Automatic Classification of Computing Literatures via Article and Reference Correlation

Oluwafemi Oriola[1, *], Lawrence Ojo[1], Ojonoka Atawodi[2]

[1]Department of Computer Science, Adekunle Ajasin University, Akungba-Akoko, Nigeria

[2]School of Computing, University of Southern Mississippi, Hattiesburg, US

Email address:

oluwafemi.oriola@aaua.edu.ng (Oluwafemi Oriola), lfojoo@gmail.com (Lawrence Ojo), erikaojonoka@gmail.com (Ojonoka Atawodi)

[*]Corresponding author

To cite this article:

Oluwafemi Oriola, Lawrence Ojo, Ojonoka Atawodi. Automatic Classification of Computing Literatures via Article and Reference Correlation. *American Journal of Computer Science and Technology*. Vol. 5, No. 4, 2022, pp. 204-209. doi: 10.11648/j.ajcst.20220504.12

Abstract: Automatic literature classification via machine learning has witnessed increasing attention in various research circles, especially computing community because of the availability of large body of research articles in diverse fields. Existing works have largely drawn features from segments of articles such as abstracts, contents and their metadata with little or no attention for references. This paper posited that correlating article and reference features would enhance the performance of machine learning algorithms. Therefore, we exploited the correlation of TFIDF of articles and references using association rule and cosine similarity-based correlation methods for classification of computing literatures. We focused on Adekunle Ajasin University Research Repository. Based on the ACM's and Denning's taxonomies, the research articles in the database were labelled by experienced computing professionals. Logistic Regression, Support Vector Machine and Multilayer Perceptron Neural Network with N-Gram features were explored as classifiers. For ACM's taxonomy, the highest accuracy and F1-score of 0.56 and 0.41, respectively were obtained for association rule-based correlation; 0.62 and 0.51, respectively for similarity-based correlation; and 0.59 and 0.46, respectively for the existing article-based classification. For Denning's taxonomy, the highest accuracy and F1-score of 0.41 and 0.40, respectively were obtained for association rule-based correlation; 0.41 and 0.36, respectively for similarity-based correlation; and 0.38 and 0.37, respectively for the existing article-based classification. These results show that both methods of correlation have better prospect than the popular abstract-based classification method in automatic classification of computing literatures.

Keywords: Computing, Research Articles, Machine Learning, Classification, Reference Features

1. Introduction

As manufacturing, construction and engineering industries are adopting artificial intelligence for automation, it is also increasingly being explored in publishing and editing [1]. Artificial intelligence components like machine learning have been severally used for abstracting, classification, author identification, text recommendation, translation, editing and bibliometric studies [2]. In fact, many of the works in natural language processing and computational linguistics are now focused on these areas.

Literatures in the context of this work refer to published works in books, book chapters, journals, conference proceedings, encyclopedia, reports, patents, monographs and theses. They usually consist of title of work, preface, body of work (articles), references or bibliography and other information. The earlier library series classification systems depend largely on manual bibliometric methods, which are impractical in this era of digitalization [2]. Across various disciplines, automatic literature classification is required because of the large volume of research articles in various publication databases and increased demand for innovativeness in the aftermath of COVID-19. The automatic classification of literature has been extensively studied in [3-7].

Supervised machine learning techniques [1, 8-10] and deep learning algorithms [11, 12] have yielded good performance in the automatic literature classification. However, much is still required in the aspect of identification and engineering of features to improve the classification algorithms' performance.

In previous works, article features such as abstracts based on text summarization and topic models [9], keywords [10] and their meta-data [1] have been explored. We posit in this work that expanding the article features with reference features through correlation would improve the feature distribution and classifier performance.

Specifically, this work focuses on an intranet-supported research repository of publications of scholars within the Adekunle Ajasin University Library. The database serves as a hub for accessing published works of members of the University community and measuring scholars' progress with respect to research investments. Therefore, this work is part of an effort to improve the query process in the database system. As a first task, we narrow the work to Computing literatures because of the comparative advantage of the discipline over others, availability of standard classification systems and benchmarking purpose.

The main contribution of this paper is the combination of article and reference features with the aid of two correlation models for improved Computing literature classification. The first relies on similarity-based method, while the other is based on association rules.

2. Related Works

Since none of the existing works has combined article and reference features for classification of research publications which is the focus of this paper, this section focuses on the review of articles-based classification models.

In order to automatically classify *Caenorhabditis Elegans biological* literatures, Support Vector Machine was used as classifier and phrase-based clustering algorithm was used to create cluster label for each class in [8]. Abstract features, topic distribution and topic words were extracted through Latent Dirichlet Allocation to classify scientific publication and the results showed that abstract features outperformed other models [9]. Keywords, authors, co-authorship, publishing journals parameters and meta-data were employed as features to classify 1.5 million research articles [1]. The proposed model outperformed the previous AdaBoost and Support Vector Machine algorithms. Keywords N-grams showed good performance having been explored to classify research articles in Microsoft Academic Research dataset [10].

In Chowdhury and Schoen [13], Support Vector Machines, Naïve Bayes, K-Nearest Neighbor, and Decision Tree classifiers were evaluated for classification of published articles into Science, Social science and Business categories. Based on Bag-of-Words and TFIDF, the results showed that every other algorithm apart from Decision Tree performed well in terms of accuracy, Precision, Recall and F1-score.

The use of abstracts has been shown to be reliable in classification of large repository of CiteSeerX's scientific literatures in [11]. The authors applied attention deep learning model and the results showed that word-level embeddings weighted by term frequency inverse document frequency outperformed character and sentence-level embedding models. Also, GloVe showed better performance than Bag-of-Words and Word2Vec in automatically labelling of Computer Science and Computer Engineering literature dataset [12].

3. Classification Model and Evaluation

This section presents the classification model and highlights the experiment carried out to evaluate the model. The architecture of the classification model is illustrated in Figure 1. The architecture consists of a research repository containing articles and references. With the aid of correlation models, including similarity- and association rules-based correlation models, the reference features are correlated with the article features. A Neural Networks algorithm is recommended for classifying the features.
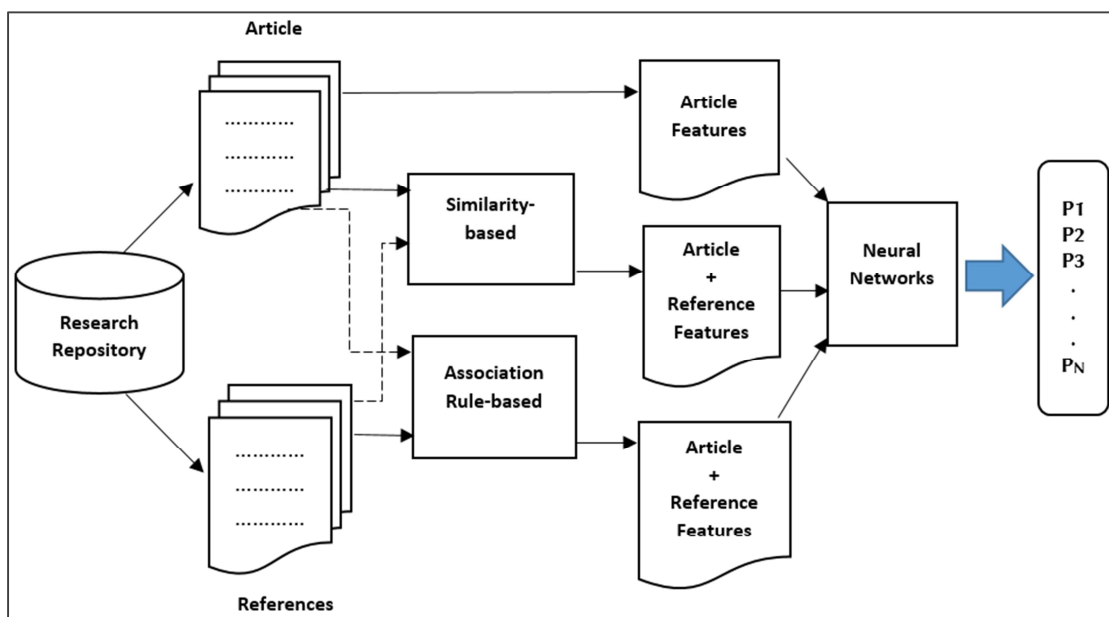

Figure 1. Architecture of the Classification Model.

3.1. Article and Reference Correlation

In the correlation process, the word features are transformed to term frequency inverse document frequency (TFIDF). The TFIDF vector for a given term t in a document (article), D is given as:

$$TFIDF = \sum_{i \in D} tf_{i,D} \times \log\left(\frac{N}{DF_i}\right) \qquad (1)$$

Where

$tf_{i,D}$ is total number of occurrence of a term t in position i in document D; DF_i is the number of documents containing a term t in position i; and N is the total number of documents in the corpus.

3.1.1. Association Rule-Based Correlation

Given set of articles, $D_i = \{a_1,\ldots, a_N\}$, set of references, $D_j = \{r_1,\ldots, r_M\}$ and set of classes, $C = \{c_1,\ldots, c_K\}$. The D_j vector is added to the D_i vector space if the following association rules hold:

$$\frac{n(D_iD_j)}{N} \geq 0.5 \qquad (2)$$

and

$$\frac{n(D_iD_j)}{n(D_i)} \geq 0.5 \qquad (3)$$

3.1.2. Similarity-Based Correlation

Given set of articles, $A = \{a_1,\ldots, a_N\}$, set of references, $R = \{r_1,\ldots, r_M\}$ and set of classes, $C = \{c_1,\ldots, c_K\}$. The D_j TFIDF is added to the D_i vector space if:

$$Sim \geq 0.5 \qquad (4)$$

The Cosine Similarity (Sim) is estimated as the cosine of the angle between the article D_i and reference D_j vectors.

$$Sim = \frac{D_i \cdot D_j}{\|D_i\|\|D_j\|} \qquad (5)$$

3.2. Model Evaluation

This involves Data Collection, Data Preprocessing, Feature Extraction, Machine Learning Classification and Performance Evaluation.

3.2.1. Data Collection

The data used for the evaluation of the models are collected from Adekunle Ajasin Research Repository, an intranet-supported database containing articles of scholars in various disciplines in the University. An assessment of the publications shows that 175 articles with 3,346 references are associated with the Department of Computer Science's scholars. By annotation involving five experienced computing professionals trained in the act of bibliometric classification, the articles are labelled according to the ACM and Denning [14] Computing Classification Systems. With selection criteria based on Fleiss Kappa Inter-Rater Agreement Scores and minimum number of instances per class of 6, 112 articles

and 2,213 references are used in ACM, while 95 articles and 1,875 references are used in Denning. The Inter-Rater Agreement Scores for ACM and Denning are 0.86 and 0.75, respectively. The labels of the ACM data fall within Artificial Intelligence (AI), Information Systems (IS), Software Engineering (SE), Programming Languages (PL), Computer Communication Networks (CCN) and Operating System (OS). The Denning's labels fall within Organizational Informatics (OI), Database and Information Retrieval System (DIR), Software Engineering (SE), Architecture (AC) and Artificial Intelligence (AI).

3.2.2. Data Pre-Processing

In order to clean the datasets, unwanted terms such as punctuations, numbers, symbols, and English stop words are removed. We also perform stemming using NLTK [15] and change the uppercase words to lower case.

3.2.3. Feature Extraction

In this paper, the N-grams (terms) are engineered by weighting Word N-gram over TFIDF.

$$D = N\text{-gram (t) x TFIDF (t)} \qquad (6)$$

Specifically, unigram, bigram and trigram features are combined.

3.2.4. Machine Learning Classification

Each of the two article datasets is divided into two samples [training sample= 70%; testing sample = 30%]. With the training sample, a training model is built using grid-search hyper-parameter optimization over 10-fold cross validation and tested with the testing sample. The training samples are added to the corresponding reference data from the correlation models. The process of testing with the previous testing sample is also carried out on association rules-based, similarity-based correlation data for both ACM's and Denning's Computing Classification Systems.

In evaluating the models, Logistic Regression (LogReg), Support Vector Machine (SVM) and Simple Neural Networks known as Multilayer Perceptron (MLP) are used for the classification and evaluated with sets of parameters as presented in Table 1 to obtained optimal performance. The classification models are implemented with Scikit-Learn Python library [16].

Table 1. *Parameters and Values of the Machine Learning Models.*

ModeL	Parameters	Values
LogReg	C	log(-4), log(4), log(20)
	Penalty	L1, L2
	Solver	Liblinear
SVM	Kernel	Linear
	C	0.001, 0.01, 0.1, 1, 10, 100
	Parameters	Values
MLP	Batch Size	32, 64, 128
	Epoch	10, 50, 100

3.2.5. Performance Evaluation

Accuracy and Macro-averaging F1-score (F1-score) are

used to evaluate the performance of the models in line with existing works. Both Precision and Recall are not used since they are captured in F1-score. The formulas for accuracy and F1-score are presented in Eq. 7-10.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{7}$$

$$F1\text{-}score = \frac{2X\,(Recall_m\;X\;Precision_m)}{(Recall_m+Precision_m)} \tag{8}$$

Given that:

$$Precision = \frac{\sum_{i=1}^{l}\frac{TP}{TP+FP}}{l} \tag{9}$$

$$Recall = \frac{\sum_{i=1}^{l}\frac{TP}{TP+FN}}{l} \tag{10}$$

Where TP is true positive; TN is true negative; FP is false positive; FN is false negative; and K is the number of classes.

4. Results

In Table 2, the test results of the prediction for ACM's Computing Classification System is presented. The results show that the highest accuracy of 0.62 and F1-score of 0.51

are recorded by similarity-based correlation model with MLP follow by article-based model with MLP. The Association rule-based model performance is worst. In each model, MLP performs better than LogReg, while SVM is worst. The best results are highlighted in bold.

Table 2. *Prediction Performance in ACM's Computing Classification System.*

Model	Classifier	Accuracy	F1-score
Article-based	LogReg	0.32	0.13
	SVM	0.29	0.08
	MLP	0.59	0.46
Association Rule-based Correlation	LogReg	0.50	0.36
	SVM	0.29	0.08
	MLP	0.56	0.41
Similarity-based Correlation	LogReg	0.44	0.28
	SVM	0.29	0.08
	MLP	**0.62**	**0.51**

Figure 2 presents a chart for the comparison of the best results for every model based on the ACM's Computing Classification System. The chart indicates that Similarity-based correlation model records the highest performance in terms of accuracy and F1-score, followed by Article-based model. The performance of Association rule-based correlation is worst.
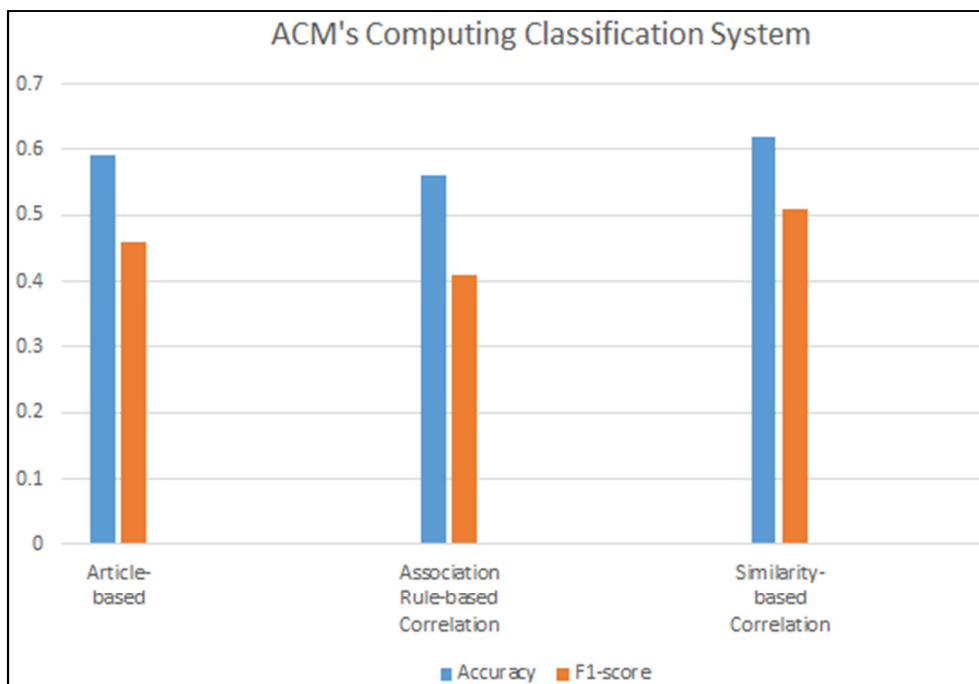
Figure 2. *Comparison of the state-of-the-art Performances for ACM's Computing Classification System.*

Table 3. *Prediction Performance in Denning's Computing Classification System.*

Model	Classifier	Accuracy	F1-score
Article-based	LogReg	0.34	0.21
	SVM	0.24	0.06
	MLP	0.38	0.37
Association Rule-based Correlation	LogReg	0.41	0.36
	SVM	0.24	0.06
	MLP	**0.41**	**0.40**

Model	Classifier	Accuracy	F1-score
Similarity-based Correlation	LogReg	0.41	0.36
	SVM	0.24	0.06
	MLP	0.41	0.35

In Table 3, the test results of the prediction for Denning's Computing Classification System is presented. The results show that the highest accuracy of 0.41 and F1-score of 0.40 are recorded by association rule-based correlation model with MLP follow by similarity-based model with MLP. The

article-based model performance is worst. In each model, MLP performs better than LogReg, while SVM is worst.

Figure 3 presents a chart for the comparison of the best results for every model based on the Denning's Computing Classification System. The chart indicates that Association rule-based correlation model records the highest performance in terms of accuracy, followed by Similarity-based correlation model and lastly, Article-based model. In terms of F1-score, Association rule-based correlation model records the highest performance, followed by Article-based model and lastly, Similarity-based correlation model.
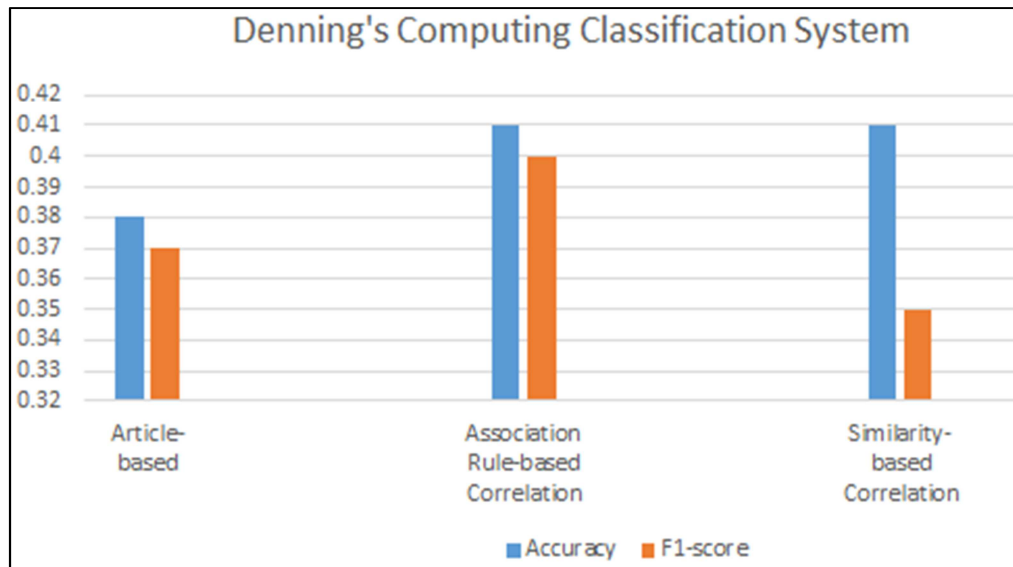

Figure 3. *Comparison of the state-of-the-art Performances for Denning's Computing Classification System.*

The implication of these results in Table 2 and Table 3 is that similarity-based correlation method has higher prospect in automatic classification of Computing literatures than association rule-based correlation method. However, association rule-based correlation method outperforms the previous article-based method in the Denning's Classification System. The best results are highlighted in bold.

The results also show that the proposed methods based on N-gram are better than previous works [9, 10].

5. Conclusion

This paper has focused on improving article-based automatic classification of Computing literatures in Adekunle Ajasin University research repository. The work formulated two models, namely association rule-based correlation and similarity-based correlation for selecting references to merge with articles features. Based on ACM and Denning (1997) Computing Classification Systems, two datasets were created from the research repository. Machine learning classifiers with Word N-gram features were employed to automatically classify the datasets. The results showed that similarity-based correlation model was better than association rules-based correlation. The results also showed that both models have better prospects compared to the popular article-based model. Specifically, neural networks recorded better performance than other machine learning algorithms evaluated.

In order to improve the performance of the proposed models, the entire articles in Adekunle Ajasin University Research Repository would be explored in our future works and this will increase the size of the datasets and the scope of the research. Also, the performance of contextual embeddings and deep learning classifiers would be explored.

Acknowledgements

References

[1] Akritidis, L., and Panayiotis, B. (2013). A Supervised Machine Learning Classification Algorithm for Research Articles. In SAC'13. Coimbra: ACM.

[2] Rivest, M., Etienne, V., and E´ric, A. (2021). Article-Level Classification of Scientific Publications : A Comparison of Deep Learning, Direct Citation and Bibliographic Coupling. PLoS ONE, 16 (5): 1–18. https://doi.org/10.1371/journal.pone.0251493.

[3] Archambault, E., Beauchesne, O. H., and Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In: Noyons, B., Ngulube, P., and Leta, J., editors. Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics, 13: 66–77. http://science-metrix.com/?q=en/publications/conference-pres entations/towards-a-multilingualcomprehensive-and-open-scie ntific.

[4]   Shu, F., Julien, C. A., Zhang, L., Qiu, J., Zhang, J., and Larivière, V. (2019). Comparing journal and paper level classifications of science. Journal of Informetrics, 13 (1): 202–25. https://www.sciencedirect.com/science/article/pii/S17511577 18303298.

[5]   Sjögårde, P., and Ahlgren, P. (2020). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. Quant. Sci. Stud.   1   (1):   207–38. https://www.mitpressjournals.org/doi/abs/10.1162/qss_a_0000 4.

[6]   Waltman, L., and van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. Journal of American Social Information Science and Technology,   63   (12):   2378–92. https://arxiv.org/abs/1203.0532.

[7]   Adele, P., and Alden, D. (2017). Classification of Journal Articles in a Search for New Experimental Thermophysical Property Data: A Case Study, Integrated Material and Manufacturing   Innovations   (2017)   6:   187–196. https://www.doi.org/10.1007/s40192-017-0096-1

[8]   Chen, D., Hans-michael, M., and Paul, W. S. (2006). Automatic Document Classification of Biological Literature, 11: 1–11. https://doi.org/10.1186/1471-2105-7-370.

[9]   Caragea, C., Adrian, S., Saurabh, K., Doina, C., and Prasenjit, M. (2011). Classifying Scientific Publications Using Abstract Features. Association for the Advancement of Artificial Intelligence. https://www.aaai.org/.

[10]   Roul, R. K., and Jajati K. S. (2017). A New Technique Classification of Research Articles Hierarchically : A New Technique. In H.S. Behera and D.P. Mohapatra (Eds.), Computational Intelligence in Data Mining, Advances in Intelligent   Systems   and   Computing   556. https://doi.org/10.1007/978-981-10-3874-7.

[11]   Kandimalla, B., Shaurya, R., Jian, W., and Giles, C. L. (2021). Large Scale Subject Category Classi Fi Cation of Scholarly Papers With Deep Attentive Neural Networks. Frontiers in Research   Metrics   and   Analytics   5   (2):   1–12. https://doi.org/10.3389/frma.2020.600382.

[12]   Pan, Z., Patrick, S., Setareh, R., Zhengtong, P., and Setareh R.. 2022. Ontology-Driven Scientific Literature Classification Using Clustering and Self-Supervised Learning. In Easychair Preprint.

[13]   Chowdhury Shovan and Schoen Marco P. (2020) Research Paper Classification using Supervised Machine Learning Techniques. (2020). Intermountain Engineering, Technology and   Computing   (IETC), https://doi.org/10.1109/IETC47856.2020.9249211

[14]   Denning, P. J. (1997). Computer Science: The Discipline, In A. Ralston and D. Hemmendinger (Eds.), 2000 Edition of Encyclopedia of Computer Science.

[15]   Bird, S., Klein, E. and Loper, E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media, Inc.

[16]   Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., and R. Weiss. (2011). Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research, 12: 2825–2830.