An Improved Approach for Fast Documents Scrapping and Classifying Using Selenium Automation and Multinomial Naïve Bayes Classifier

Sunny Mehta, Gayatri Jain (Pandi)

1Master of Engineering Student, 2Head of Department
1Computer Engineering,
1L.J.Institute of Engineering and Technology, Ahmedabad, India.

Abstract: Generally, Selenium Automation is used for testing purpose and detecting errors and defects of the system in development. However, we will use Selenium for making a list of required web element from web page. In addition, by using that list selenium will identify the new documents from web pages for scrapping data. For example, we consider any tenders site, there may be thousands of tenders getting published every day, so it is very hard for the user to surf every tender one after another to get the tender of his/her need. But, in our method firstly, we are using bag of words method to gather test data for further classification. Secondly, we are using Multinomial Naïve Bayes Classifier to classify our documents industry wise which will be useful for the user to pick up his category fresh tender. For picking up a fresh tender, user will access the folder created on the desktop where the scraped fresh documents will be stored in a technology wise folder. In the last, Confusion Matrix will be built and detailed accuracy by class for the technology category will be calculated and shown. This approach helps the larger service providing business organizations to provide their clients the documents of their needed categories.

Index Terms – Selenium Web Driver, Multinomial Naïve Bayes Classifier, Bag of Words, Stream Writer, Web Scrapping.

I. INTRODUCTION

1.1 Selenium

Selenium is a portable framework for testing web apps. It also provides a Selenese - a specific language to write tests in a number of popular programming languages, including C#, Groovy, Java, Perl, PHP, etc. The tests can then run on most web browsers. It runs on Windows, Linux, and macOS. It is open source software released under the Apache License 2.0 [1].

Fig-1.1: Modern browsers use to run tests of Selenium Framework.
1.2 Difference between Naïve Bayes and Multinomial Naïve Bayes Classifier

Naive Bayes classifier is the term which refers to conditional independence of each of the features in the model, while Multinomial Naïve Bayes classifier is the specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features [2].

It’s a family of probabilistic algorithms based on applying Bayes theorem with the naïve assumption of conditional independence between every pair of a feature. Bayes theorem calculates probability $P(c \mid x)$ where $c$ is the class of the possible outcomes and $x$ is the given instance which has to be classified, representing some of the certain features [3].

$$P(c \mid x) = \frac{P(x \mid c) \cdot P(c)}{P(x)}$$

1.3 Difference Between Multivariate Bernoulli and Multinomial Naïve Bayes

Multinomial cares about the counts for a multiple feature that do occur, whereas Bernoulli cares about the counts for a single feature that do occur and counts for the same feature that do not occur.

It means that, for example, Multinomial will classify a document based on the counts it finds of the multiple keywords. Whereas, Bernoulli can only focus on a single keyword but, will also count how many times that keyword does not occur in the document.

However, they do model slightly different things. If you have discrete multiple features to worry about, you have to use Multinomial Naïve Bayes Classifier.

1.4 Bag of Words

The bag of words model is a simplifying representation used in natural language processing and information retrieval. In this model, it represents text as the bag of its words, disregarding grammar and even word order but keeping multiplicity. In addition, it has also been used for computer vision.

Generally, this model is commonly used in methods of document classification where the occurrence of each word is used as a feature for training a classifier. [4]

The Bag of Words Representation

![Fig-1.2: Bag of Words Representation.](image)

II. RELATED WORK

Generally, in recent trends in selenium is for the better automated testing purpose and for efficient test results rather than doing a manual testing which is more time consuming and human effort is needed. System automatically learns what actions to be done where on its own. For example, if there is a button it will click there or else if there is a text box then it will enter the required text etc. [6]. System automatically learns what actions to be done where on its own [7]. For example, if there is a button it will click there or else if there is a text box then it will enter the required text etc. The quality of web application is one of important factor while deploying the web applications [8].

Since there is a consistent growth in the volume of digital documents, both on the internet and within organizations, the need to classify them into categories is obvious. So, bag of words technique to represent the tender documents. Classification was implemented by Naïve Bayes classifier [5].
III. PROPOSED MODEL

Figure 3.1 illustrates the working principles of this paper proposed system. Firstly, the system starts with preprocessing the data where the website is automatically simply scrapped documents. After that, text from every document is extracted to create the instances.

Secondly, whole website is started scrapping by selenium automation framework for classifying documents in the respective technological categories.

Thirdly, by using bag of words approach gathering the text for test data and afterwards that data is used in classification of documents in their technological category by using Multinomial Naïve Bayes classifier. In addition, documents will be stored in technologically classified folders.

At last, system will calculate the detail accuracy by class for technological category.

IV. IMPLEMENTATION

The proposed approach mainly focuses on improving the accuracy for classification based on technological category. The detailed accuracy by class for technology category is calculated based on the following formulas.

1.) \[ TP \text{ Rate} = \frac{TP}{(TP + FN)} \]
2.) \[ FP \text{ Rate} = \frac{FP}{(FP + TN)} \]
3.) \[ Precision = \frac{TP}{(TP + FP)} \]
4.) \[ Recall = \frac{TP}{(TP + FN)} \]
5.) \[ F1 \text{ Score} = 2 * \left( \frac{Precision \times Recall}{Precision + Recall} \right) \]

The proposed method yields the accuracy of 81.481% using which is more than the existing system accuracy using Multinomial Naïve Bayes classifier. The detailed step wise implementation with its snapshots are shown below.
4.1 Preprocessing

Firstly, using Selenium Automation all the DRDO documents from the website are automatically downloaded and from that a list of words was prepared for each file and the unwanted words (and, or, the, but, …) were further removed. A list of 1,49,245 unique words was then built for classification after removing the duplicates from the collective word list of all the documents.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Product_Name</th>
<th>SubIndustry_Name</th>
<th>Industry_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Printer</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>Defence Of Manohar Caves</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>Lighting Panel Scrapping</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>Revival To Pave Scheme At</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Abrasive &amp; Abrasive Media</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>3D Printing System</td>
<td>Abrasive Wheels</td>
<td>Abrasive Wheels</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>Delivery Of Sate Tapes</td>
<td>Abrasive Wheels</td>
<td>Abrasive Wheels</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>Lighting Poles</td>
<td>Abrasive Wheels</td>
<td>Abrasive Wheels</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>Research</td>
<td>Abrasive Wheels</td>
<td>Abrasive Wheels</td>
<td>Machinery &amp; Tools</td>
</tr>
<tr>
<td>2D Image Scanners</td>
<td>Aadhaar Enabled Payment System</td>
<td>Aadhaar Enabled Payment System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Degradation Units</td>
<td>Aadhaar Enabled Payment System</td>
<td>Aadhaar Enabled Payment System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Lighting equipment and electric lamps</td>
<td>Aadhaar Enabled Payment System</td>
<td>Aadhaar Enabled Payment System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Revival Of MeS Scheme</td>
<td>Aadhaar Enabled Payment System</td>
<td>Aadhaar Enabled Payment System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Abrasion Powder</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Dental Equipments and Stores</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Lime Preparation Plants</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Rhinoplasty Sets</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Abortion Equipment</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Dental equipment</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Lime Kin Repair</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Rhinoplasty Instrument</td>
<td>Accounting Software</td>
<td>Accounting Software</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Abrasion Powders</td>
<td>Accounting System</td>
<td>Accounting System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>dental equipments</td>
<td>Accounting System</td>
<td>Accounting System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Lime Quick</td>
<td>Accounting System</td>
<td>Accounting System</td>
<td>IT And Telecommunications</td>
</tr>
<tr>
<td>Rhinoplasty Sets</td>
<td>Accounting System</td>
<td>Accounting System</td>
<td>IT And Telecommunications</td>
</tr>
</tbody>
</table>

4.2 Chrome browser is controlled by automated testing software

![Chrome browser controlled by automated testing software](image1)

**Fig-4.1:** Dataset Sample.

**Fig-4.2:** Chrome browser is controlled by automated testing software.
4.3 Gathering text from documents using Bag of words approach for classification

**Fig-4.3**: Gathering text from documents using Bag of words approach for classification.

4.4 Proposed system output

**Fig-4.4**: Proposed system output.
4.5 Confusion matrix for technological category classification

Fig-4.5: Confusion matrix for technological category classification.

4.6 Detail accuracy by class for technological category

Fig-4.6: Detail accuracy by class for technological category.
V. CONCLUSION

Here the proposed system shows that Selenium Automation can be used not only for testing purpose but it is used for Scrapping purpose also which will be helpful for many users for less time consuming and save manual efforts in browsing data from web pages. Similarly, by using Multinomial Naïve Bayes Classifier, this system can divide each and every document in an appropriate Technological category and improving the accuracy more from the existing methodologies.

REFERENCES

[5] Sumit Goswami, Sunaina Kapoor, Prakriti Bhardwaj "Machine Learning for Automated Tender Classification" IEEE.