

Clustering Source Code Comments

Welisarage Sudara Fernando
Dept. of Electrical and Computer Engineering
University of Auckland
Auckland, NZ 1010
wfer594@aucklanduni.ac.nz

Abstract—Text clustering and classification techniques are used in characterising text documents into one or more categories based on various factors. This study focuses on classifying source code comments into a set of categories by utilising several natural language processing techniques such as TF-IDF vectorising, stop words elimination and frequency distribution count of "Part of Speech" tags. Furthermore, the comments to perform classification are also collected by implementing a source code scanner for Java.

Index Terms—text classification, comment, tf-idf vectorising

I. INTRODUCTION

Data classification highlights the relationships between data points. It uncovers information otherwise hidden in plain-sight which can be visualised and processed to make better-informed decisions. Text clustering is a form of data classification which divides a set of text documents into one or more categories based on their underlying linguistic content. It can be visualised as a mapping function which maps the given documents into categories. It can be one-to-one as well as one-to-many [1]. The semantic factors that influence text classification results are discussed, focusing on the ones that apply on "short text"s, such as source code comments which are usually in the range of 100 to 150 words.

Natural Language Processing (NLP) has been a vital topic in computer science. It involves analysing human languages and recognising the connotative idea of text documents. NLP is a subfield of Artificial Intelligence (AI) and Machine Learning (ML) which extends a computer program's understanding of human languages. It powers several linguistic tasks such as text summarization, classification, translation and manipulation. Human languages are inherently complex. Some complexities of the English language are different literal meaning from the actual implied meaning (idioms and slang), different meaning of the same word depending on the context and different style of writing (American and British). However, advanced NLP techniques have progressively enabled the understanding of the intricate languages [2] which innately enables excellent text manipulation using computer systems. This paper explores different NLP techniques and algorithms which can be utilized for meaningful text classification.

The clustering algorithms discussed in this research are based on *unsupervised learning* methods which utilises a set of features $F_1, F_2, F_3 \dots F_n$ obtained from n observations to produce a meaningful visualization of data. This is opposed to *supervised learning* which predicts a response R over the feature set

$F_1, F_2, F_3 \dots F_n$ using a set of f features $\{F_1, F_2, F_3 \dots F_n\}$ measured from n observations and a response relation R_R which is also measured on the same n observations [3].

The only input for an *unsupervised learning* approach is the dataset. The dataset for this project is obtained through a source code scanning program designed to scan through Java source code. An *unsupervised learning* procedure does not require a pre-determined response relation (a model) since the process does not involve in any prediction. In a *supervised learning* environment, the results can be cross-validated with the predetermined model. However, in an *unsupervised* approach, there is no universally accepted mechanism to validate the results since the problem itself was unsupervised without an initial model [3]. The algorithms used in this research visualise the results, but unable to cross-validate due to the unsupervised nature of the problem.

Two main outcomes of this project are retrieving comments from source code files and classifying them into several categories.

A. Retrieving source code comments

The dataset was obtained through a source code scanner program which was designed to scan through a given set of Java source code files collecting both single-line as well as multi-line comments. More than 5,000 Java source files were obtained from several well-established and maintained open source projects such as Firebase Admin API (Java), Android Open Source Project and Apache Hadoop. The comments scrapped from the used files were written in English which adheres to correct English language grammar and semantics.

B. Comment clustering

The collected comment data is then processed using various NLP techniques and data cleansing methods. After cleansing it is represented it as $n \times n$ matrix. The matrix is then vectorised using a TF-IDF Vectoriser and then fed into the algorithm. The DBSCAN clustering algorithm was used as the clustering algorithm for this project. Finally, the result set is then examined to identify and visualise the cluster output.

II. LITERATURE REVIEW

A. Retrieving source code comments

Source code parsing and lexical analysis of the programming languages are used in many software products such as Integrated Development Environments (IDE) and other text

editors to provide visual feedback of the code that a developer is working on. These tools utilise regular expressions and other language-specific tools to construct a data structure called *abstract syntax tree* (AST) which is composed of a sequence of characters [4]. Several authors have attempted to design software tools to obtain an AST from Java source code.

One such implementation is to obtain an XML view of the Java source code, named JavaML [4]. The authors formalise a *document type definition* (DTD) which describes the conversion of Java grammar into XML nodes. Then a computer program based on IBM Jikes Java compiler framework [5] was used to parse the standard Java source files and convert them into JavaML documents. These JavaML documents can be utilised by other programs which support XML file parsing. Java is a live language which gets continuously updated. Hence it is vital to keep the language parsers up to date which is also being discussed by the authors in their paper. However, the proposed algorithm only supports Java 5.0 and is unable to process the new Java lexical structure introduced in newer Java versions. Also, the suggested procedure generates a complex and bulky XML document even for simpler Java source code. XML itself is verbose and converting Java source with all its individual properties result in a massive XML document which makes it challenging to cross-validate.

Contrary to AST, Token-based source code feature identification is also found in the literature. The token-based approach is linear complexity in both time and space and does not require the whole document to be parsed at once [6]. It even works when the source code is syntactically incorrect or incomplete. It is more suitable to identify a particular language feature (such as individual methods, variables or comments). AST-based procedures usually traverse the whole document several times to classify the distinct language components whereas the token-based approaches tokenise the source code to recognise the required language feature.

An open-source library named "*javalang*" provides a lexer and a parser for Java 8.0 syntax. "*javalang*" uses token-based feature extraction to scan and tokenise a Java source code and represents it in a tree data structure [7]. It also provides a Python API to access the properties and methods listed in the Java source code. However, one shortcoming of this library was that it was unable to recognise multi-line comments.

B. Comment clustering

Several text clustering algorithms such as K-Means, DBSCAN and FTSC exist in the current literature. These algorithms are mainly categorised into prototype-based, density-based and hierarchical clustering. Prototype-based algorithms such as K-Means assigns each document into its nearest initial prototype cluster. Density-based clustering algorithms such as DBSCAN, separate documents into subsets of similar densities [8]. These algorithms differ from each other based on several factors such as scalability, use case and geometry. Each algorithm also requires a different set of parameters as inputs.

The frequent terms in text clustering is a common factor analysed by several algorithms. Frequent term Set-based clustering (FTSC) was introduced by X Liu et al. which uses frequent term sets for text clustering [9]. FTSC uses a part-of-speech tagging system to segment the terms of a text document and use a discrimination value (DV) of feature terms to find words which can reflect the caption of a text. The DV value represents a word token's ability to distinguish unique document with different contents when it is used to identify the content. Furthermore, Frequent Term Set-based Hierarchical Clustering algorithm (FTSHC) modelled after FTSC can produce overlapping clustering results which increase its reliability and accuracy. Authors claim that it is both faster and reliable than the K-Means algorithm.

FTSC improves the efficiency of clustering by choosing a set of frequent terms based on its appearance frequency as the candidate set without immediate clustering of the documents with high dimensions. Mining of the frequent terms is a two-step process where all the frequent items found in the text are filtered according to a predetermined threshold (internally known as *min_sup*) [9]. Then the general association rules are obtained for each filtered item set adhering to the rules defined in the Apriori algorithm. Each frequent term in frequent term set $\{F\}$ forms a clustering category where each category $\{f_i\}$ has the least mutual overlap [9]. It shows that there is no much difference in the clustering results obtained from FTSC and K-Means algorithms according to the quantitative analysis conducted by the authors using F-Measure on two datasets (namely *NTCIR-3* and *Reuters-21578*). However, due to the nature of the frequent term selection algorithm used in mining the datasets, FTSC is more optimised for text documents with longer and larger text data [9].

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm which is heavily adopted for text clustering. It requires two main parameters; a point neighbourhood and the minimum number of points (*min_pts*) in the neighbourhood. The cohesion of the result depends on this minimum point value [8]. The algorithm starts at an arbitrary position and if $min_pts \leq$ number of points in the initial position, then a cluster is formed. Otherwise, the point will be marked as noise and will select another point to repeat the process. DBSCAN produces clusters of arbitrary shape and is more sensitive for datasets with noise.

III. DATASET PREPARATION

The primary goal of this research is to mine comment data from a given set of Java source files and to cluster them based on their linguistic content. A Java source code scanning program was developed in Python to scan through source files and collect both single-line as well as multi-line comments. The program uses a token-based source code scanning approach since it only targets a specific language component (i.e comments). AST traverses the source file several times to identify all the language components which is an overkill for this research. Token-based approach also dramatically reduces the time and memory requirements of the

scanning program [6] which is essential because the program has to scan an adequate amount of source files to obtain comment data for a meaningful clustered result set.

A. Comment data collection

The program scans the source files and identifies the comment start and end tokens depending on the comment type. It then starts capturing the comment as soon as it detects the start character and ends the comment when it meets the end character. The start character may appear in any arbitrary location in the scanned line. In case of the multiline comments, the star (*) character that appears in between the lines are omitted. To minimise unwanted omissions, it gets excluded only if it is followed by a line-end character.

TABLE I
START AND END TOKENS FOR JAVA COMMENTS

Comment Type	Start Character	End Character
Single-line	//	Line end character
Multi-line	/*	*/

The collected comment data is then stored in a Python dictionary along with a reference to the file.

```
package com.sudara.test;
import com.apple.eawt.*;
// Testing comments
- public class TestingComments {
    /**
     * This is a multiline comment - on a field
     */
    public String a;
    // This is a single line comment - on a field
    public String b;

    /**
     * This is a multiline comment on a method
     */
    public void method1() {
    }

    // This is a single line comment on a method
    public void method2() {
        int i = 0; //This is a comment within a line
    }
}
```

```
[ntk_data] downloading package stopwords to
[ntk_data] /Users/Sudara/ntk data...
[ntk_data] Package stopwords is already up-to-date!
Getting files in the path /Users/Sudara/Offline/source-code-analysis
Found 1 files in the path /Users/Sudara/Offline/source-code-analysis
Found 6 comments in file /Users/Sudara/Offline/source-code-analyse
[FILE] : /Users/Sudara/Offline/source-code-analyser-resources/test
[COMMENT] : Testing comments

[COMMENT] : This is a multiline comment - on a field /

[COMMENT] : This is a single line comment - on a field

[COMMENT] : This is a multiline comment on a method /

[COMMENT] : This is a single line comment on a method

[COMMENT] : This is a comment within a line
```

Fig. 1. Comment scanning example

B. Data cleaning

Data cleaning is an essential step in text clustering process. Text clustering, specially density-based clustering algorithms cluster documents depending on the neighbourhood spatial size. Several common English language terms such as articles, pronouns, prepositions and conjunctions can skew the results

as they commonly occur in text documents. Also in most cases, the less frequent words in a sentence convey the meaning rather than the most occurring articles and pronouns.

1) *Filtering stop words*: Stop words with a high count frequency occur in a variety of languages and texts. These words can skew the frequency counting and density based classifications. They appear in several places of the sentence yet fail to convey the meaning without other content words accurately. Examples for stop words in the English language include "of", "is", "the" etc. There are two possible ways to filter the stop words. One is to use a precompiled stop words dictionary and to refine the text document excluding the words found in the dictionary. The other option is to use a word frequency threshold where words with higher frequencies are considered stop words [10].

The NLTK Python package, which is used for the clustering process uses a precompiled stop words dictionary. Once the "stopwords" package is downloaded, the words can be filtered by specifying the language.

2) *Stemming and lemmatization*: Both stemming and lemmatisation are used in English language text processing. Reducing a word by analysing its prefix or suffix into its stem or root is known as stemming. Words with the same meaning with different prefixes and suffixes can affect count frequency. For example, words "running", "ran" and "runs" represent the root word "run". These words would be counted as different unless they were stemmed into its roots. The information retrieval of the English documents is improved by 6% - 8% after using a stemming process [11].

EnglishStemmer from nltk.stem.snowball package was used in the project for stemming the words into its root representation.

IV. DOCUMENT FEATURE EXTRACTION

After obtaining the dataset, the next step is to extract several document features. In this step, the raw text data will be transformed into feature vectors to be analysed by the clustering algorithm.

A. Bag of words representation

The raw corpus data cannot be fed directly into most of the algorithms for analysis. It will have to be converted into fixed size numerical feature vectors for it to be processed and analysed. This conversion of raw text data into a collection of numerical feature vectors is known as "Bag of Words" ("Bag of n-grams") representation. The project applies a three-step conversion function. Firstly the words are tokenised into several token terms. Each of the tokens is given an integer id for later identification. Secondly, each token occurrence is counted and stored in a matrix (3D array). Finally, the counted tokens are weighted using TF-IDF as described below. The final matrix represents the input corpus as a set of numerical vectors. This matrix representation is then used by the DBSCAN algorithm to analyse and form the clusters. [12]

B. Count vectors

The count vector represents the data set as a matrix notation. In the matrix, a column is represented by a word or a term in the corpus and a row is represented by an individual document in the corpus. By examining each cell, we can identify the frequency count of a given word in the given text document.

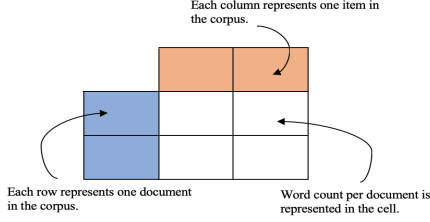


Fig. 2. Count vector representation in NLTK

The *CountVectorizer* class in the *sklearn* package was used to obtain the count vector of the collected dataset.

```
CountVectorizer Array:
[[0 1 0 0 0 0 0 0 1 0 0]
 [1 0 1 1 0 0 1 1 0 0 1 0]
 [1 0 1 1 1 0 0 1 1 0 1 0]
 [1 0 0 1 0 1 1 1 0 0 1 0]
 [1 0 0 1 1 1 0 1 1 0 1 0]
 [1 0 0 1 1 0 0 0 0 0 1 1]]
CountVectorizer vocabulary:
{'testing': 9, 'comments': 1, 'this': 10,
 'is': 3, 'multiline': 6, 'comment': 0, 'on':
 ': 7, 'field': 2, 'single': 8, 'line': 4,
 'method': 5, 'within': 11}
```

Fig. 3. Count vector representation in NLTK

C. TF-IDF vectors

The TF-IDF (term-frequency times inverse document-frequency) represents the weight of a term in the entire document. TF-IDF weighting is essential because some scarce words in the language provide more meaning in the sentence than the other commonly occurring words. This aspect is ignored if we only count the occurrence frequency of a term rather than estimating the weighted value for contributing words.

TF-IDF is a two-step calculation where the TF and the IDF scores are first calculated which is then multiplied to get the final score. TF (which stands for term frequency) of a term t is calculated by dividing the count vector of t in the document (d_x) by the number of documents in the entire corpus (d) with term t . IDF (which stands for inverse document frequency) of a term t is computed as the logarithm value of the total number of the documents in the corpus divided by the total number of documents in the entire corpus (d) with term t . [12]

$$TF(t, d, d_x) = \text{count}(t, d_x) / \text{count}(t, d) \quad (1)$$

$$IDF(t, d) = \log_e(\text{count}(d) / \text{count}(t, d)) \quad (2)$$

```
TF-IDF Vectorizer Array:
[[0.40572238 0. 0. 0. 0. 0.
 0.91399636]
 [0.35748457 0.66038049 0. 0. 0.66038049 0.
 0. ]
 [0.31223453 0.57679018 0.4869659 0. 0.
 0. ]
 [0.35748457 0. 0. 0.66038049 0.66038049 0.
 0. ]
 [0.31223453 0. 0.4869659 0.57679018 0.
 0. ]
 [0.53976033 0. 0.84181874 0. 0.
 0. ]]
```

Fig. 4. Count vector representation in NLTK

$$TF - IDF(t, d) = TF(t, d, d_x) * IDF(t, d) \quad (3)$$

The term t in the TF-IDF score can be scoped into three levels; word level, n-gram level and character level. When the TF-IDF score is calculated considering the word level, it examines each word in all the documents. The n-gram level considers n number of words determined by the user. Finally, at the character level, every character is recognised for the calculation. However, it seems that setting the scope depending on the size of the dataset yields better results. Character level should be considered with a small number of documents, while n-gram level should be considered with a more significant corpus.

TABLE II
TF-IDF SCOPING AND CLUSTER RESULTS

TF-IDF scope level	Corpus size	Number of clusters	Average documents per cluster
Word	14,243	1,354	10.5
N-gram (2)	14,243	654	21.71
N-gram (3)	14,243	432	32.81
N-gram (4)	14,243	371	38.18
N-gram (5)	14,243	332	42.64

V. DISCUSSION

The DBSCAN algorithm was used to cluster a sequence of comment documents. During the initial stage, the number of clusters is unknown. DBSCAN can cluster the documents by analysing the cluster density specified by the neighbourhood size. It creates an arbitrary number of clusters depending on the available data. Since the initial number of clusters is not available or restricted to a predetermined value, the algorithm can analyse the high-density areas to produce uneven and non-flat geometry cluster sizes. DBSCAN is also resistant to noise and robust to outliers. [12]

A. Accumulated noise

One crucial matter I noticed in the project was that the algorithm labels more data as noise if the dataset is small. This behaviour depends on the *min_samples* variable which determines the minimum documents required to form a cluster. The higher *min_samples* value, higher the number of documents classified as noise. If the *min_samples* value is 1, lots of clusters are generated with one document, which is expected since there aren't enough data in a small dataset. This phenomenon declines with the corpus size. Nearly 84% of the documents in a corpus with around 3,000 documents were classified as noise. However, only 72% were classified as noise in a corpus of size 14,000.

TABLE III
NOISE VS CORPUS SIZE (*min_sample*=2)

Corpus size	Noise %
1,231	95%
2,156	89%
3,564	84%
4,675	81%
7,435	79%
13,923	72%
20,328	55%

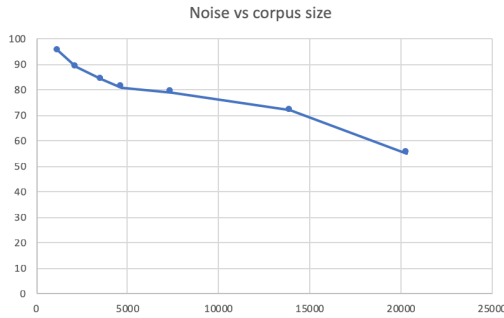


Fig. 5. Noise vs corpus size

Fig. 5 shows that the noise value decreases when a more significant corpus is used. This is expected because it should have access to a bigger dataset for the algorithm to form meaningful clusters. Most notably, the noise percentage started decreasing in an exponential rate when the corpus size was gradually increased up to 20,000. This must be because the algorithm has more data to form more clusters which otherwise would have labelled as noise. Since *min_samples* was set to 2, there wasn't enough data to create a valid cluster when the corpus size was below 5,000.

B. Cluster accuracy and cohesion

The other parameter *eps* value determines the similarity between two different samples. The accuracy of the classification mainly depends on this value. Less value indicates higher

cohesion between the clusters and a higher value indicates a lower coherence. It is advised to determine this parameter depending on the dataset. By testing the algorithm with the obtained comment data (corpus of size nearly 14,000), it seemed that 0.7 is a better value for balancing the cluster cohesion and the validity.

TABLE IV
eps VALUE VS CLUSTER FORMATION

<i>eps</i> value	Corpus size	Number of clusters	Average documents per cluster
0.1	14,252	26	4.4
0.2	14,252	31	4.2
0.3	14,252	41	4.0
0.4	14,252	83	4.16
0.5	14,252	154	4.2
0.6	14,252	258	4.3
0.7	14,252	398	4.42
0.8	14,252	561	4.78
0.9	14,252	728	5.60

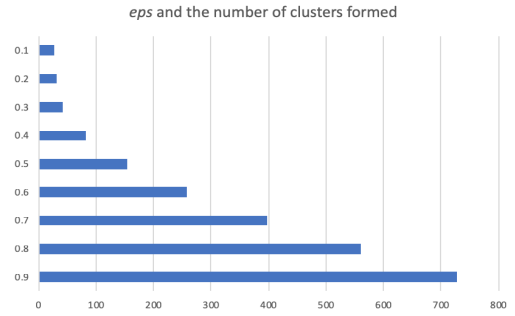


Fig. 6. Noise vs corpus size

The attached Appendix B lists examples of the clusters formed with the *eps* value of 0.7 and *min_sample* of 3.

VI. EVALUATION

A. Threats to validity

As described in Section V, the clustering performance and accuracy is improved with a bigger corpus size. The more data the algorithm has access to, better the results are. However, text classification seemed a very resource intensive task which requires decent computing power. The algorithm was tested with a comment set of around 20,000 comments obtained from the Android Open Source Project, Firebase Admin API and Apache Hadoop Project. A typical computer with Dual-Core i5 processor and 8GB of RAM takes around 1 hour to process the given 20,000 comments. Using a variety of projects with a lot more comment data would produce a better-unbiased results set.

B. Future Work

1) *Improving the model for supervised learning:* This project attempts to cluster a given sequence of comment data. As described in SECTION I, this technique is known as unsupervised learning. As opposed to the unsupervised approach, the supervised learning approach uses a pre-clustered set of documents to model its dataset. The clusters formed by this project can be used as a model for future supervised learning techniques to predict and identify more comment patterns that exist in the current software projects.

2) *Performance improvements:* The program requires a computer with reasonable processing power to process an adequate amount of data. The main focus of this project was to develop a program to collect comments in a source file and to cluster them. There may be places in the software which can be optimised further to perform document clustering with limited resources. The program utilises around 5-6GB of RAM to handle a document set of approximately 20,000 documents. Currently, all comment data and generated Count and TF-IDF vectors are stored in the primary memory. This data can be cached elsewhere to release temporary memory for other tasks.

3) *Extending to other languages:* The proposed source code scanner only works with Java source files with the file extension `.java`. This can be improved furthermore to capture the comments in other languages as well. Adopting the algorithm for Java-like languages which uses the same lexical structure for the comments should be a simple modification to enable the algorithm to process different file extensions. Java Source Scanner class (class responsible for collecting comments from a source file) can be modified for other languages which use a completely different syntax. It is implemented as a Python class, and by utilising OOP concepts, we can extend the same class to provide support for other languages.

REFERENCES

- [1] Z. Faguo, Z. Fan, Y. Bingru, and Y. Xingang, "Research on short text classification algorithm based on statistics and rules," in *2010 Third International Symposium on Electronic Commerce and Security*, July 2010, pp. 3–7.
- [2] H. Isahara, "Resource-based natural language processing," in *2007 International Conference on Natural Language Processing and Knowledge Engineering*, Aug 2007, pp. 11–12.
- [3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, ser. Springer Texts in Statistics. Springer, 2017, no. DOI 10.1007/978-1-4614-7138-7.
- [4] G. J. Badros, "Javaml: a markup language for java source code," *Computer Networks*, vol. 33, no. 1, pp. 159 – 177, 2000. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1389128600000372>
- [5] IBM, "Ibm jikes java compiler framework." [Online]. Available: <http://jikes.sourceforge.net>
- [6] R. Koschke, R. Falke, and P. Frenzel, "Clone detection using abstract syntax suffix trees," in *2006 13th Working Conference on Reverse Engineering*, Oct 2006, pp. 253–262.
- [7] C. Thunes, "Pure python java parser." [Online]. Available: <https://github.com/c2nes/javalang>
- [8] Q. Li and X. Huang, "Research on text clustering algorithms," in *2010 2nd International Workshop on Database Technology and Applications*, Nov 2010, pp. 1–3.
- [9] X.-W. Liu, P.-L. He, and H.-Y. Wang, "The research of text clustering algorithms based on frequent term sets," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 4, Aug 2005, pp. 2352–2356 Vol. 4.
- [10] Z. Yuhang, W. Yue, and Y. Wei, "Research on data cleaning in text clustering," in *2010 International Forum on Information Technology and Applications*, vol. 1, July 2010, pp. 305–307.
- [11] P. Han, S. Shen, D. Wang, and Y. Liu, "The influence of word normalization in english document clustering," in *2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, vol. 2, May 2012, pp. 116–120.
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.