
Version	Sample 19	85.71%
Word	Sample 1	71.42%

We give 20 papers abstract and thus have 3,292 numbers of the token. After completing the entire clustering process; system shows that there have 35 clusters. In this 35 cluster, all sample text files associated except sample 3 and sample 16. Those two samples do not have enough maturity to assign a cluster. Our clustering process oriented for high quality. We focused on quality (accuracy of the cluster). To determine and quality out clusters passed from standard quality measurement technique. We apply f-measure to our 35 groups. Minimum accuracy was for our cluster was our overall accuracy for all clustering was71.42%.

Discussion

In this experiment, there have 20 samples from 20 different papers abstract. After removing the tags from the sample text, when it tokenized, it has 3292 tokens. There tend to continue the operation on those 3292 token. Once obtaining the token there tend to forward to get rid of the stop words from those tokens, then it trends towards finding that 1524 token is from stopwords. It tends to remove those tokens from the total tokens. Those can be an enormous range of the token. That square measure removed. Those are 46.29% of the entire token. Once take away there has 1748 token. Those 1784 tokens are sent to WordNet separately and get the synsets (synonyms). Then the synsets have replaced the word with its signifier. Then it steps forward to the process of lemmatization. This system sends original word for lemmatization. When it has a trend towards complete outset margining and lemmatization method, and it returns 672 tokens to the system. The system gets 3292 tokens from the input, and currently, it has solely 672 tokens to cluster, and it is 20.41% of the total inputted token. Document pre-processing, normalized knowledge, terribly swimmingly and currently, it tends to have one fifth of the inputted knowledge solely. In those 672 token numbers of the distinctive token square measure a hundred and forty-four. They seem many times on those entire twenty sample inputs. Most twenty-two times it was found a word. Some number of the phrase seems only once. When COBWEB algorithmic rule clusters those 672 tokens; then it gets thirty-five clusters. During this thirty-five cluster, all sample documents associated except sample three and sample sixteen. Those two inputs do not have enough maturity to assign a cluster.

This agglomeration method familiarized for prime quality. It was time to center on quality (accuracy of the cluster). To see and quality out clusters square measure passed from conventional quality mensuration technique. In this system, By apply f-measure in the thirty-five clusters, the accuracy can be assured. Some clusters square measure 100% correct, however, System take minimum accuracy consider overall accuracy for all agglomeration was 71.42%.

Conclusion

In this age of information technology, information is power. The textual document has too much information those are practical and relevant to our daily life. We are trying to find information from the Textual documents. This proposed framework to do valuable clustering on textual documents for grab the secret information from unsupervised, unclassified text. This Methodology proposed and developed a system with the capability to work with the semantic meaning of textual data. There Word Net used to ensure the semantic value of data and maintain relation semantically. In this paper, it was trying to deliver a very quality full, accurate clustering. F-measure evaluation and testing assure that our clusters are so accurate. All over we achieve 71.42% of cluster accuracy. Semantic clustering with Word Net gives us a robust semantic relation clustering and by f-measure ensures the quality. In this framework, it figured out the general context and development; it developed with Python program with COBWEB algorithm. It takes help from Word Netsynset. Several systems can update our development. In the future, it can focus on some points those can make a semantic document clustering more eligible. There have some chance to use the new version of Conceptual clusterings like COBWEB/3 or ITERATE or LABYRINTH. It designed for word token; in the future, there has some chance to work with sentence token. There have used an only synset feature of WordNet. There have much more tools on WordNet. Like; type, semantic meaning. It can use them for future research.

References

Aggarwal, C. C., &Zhai, C. (2012).A survey of text clustering algorithms. In Mining text data (pp. 77-128). Springer US.

Biswas, G., Weinberg, J. B., Yang, Q., &Koller, G. R. (1991, June).Conceptual clustering and exploratory data analysis.In Proceedings of the Eighth International Conference on Machine Learning (pp. 591-595). Morgan Kaufmann Publishers Inc..

Chim, H., & Deng, X. (2008).Efficient phrase-based document similarity for clustering.IEEE Transactions on Knowledge and Data Engineering, 20(9), 1217-1229.

Fahad, S. A., &Alam, M. M. (2016).A Modified K-Means Algorithm for Big Data Clustering.International Journal of Science, Engineering and Computer Technology, 6(4), 129.

Fahad, S. A., &Yafooz, W. M. (2017). Review on Semantic Document Clustering. International Journal of Contemporary Computer Research, 1(1), 14-30.

Ferrari, D. G., & De Castro, L. N. (2015). Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. Information Sciences, 301, 181-194.

Fisher, D. H. (1987).Knowledge acquisition via incremental conceptual clustering.Machine learning, 2(2), 139-172.

Fisher, D., & Langley, P. (1985).Approaches to Conceptual Clustering (No.UCI-ICS-85-17).CALIFORNIA UNIV IRVINE DEPT OF INFORMATION AND COMPUTER SCIENCE.

Fisher, D., & Langley, P. (1986).Conceptual clustering and its relation to numerical taxonomy.In Artificial intelligence and statistics.

Fodeh, S. J., Punch, W. F., & Tan, P. N. (2009, March).Combining statistics and semantics via ensemble model for document clustering.In Proceedings of the 2009 ACM symposium on Applied Computing (pp. 1446-1450).ACM.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Biometrics, 21, 768-769.

Fung, B. C., Wang, K., & Ester, M. (2003, May).Hierarchical document clustering using frequent itemsets.In Proceedings of the 2003 SIAM International Conference on Data Mining (pp. 59-70).Society for Industrial and Applied Mathematics.

Gaigole, P. C., Patil, L. H., &Chaudhari, P. M. (2013).Preprocessing Techniques in Text catagorization. In National Conference on Innovative Paradigms in Engineering & Technology (NVIPET-2013), Proceedings published by International Journal of Computer Applications (IJCA).

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In Proceedings of the New Zealand computer science research students conference (pp. 57-64).

Glavaš, G., Nanni, F., &Ponzetto, S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. Association for Computational Linguistics.

Han, J., Pei, J., &Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hartigan, Jhon A., &Hartigan, J. A. (1975).Clustering algorithms (Vol. 209). New York: Wiley.

Hung, C. C., Peng, W. C., & Lee, W. C. (2015). Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. The VLDB Journal—The International Journal on Very Large Data Bases, 24(2), 169-192.

Hurst, N., Marriott, K., &Moulder, P. (2003). Cobweb: a constraint-based WEB browser. In Proceedings of the 26th Australasian computer science conference-Volume 16 (pp. 247-254). Australian Computer Society, Inc..

Karthikeyan, M., &Aruna, P. (2013).Probability based document clustering and image clustering using content-based image retrieval. Applied Soft Computing, 13(2), 959-966.

Kaufman, L., &Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

King, B. (1967). Step-wise clustering procedures. Journal of the American Statistical Association, 62(317), 86-101.

Kohonen, T. (2012).Self-organization and associative memory (Vol. 8).Springer Science & Business Media.

Lee, J. G., Han, J., &Whang, K. Y. (2007, June). Trajectory clustering: a partition-and-group framework. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data (pp. 593-604).ACM.

Li, T. (2005).A general model for clustering binary data.In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 188-197).ACM.

Li, Y., Chung, S. M., & Holt, J. D. (2008). Text document clustering based on frequent word meaning sequences. Data & Knowledge Engineering, 64(1), 381-404.

Li, Y., Luo, C., & Chung, S. M. (2008). Text clustering with feature selection by using statistical data. IEEE Transactions on Knowledge and Data Engineering, 20(5), 641-652.

Liu, X., & Croft, W. B. (2004).Cluster-based retrieval using language models.In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 186-193).ACM.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., ... & Sigelman, S. (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In Proceedings of the second international conference on Human Language Technology Research(pp. 280-285). Morgan Kaufmann Publishers Inc..

Michalski, R. S., &Stepp, R. E. (1983).Learning from observation: Conceptual clustering.In Machine learning (pp. 331-363).Springer Berlin Heidelberg.

Miller, G. A. (1990). Nouns in WordNet: a lexical inheritance system. International journal of Lexicography, 3(4), 245-264.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Misra, H., Yvon, F., Jose, J. M., &Cappe, O. (2009). Text segmentation via topic modeling: an analytical study. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1553-1556).ACM.

Moravec, P., Kolovrat, M., &Snasel, V. (2004).LSI vs. Wordnet Ontology in Dimension Reduction for Information Retrieval.In Dateso (pp. 18-26).

Oikonomakou, N., &Vazirgiannis, M. (2009).A review of web document clustering approaches. In Data Mining and Knowledge Discovery Handbook(pp. 931-948). Springer US.

Pantel, P., & Lin, D. (2002).Document clustering with committees.In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 199-206).ACM.

Qiu, J., & Tang, C. (2007).Topic oriented semi-supervised document clustering. In Proceedings of SIGMOD 2007 Workshop IDAR.

Rangel, F., Faria, F., Lima, P. M. V., & Oliveira, J. (2016).Semi-Supervised Classification of Social Textual Data Using WiSARD.ESANN.

Recupero, D. R. (2007).A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. Information Retrieval, 10(6), 563-579.

Sathiyakumari, K., Manimekalai, G., Preamsudha, V., & Scholar, M. P. (2011).A survey on various approaches in document clustering. International Journal of computer technology and application (IJCTA), 2(5), 1534-1539.

Sedding, J., &Kazakov, D. (2004).Wordnet-based text document clustering.In proceedings of the 3rd workshop on robust methods in analysis of natural language data (pp. 104-113).Association for Computational Linguistics.

Sneath, P. H., &Sokal, R. R. (1973).Numerical taxonomy. The principles and practice of numerical classification..

Steinbach, M., Karypis, G., & Kumar, V. (2000, August).A comparison of document clustering techniques.In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526).

Sun, Q., Li, R., Luo, D., & Wu, X. (2008).Text segmentation with LDA-based Fisher kernel. In Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: short papers (pp. 269-272). Association for Computational Linguistics.

Termier, A., Sebag, M., &Rousset, M. C. (2001).Combining Statistics and Semantics for Word and Document Clustering.In workshop on ontology learning.

Thompson, K., & Langley, P. (1991).Concept formation in structured domains. Concept formation: Knowledge and experience in unsupervised learning, 127-161.

Van Dongen, S. M. (2001). Graph clustering by flow simulation (Doctoral dissertation).

Vester, K. L., &Martiny, M. C. (2005).Information retrieval in document spaces using clustering (Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark).

Wang, Y., & Hodges, J. (2006).Document clustering with semantic analysis.In System Sciences, 2006.HICSS'06.Proceedings of the 39th Annual Hawaii International Conference on (Vol. 3, pp. 54c-54c).IEEE.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Yoo, I., Hu, X., & Song, I. Y. (2006).Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering.In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 791-796).ACM.

Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on computers, 100(1), 68-86.

Zhai, C., &Massung, S. (2016). Text data management and analysis: a practical introduction to information retrieval and text mining. Morgan & Claypool.

Zhang, Q., Kang, J., Qian, J., & Huang, X. (2014). Continuous word embeddings for detecting local text reuses at the semantic level. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 797-806).ACM.