Authors

Ananth Nath

Prince Kumar

Rudraksh Syal

Nikeeta Brijwasi

Amrathansh Sharma

success Transformation from text architecture

Team Inspire Analytica

**Abstract :**

Inspire Analytica consults Craigslist on ways to increase the traffic on their job portal. One of the main reasons identified is the lack of structure in job postings. Inspire Analytica, a group of students, suggest ways to improve the structure of these web sites using NLP and machine learning techniques that they acquired over the period of their course ‘Analyzing Unstructured Data’.

**About Craigslist :**

Craigslist was started in 1995 as an e-mail distribution list featuring local events in the San Francisco area. In 1996, It transformed itself to a web-based service to help users find cars, apartments, jobs etc. Since then it has expanded into multiple categories and has become one of the largest online classified portals in the US. In 2000, it started expanding to other US cities and now covers more than 700 US cities and has also expanded in 70 countries.

Craigslist has more than 50 million unique monthly users with more than 20 billion-page views per month making it one of the most visited websites in the USA. The company recorded a revenue of USD 694 million in 2016, with a net profit of USD 500 million.

**About Job Search Market :**

As per the Bureau of Labor Statistics, more than 2 million Americans voluntarily leave their jobs every month and look for new opportunities. Combining these numbers with involuntary unemployment, means that there are a lot of people looking for jobs at any given point of time. The top channels people use to look for new jobs are online job boards (60%), social professional networks (56%), and word of mouth (50%). Global job portal revenue rose 9% in in 2016, reaching $12.4 billion with LinkedIn and Monster as the largest players in the US. Other job portals in the US include indeed.com, simplyhired.com.

**Problem Statement :**

Craigslist also offers the facility of posting advertisements related to jobs on their portal, but this category is not nearly as popular as similar websites like monster.com or indeed.com. It has a lot of job listings that are classified into very broad categories. When a user navigates through the listings, he/she can only see the date on which the job was posted and the title. The title is a user’s free input text, thus, there is not format or structure. Recruiters use fancy words to attract traffic. But they lack the basic information about the position for which the job is posted.

Additionally, users can’t find most of the relevant information such as employment type, compensation, the name of the organization unless he clicks on the link. This lack of information on the search page forces the user to click on the links only to find that the job is not suitable for him, which could be for any number of reasons. Given a large number of job postings, it is humanely impossible to navigate through 100s of pages. Thus, even when there are 1000s of jobs available, because of the lack of text architecture, the purpose of matching a recruiter’s need to user’s remains difficult.

**Proposal :**

Improve the quality and presentation of the job listings on Craigslist website to enhance the user experience, drive more traffic to the website and thus increase the revenue.

We are proposing to change the way these posts are listed on Craigslist so that job-seekers have an easy time navigating the portal and can easily find the relevant jobs. We believe that this will lead to an enhanced user experience, which will drive more traffic to the website and encourage job-providers to list on the website. Considering that job-listings are charged by Craigslist, it will increase the revenue of the client.

**Business Analysis :**

Our project aims at solving the problem by using text mining principles and methods, which will improve the listings of the search result so that they follow a well-defined text architecture and contain information most pertinent to the job.

For comparison, let us look at how the postings look on Craigslist and how it looks on a different portal (indeed.com).

|  |
| --- |
|  |

Fig. 1 : Screenshot Job Postings - Craigslist.org

|  |  |
| --- | --- |
|  |  |

Fig. 2 : Screenshot Job Postings – indeed.com

We can clearly observe that the job postings on indeed.com provide much more relevant information that can assist the job seeker in making their mind and if interested, find more details about the job by clicking on the link. In contrast, Craigslist provides very limited information and one needs to click on the link to find even basic details like salary and company. This drawback limits the use of jobs categories. This also creates issues for employers as getting no response they post the same ad multiple times, but since the ad still provides no new information, response rate is still low. This drives employers away from the portal. Both the scenarios lead to a loss of revenue for the company.

One way of bringing about the change is to modify the format in which users are asked to input the details of the jobs. But since Craigslist has used the same free-style format for a long time, there may be some pushback from the employers on any such change in format. It will also mean more work on the part of the employer, hence they would be reluctant to post in the new format.

To overcome the above challenge, we are proposing to implement a machine learning layer that can take the input provided in the job postings and extract the relevant contents such as job position, type of job, salary etc and create the title from these data.

We have chosen ‘jobs’ section as it is one of the few categories that is monetized by Craigslist and an increase in the postings here will have a direct impact on the revenue. For developing the model, we focused on ‘Business and Management’ section. Although we will be conducting the analysis on one section in the job category, the model can be easily scalable to include other sections.

**Data Analysis :**

We have performed the following steps to complete the data analysis:

1. Data Extraction
2. Modeling and Training
3. Validation

**Data Extraction:**

We have used 2 data sources: Craigslist.org and indeed.com.

From Craigslist.org, we have extracted data about the job postings listed in the business and management category. We have used this dataset to test the model provided by training the indeed dataset. We have used ‘scrapy’ for this purpose.

From indeed.com, we have extracted the title and description of more than a thousand jobs. Security features on the website restricted us from getting a bigger dataset. We tried using ‘scrapy’ but the security features of the website did not allow to scrape the listings. Hence we used ‘BeautifulSoup’ along with the ‘urllib’ to extract information about 1035 data points.

**Data Cleaning :**

On further analysis we found that some of these jobs were repated and hence we removed them from our dataset, leaving us data of 674 postings. These 674 data points had varied variety of titles for the same job position, for example, ‘sales associate’ and ‘sales associates’, ‘Shift Supervisor’ and ‘Shift Manager’. To reduce the number of job titles (classes), we clubbed such titles together. On proceeding, we were left with 293 uniquely identified job titles (classes).

**Modeling and Training :**

We divided the indeed dataset into two parts: we used 70% of the data for training the model and we kept the remaining 30% of the data for testing or model validation purposes.

Steps to pre-process the data to transform to a mathematical form:

1. Tokenize each job description using the ‘tokenize’ function from the ‘nltk’ library.
2. Lemmatize all the tokens generated in Step 1 using ‘lemmatizer.lemmatize’ to get the root of all the words.
3. Remove all the stop words and punctuations in tokens generated from Step 2
4. Combine all the tokens in Step 3, to get a string of tokens to be used for creation of term document matrix
5. Create a term document matrix using the ‘CountVectorizer’ with n-gram range from 1 to 3.
6. Create a TF-TDF matrix using the ‘TfidfVectorizer’ with n-gram range from 1 to 3.
7. Repeat above steps for the test set.

Once we had the data in tokenized form, we ran multiple machine learning models to extract the keywords which can be used in the title of the job. We ran multiple models in search of best models. The models that we tried are:

1. Naïve – Bayes
2. Logistic Regression
3. Random Forest
4. Support Vector Machine

**Naïve - Bayes**

Naive Bayes is one of the simpler techniques for constructing classifiers based on the Bayes theorem. Naive Bayes is used when we have limited resources in terms of computing power and memory. We also use this when the training time is a crucial factor, as it can be trained very quickly. This model is generally outperformed by other more advanced models.

We trained a model that can assign class labels to the job description, which are provided as vectors of feature values. We have done this using MultinomialNB function from

‘sklearn.naive\_bayes’ library.

**Logistic Regression**

Logistic regression tries to classify the text in one of the categories calculating the estimated probabilities using a logistic/sigmoid function. Logistic Regression performs better when the training data size is very large.

We have performed logistic regression using the ‘sklearn.linear\_model. LogisticRegression’ function.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig. 3: Naïve Bayes Fig 4: Logistic Regression

**Random Forest**

Random forests for classification builds decision trees (no. Each tree outputs a category. Then, the final category is decided by a majority voting process. Random Forests are popular because of their interpretability.

We have used ‘sklearn.ensemble.RandomForestClassifier’ with number of estimators as 10 and maximum depth as 25 to generate the model.

**Suppport Vector Machines**

Support Vector Machines is a very popular classification algorithm that extracts a best possible hyper-plane / line that segregates the classes. SVMs are very effective even in high dimensional spaces and in cases where number of dimensions is greater than the number of samples.

We have used ‘sklearn.svm.LinearSVC’ to train a SVM model.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig 5: Random Forest Fig 6. Support Vector Machines

**Validation :**

**On test data-set (30% of indeed data) :**

We applied the above models on the test data and generated the output predicted by the model. Then we compared the results of the prediction to the actual output provided by the models. The results are summarized below:

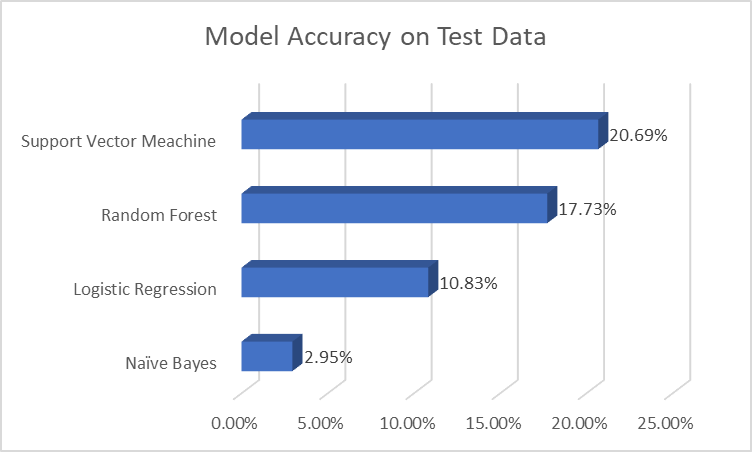


Fig 7. Model Accuracy on Test Data

Based on the accuracy measures as seen in the above table, we decided to use Support Vector Machines to build our tagging model.

Besides accuracy, other reasons we chose SVM Model were that it is time efficient unlike randome forest. The regularization parameter also avoids overfitting on the training data.

**Solution :**

We used the SVM model to predict the titles for jobs listed on Craigslist using the job descriptions. We had a total of 120 data points under ‘Business and Management’ category. Since, the job titles were not available to cross-validate, we manually created these labels based on job description.

Using these manually labelled titles as original data, we calculated the accuracy using predicted values from SVM model.

In the model, we observed that we were able to predict the Job Title accurately for 42 cases out of 120. This puts our accuracy for Craigslist at 35%. These are followed by those jobs who were tagged partially. There were 34 such cases, which accounted for 28% of the jobs.

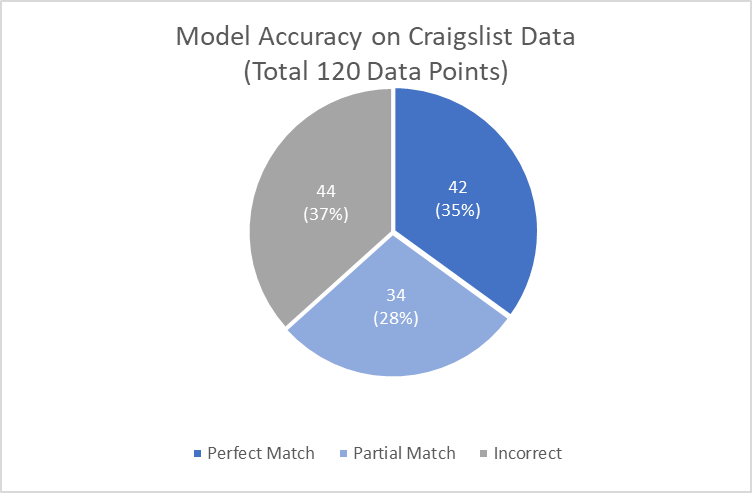
****

Fig 8: Model Accuracy on Craigslist Data

To better understand this behavior, we must first remember that the model works on Key-word tagging. Once, tokenized, sentences no longer retain contextual meaning. This makes it difficult for NLP algorithms to predict all entities in a sentence. In our case, we had 1000 job listings from Indeed and over 300 titles to which these jobs belonged.

Due to a lack of data, the model is not able to identify all keywords associated with various job titles. This is reflected in its behavior while partially tagging reviews.

The performance of the model can be improved by training it on more data. Currently, more than a third of the data consists of unique labels. For building an accurate model we will need more data points with fewer classification labels.

|  |  |
| --- | --- |
| Before | After |

Fig 9: Craigslist site after replacing current job titles with the predicted job titles.

**Conclusion** :

We believe that implementing the model will be very helpful to the Craigslist. It will have an impact on employers as well as the job-seekers. Since the jobseekers will have an easy time navigating the platform and extracting the relevant information without even clicking on the link, they will be able to browse through a larger number of jobs. If the job matches their criteria, they can look for further details by clicking on the link. In the current scenario, they need to click on all the links to find out the details and then decide which leads to lots of missed opportunities. Increased interest from the jobseekers will encourage more employers to post jobs, thus perpetuating a cycle of growth.

**Future Scope :**

The model can be improved even further if we are able to provide tags based upon the generated labels. We can also add tags for Employment Type, Salary etc. With more granulated data, the need to navigate through multiple pages will reduce to just glancing at the search results and filtering out the ones that don’t fit the needs of the user. The advanced model can use our current model for it’s underlying logic. These steps will help Craigslist be a pioneer job hunt website.

**References** :

more than 2 million Americans are voluntarily leaving their jobs every month

https://www.forbes.com/sites/alanhall/2013/03/11/im-outta-here-why-2-million-americans-quit-every-month-and-5-steps-to-turn-the-epidemic-around/#195d94936484

https://www2.staffingindustry.com/site/Editorial/Daily-News/List-ranks-largest-job-boards-global-job-board-revenue-up-9-43941?

https://business.linkedin.com/content/dam/business/talent-solutions/global/en\_us/c/pdfs/Ultimate-List-of-Hiring-Stats-v02.04.pdf