An Investigation of Term Weighting and Feature Selection Methods for Sentiment Analysis

Tuba Parlar[1], Selma Ayşe Özel[2]
1- Department of Mathematics, Mustafa Kemal University, 31060, Hatay, Turkiye.
Email: tparlar@mku.edu.tr (Corresponding author)
2- Department of Computer Engineering, Cukurova University, 01330, Adana, Turkiye.
Email: saozel@cu.edu.tr

Received: November 2017 Revised: January 2018 Accepted: February 2018

ABSTRACT:
Sentiment analysis automatically classifies the opinions, which are expressed in a document, usually as positive or negative. A review document in general, reflects its author's opinion about the objects mentioned in the text. Therefore, it can have many useful applications such as opinionated web search and automatic analysis of reviews. Although sentiment analysis is a kind of text classification problem, structures of review documents are different from texts like news, articles, or web pages; so that techniques applied for text classification are needed to be re-experimented for the sentiment analysis. Assigning appropriate weights to features is important to the performance of sentiment analysis so that important features can receive higher weights for the feature vectors. Feature selection reduces feature vector size by eliminating redundant or irrelevant features to improve classification accuracy. In this study, our aim is to examine the effects of term weighting methods on newly proposed Query Expansion Ranking (QER) feature selection method and also compare the classification results with one of the well-known feature selection method namely Chi-square statistic. We use three popular term weighting methods (i.e., term presence, term frequency, term frequency and inverse document frequency-*tf*idf*) and perform experiments using multinomial Naïve Bayes classifier. The experimental results show that when QER feature selection method is used with *tf*idf* term weighting method, the classification performance improves in terms of F-score.

KEYWORDS: Sentiment Analysis, Feature Selection, Term Weighting, Text Classification.

1. INTRODUCTION

With the rapid growth of the internet and the widespread use of social media applications, it has become increasingly important to evaluate users' opinions. Users opinions which can be on different topics such as a product, a movie or an idea, generate a great deal of information. Automated methods are needed for evaluation and analysis of these large volume documents. Sentiment analysis is an important field of study in which natural language processing and artificial intelligence techniques are applied. Sentiment analysis automatically classifies opinions expressed in review documents as positive or negative using traditional text classification methods. Many researchers prefer supervised machine learning methods because of their ease of use and high classification performance. For example, Pang et al. [1], [2] use Naïve Bayes (NB) and support vector machines (SVM), Nicholls and Song [3] use maximum entropy algorithms for sentiment classification.

In text classification, the methods used to construct feature vectors have impacts on the performance of classification. Pang et al. [1] classify movie reviews dataset using unigrams and bigrams with term frequency and term presence weighting methods. They find that unigrams with term presence weighting method achieve better than the others. In another study [4] researchers investigate the performance of supervised and traditional term weighting methods for sentiment analysis in Turkish. They find term frequency-relevance frequency supervised weighting method achieves better than the other methods.

To classify the increasing number of opinion documents, using feature selection methods for sentiment analysis has become important. Researchers also propose new feature selection methods that can work more effectively for sentiment analysis than well-known statistical methods such as chi-square or information gain [3], [5], [6]. Nicholls and Song [3] propose document frequency difference (DFD) feature selection method and compare its performance with chi-square, count difference [7] and optimal orthogonal centroid [8] methods using maximum entropy classifier. Their DFD feature selection method achieves better results with reduced number of features.

In this study, we examine the effects of term weighting methods on newly proposed QER feature selection method and also compare the classification results with one of the well-known Chi-square feature selection method. After summarizing the previous studies on sentiment analysis, in section 2, we present the methodology that we use for term weightings, feature selection, and classification. In section 3, we introduce the dataset, experimental settings, and evaluation criteria and then discuss our experimental results. In section 4, we provide our conclusions.

2. METHODOLOGY

In this study, our aim is to investigate the relationship between term weighting methods used in term vector construction with the newly proposed QER feature selection method. To reach our goal we apply three term weighting methods that are term presence, term frequency, and term frequency inverse document frequency to construct feature vectors. We use two rank based feature selection methods that are chi-square and QER to show the interactions between the feature selectors and term weighting methods.

2.1. Feature Selection Methods

Feature selection methods try to find a subset or good features by using wrappers or rank-based techniques. Wrappers select the best feature set by applying a search method with a classifier over a training set. Rank-based methods, on the other hands, are independent of classifiers; they assign ranks to features according to an algorithm and highly ranked features are selected for the classification process. Thus, the classification accuracy and effectiveness are improved by choosing the best subset or the top ranked features which contain more meaningful information.

Chi-square ($\chi 2$)

Chi-square statistic is a commonly used feature selection method in sentiment analysis [3], [8]. Chi-square statistic of a feature *f*, is the weighted sum of chi-square scores of feature *f* for each class c_i in the dataset. If the feature has a low score, then the feature can be eliminated because it contains less information.

The chi-square score for each feature *f* in the dataset is calculated as follows:

$$\chi^2(f) = \sum_{i=1}^{m} P(c_i)\chi^2(f, c_i) \qquad (1)$$

where *m* is the number of classes in the dataset, $P(c_i)$ is the probability of class ci in the dataset, and $\chi^2(f, c_i)$ is the chi-square score of the feature *f* for the class c_i which is computed according to the 2-by2 contingency matrix for the feature and the class.

Query Expansion Ranking (QER)

Information retrieval is an important field of study aimed at selecting relevant documents or texts according to a given query. Researchers have developed a variety of query expansion techniques for finding more relevant documents to a given query. Harman [9], [10] works on assigning scores to terms extracted from relevant documents to expand the original query and improve precision of information retrieval strategies. Therefore, terms extracted from relevant documents are scored and top scored terms are chosen as the most valuable terms to include expanded query.

QER feature selection method [6] is developed based on these scoring formulas proposed by Harman [10]. In this method, score of a feature is computed according to (2), where the probability of feature *f* for positive class documents is calculated by p_f, while the probability of feature *f* being in negative class is calculated by q_f. Features are then ranked by using the score formula:

$$score_f = \frac{p_f + q_f}{\left|p_f - q_f\right|} \qquad (2)$$

$$p_f = \frac{df_+^f + 0.5}{n^+ + 1.0} \qquad (3)$$

$$q_f = \frac{df_-^f + 0.5}{n^- + 0.5} \qquad (4)$$

Positive and negative class probabilities for feature *f* are calculated by using (3) and (4) where df_+ is the number of documents that contain *f* in the positive class, df_- is the number of documents that contain *f* in the negative class, n^+ is the number of documents in the positive class, n^- is the number of documents in the negative class.

2.2. Term Weighting Methods

It is important to create feature vectors in text classification. Each document *d* is represented by a feature vector (f_1,f_2,…f_m) where a weight value is assigned to each feature in a document. In vector space model, different methods are used to assign weight values. The most commonly used methods are term frequency (*tf*), term presence (*tp*), and term frequency*inverse document frequency (*tf*idf*).

Term presence (*tp*) deals with the existence of a feature in the document, it does not matter how many times it occurs. Term Frequency (*tf*) is concerned with the number of times a feature is observed in a document. Although a feature that is more frequent than other features is considered more valuable, it may not be a valuable term for the document because it may be a stop word. Term frequency*inverse document frequency (*tf*idf*) value is calculated highest when a feature occurs many times in a document but is less common in other documents. Inverse document frequency is calculated as in (5).

$$idf_f = \log \frac{|D|}{df} \quad (5)$$

where *df* gives the number of documents that contain the feature *f*, and |D| is the number of documents in the dataset. Therefore, *idf* gives the specificity of the feature.

2.3. Classification

Sentiment analysis classifies the opinions in the review documents as positive and negative. Machine learning techniques are widely used in sentiment analysis such as Naïve Bayes, support vector machines, maximum entropy. Naïve Bayes classifier has two models: multinomial and Bernoulli. Naïve Bayes Multinomial (NBM) model is preferred in this study as it has more successful results for text classification.

3. EXPERIMENTAL RESULTS

3.1. Dataset

Experiments are conducted on Turkish movie and products review datasets. The Turkish movie reviews dataset is collected from beyazperde.com. This dataset consists of 1057 positive and 978 negative review documents [11]. Turkish product reviews dataset is collected in four categories: book, DVD, electronics, and kitchen products from hepsiburada.com. Each category contains 700 positive and 700 negative review documents [12].

3.2. Evaluation Criteria

The evaluation of text classifiers is defined by some terms such as precision (*P*), Recall (*R*), and F-score (*F*). Precision is the ratio of correctly classified documents to a class among all documents by the classifier. Recall is the ratio of correctly classified documents belonging to a specific class to all documents belonging to this class. The F-score is calculated as the harmonic mean of precision and recall as follows [13]:

$$F = 2 \times \frac{P \times R}{P + R} \quad (6)$$

3.3. Results and Discussion

In this study to examine the impacts of the term weighting methods on newly proposed QER feature selection method, features are obtained from the training set with bag-of-words method using only alphabetic characters. For this purpose, we develop a software using Python NLTK library [14]. For the baseline results, we run against NBM classifier for all features using five-fold cross validation. Table 1 shows the total number of features and classification results for each term weighting method according to the NBM classifier. Fig.1 shows the classification performances of three term weighting methods for all datasets.

Table 1. Baseline results in F-scores using NBM classifier for the datasets.

Dataset	Total Feature Size	*tp*	*tf*	*tf*idf*
Movie	18565	**0.834**	0.826	0.797
Book	10500	0.832	**0.832**	0.789
DVD	11334	0.792	**0.793**	0.751
Electronics	10901	0.813	**0.815**	0.800
Kitchen	9436	**0.781**	0.777	0.768

As can be observed from Table 1 and Fig.1, the classification results obtained by *tf* and *tp* weighting methods are very close to each other while *tf*idf* results lack behind them.

Feature selection methods aim to increase the classifier's performance by eliminating non-informative features. To analyze the impacts of term weighting methods over the feature selection process, we reduced the features according to the feature selection methods and then computed feature vectors using each term weighting method. After that we compared the results with the baseline cases given in Table 1 and Fig.1 where we did not apply any feature selection. We chose the most valuable five feature sizes from 500 to 2500 for each method. Fig. 2-6 show the classification results for each term weighting and feature selection methods with reduced feature sizes.

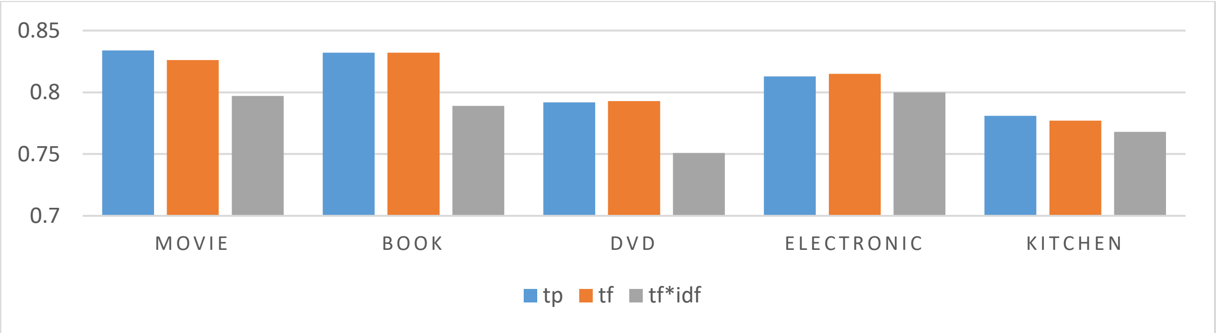

Fig. 1. Effects of term weighting methods using NBM classifier for all features (without any feature selection) in the datasets.

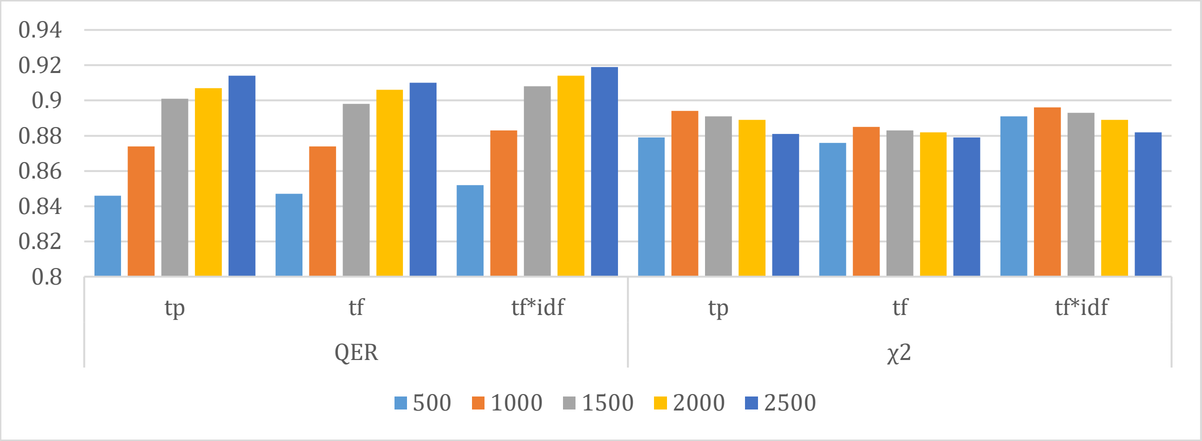

Fig. 2. Effects of reduced feature sizes for term weighting methods using NBM classifier on the movie dataset.

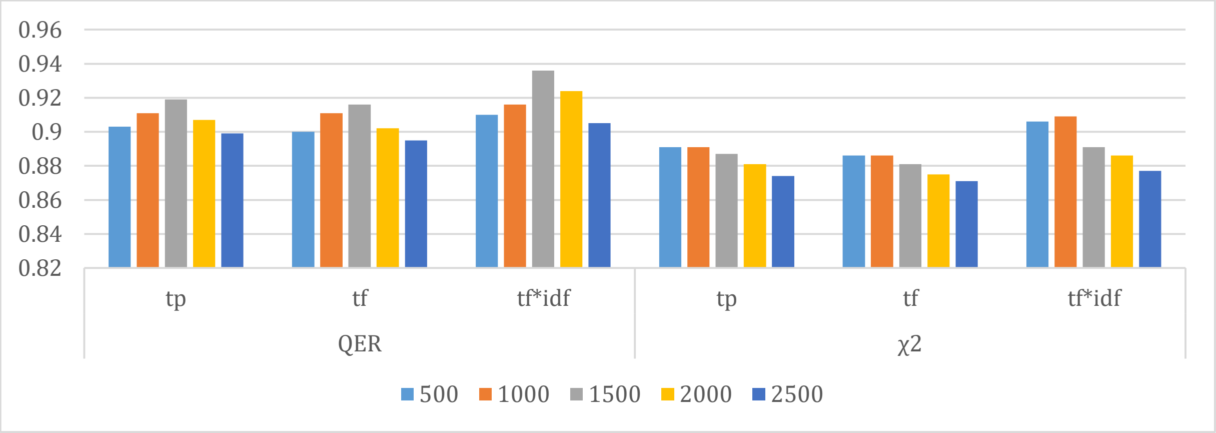

Fig. 3. Effects of reduced feature sizes for term weighting methods using NBM classifier on the book dataset.

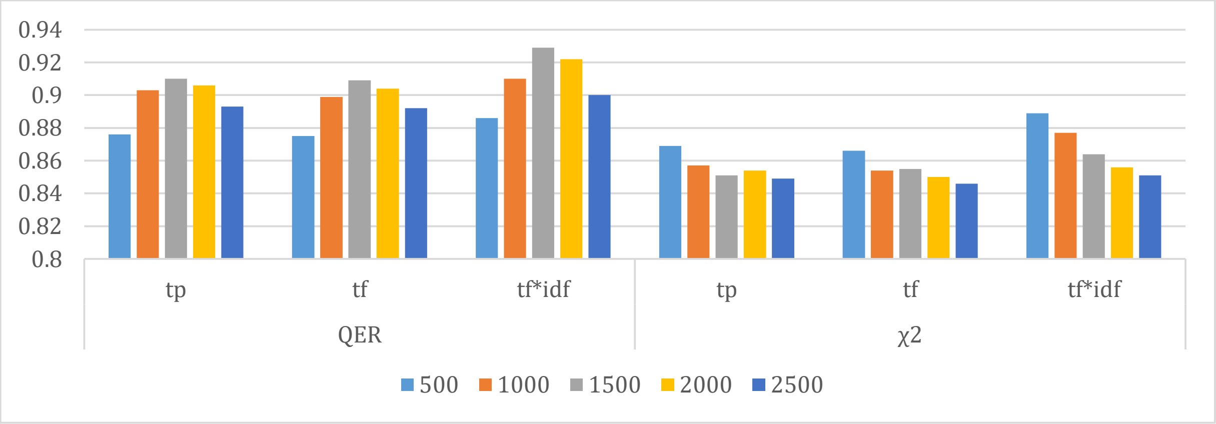

Fig. 4. Effects of reduced feature sizes for term weighting methods using NBM classifier on the DVD dataset.

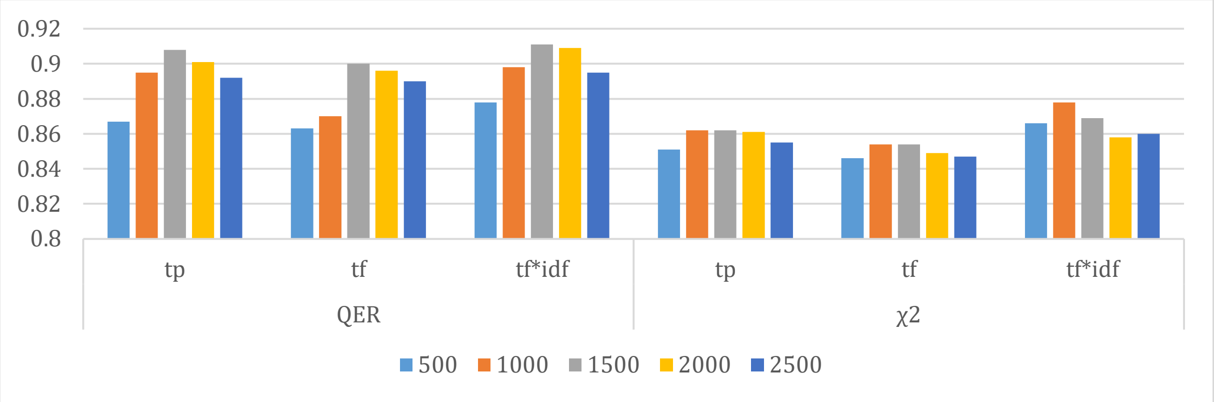

Fig. 5. Effects of reduced feature sizes for term weighting methods using NBM classifier on the electronics dataset.

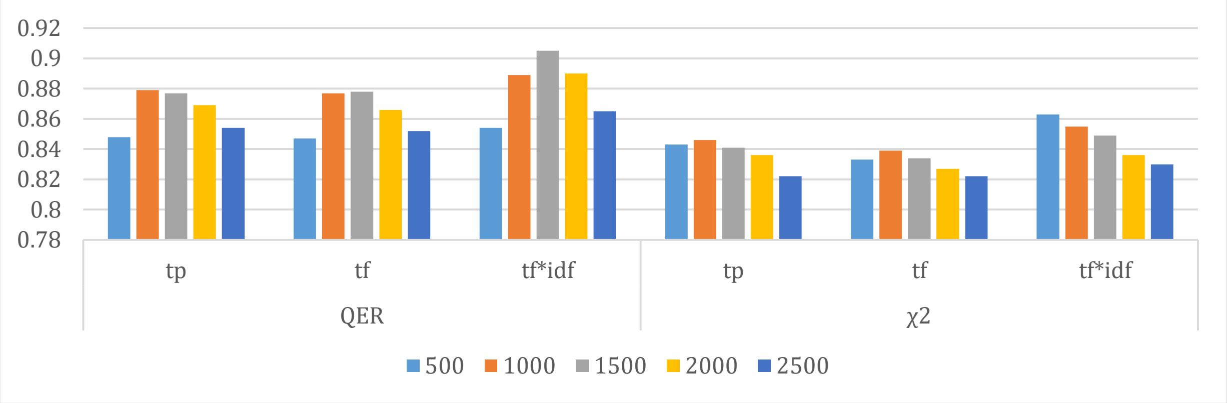

Fig. 6. Effects of reduced feature sizes for term weighting methods using NBM classifier on the kitchen dataset.

As can be observed, the QER method is more successful than $\chi2$ method in terms of F-score for each dataset. QER method achieves the best classification results with *tf*idf* for each dataset with feature sizes 2500 for the movie dataset and 1500 for the product datasets. For the movie dataset, the feature size is larger than the products datasets because the movie review dataset has 18565 features while the product dataset has maximum 11334 features in one dataset (Table 1).

As observed in Fig.1-6, the performance of the classification results has been increased significantly over the baseline results. For example, the F-score of the movie dataset is increased from 0.834 to 0.919 with *tf*idf* using QER feature selection method. Also, it can be observed that $\chi2$ method achieved good results with *tf*idf* method for each dataset.

4. CONCLUSION

In this study, we investigated the effects of term weighting methods on newly proposed QER feature selection method and also compared the classification results with one of the well-known feature selection method namely Chi-square statistic. For baseline, the F-score results using *tp* and *tf* term weighting methods are very similar. As can be observed in figures, the classification performances have been increased significantly over the baseline results. Furthermore, *tf*idf* method has the best classification results with QER method for each dataset. We conclude that *tf*idf* term weighting method can be more discriminating among the selected features. Also, QER method performed better results than $\chi2$ method. Therefore, we can conclude that if feature selection is applied, *tf*idf* weighting method should be preferred, otherwise *tp* or *tf* weighting methods should be used for feature vector computation.

5. ACKNOWLEDGMENT

The author gratefully acknowledges Mustafa Kemal University Academic Research Project Unit (no: 15426) for its support on this paper.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "**Thumbs up?: Sentiment Classification using Machine Learning Techniques**," in *Empirical methods in natural language processing*, 2002, pp. 79–86.

[2] B. Pang, L. Lee, "**A Sentimental Education:**

Sentiment Analysis using Subjectivity Summarization Based on Minimum cuts,” in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 271–278.

[3] C. Nicholls, F. Song, “**Comparison of Feature Selection Methods For Sentiment Analysis**,” in *AI’10 Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence*, Vol. 10, No. 3, pp. 286–289.

[4] M. Cetin, M. F. Amasyali, “**Supervised And Traditional Term Weighting Methods For Sentiment Analysis**,” *21st Signal Process. Commun. Appl. Conf.*, pp. 1–4, 2013.

[5] B. Agarwal and N. Mittal, “**Prominent Feature Extraction For Review Analysis: An Empirical Study**,” *J. Exp. Theor. Artif. Intell.*, Vol. 28, No. 3, pp. 485–498, May 2016.

[6] T. Parlar, S. A. Ozel, “**A New Feature Selection Method For Sentiment Analysis Of Turkish Reviews**,” in *International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1–6.

[7] J. Cai, F. Song, “**Maximum Entropy Modeling With Feature Selection For Text Categorization**,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 4993 LNCS, pp. 549–554, 2008.

[8] J. Yan, N. Liu, B. Zhang, S. Yan, Z. Chen, Q. Cheng, WFan, W. Ma., “**Ocfs: Optimal Orthogonal Centroid Feature Selection For Text Categorization**,” in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’05*, pp. 122, 2005.

[9] D. Harman, “**Relevance Feedback Revisited**,” in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’92*, pp. 1–10, 1992.

[10] D. Harman, “**Towards Interactive Query Expansion**,” in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 321–331, 1989.

[11] B. İ. Sevindi, “**Türkçe Metinlerde Denetimli Ve Sözlük Tabanli Duygu Analizi Yaklaşimlarinin Karşilaştirilmasi**,” Msc. Thesis, Gazi University, 2013.

[12] E. Demirtas, M. Pechenizkiy, “**Cross-Lingual Polarity Detection With Machine Translation**,” in *WISDOM’13*, pp. 1–8.

[13] J. Han, M. Kamber, **"Data Mining: Concepts and Techniques"**, Vol. 54, No. Second Edition. 2006.

[14] S. Bird, E. Klein, and E. Loper, **"Natural Language Processing with Python"**, Vol. 43. 2009.