

EMPLOYEE SURVEY ANALYSIS (ESA) SCRIPTS YET ANOTHER NATURAL LANGUAGE PROCESSING APPLICATION

¹Aman Sharma, ²Anuj Yadav

^{1,2}Electronics and Computer Engineering
Dronacharya College of Engineering Gurgaon
Gurgaon, (India)

ABSTRACT

With this paper of our, we have particularly worked on one of the application of Data Analysis. We have proposed a novel method for finding out valuable information out of the bunch of raw data using Python and NLTK libraries. We have processed the Comments of the various Employees of a Company in the form of Raw Data. Each Comment follows different steps such as Cleansing which removes all the mistakes in the comments made by the user, Tagging which tags word according to the different types of verbs or adjectives used in the comments, Chunking which includes selecting a perticular phrase out of the cleansed comment by use of a appropriate grammar rule, Category Generation which includes different types of category generated for the words which can be used for producing different categories user comments. This includes the use of Python as a tool where NLTK is added as a Natural Language Processor which is used for different kinds of languages. You may find the detailed explanation about our methodologies in the later parts of this paper.

Keywords: Python, NLTK, Tokenizer[8], Lemmatizer[9], Stemmer[9], Chunker, Tagger.

I INTRODUCTION

With the growth of IT sector over the past few years, data handling and its analysis had become very difficult. Many companies deals with a large amount of data and they have purchased different tools from different companies like IBM, Microsoft, etc for data storage and its analysis. Data Analysis basically provides us the method to extract some valuable information out of some raw facts. It contains several fields which are required to be undertaken such as removing all the mistakes, converting it into that form which our tool can understand, stating rules for it usage, finding the outcomes and take supportive actions on the basis of these outcomes. The field of Data Analytics is pity vast and have many approaches related to data extraction and modeling and in this paper we will be discussing on the one of the important application of Data Analytics.

Let us better understand what Data Analysis is with example of a person named Lee who had a habit of writing dairy. He started noting each and every incident of his life starting from his birth till now. With the course of time, he have written a lot of information about himself which reflects different stages of his life. Suppose if another person goes through each and every incident of Lee's life and analysis what he used to like when he was below 10 years of age or which part of his life was unforgettable. This analysis of the raw information and find out the valuable information out of it is categorized with the term Data Analytics. I think now we are in a

position to understand the relevant terminologies used in this paper. So I would like to describe the real methodology of our research paper.

II A BRIEF METHODOLOGY

This paper demonstrates a novel method which help user to extract useful information from the bunch of a raw data. It includes a method/ codes which include the use of set of classes and functions which help in extracting a useful data out of input data. There are many useful functions which help in extracting information that are included here. Some of them may be named as, Tokenizer, Taggers, Chunkers, Stemmers, Transformation of Chunkers and Taggers and many more. These methods or classes work on the tool Named as Python 2.7.6 which is required to be downloaded and well configured in the system. Every Code that is executed required to be imported through various packages present in the library. In this project, we have processed the data and produced the different categories out of it and through that we have extracted what user actually meant to say in his/her comments. You may find the detailed explanation as what this paper is all about in later part.

2.1 Python

Python[1] is considered as a high level language, a level ahead of C and C++. It is basically developed for developing applications or scripts for transforming different forms of languages like English, French, German and many more. Python have a unique feature which differentiate it from other languages like C, C++ or java is that it uses white space indentation rather than curly brackets. Currently, the latest version of python in the market is Python 3.4.1 was released on May 18th, 2014. But we have used Python 2.7.6 .

2.2 NLTK

NLTK[3] is described as Natural Language Tool Kit. It comprises of library files in different languages that Python may use for data analysis. One is required to import the NLTK package in the Python Shell so that its library files can be used by the programmer. NLTK includes several features like graphical demonstration of data. Several books have been published on the exotic properties and facilities of NLTK which clearly explains things off to any programmer who is either beginner with python or NLTK or just an expert. NLTK finds several applications in research work when it comes to Natural Language Processing. It helps in processing text in several languages which itself is a big positive for modern researchers.

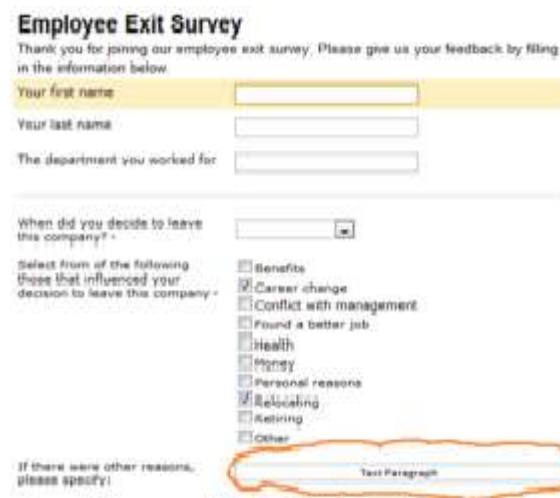
III IMPLEMENTATION OF EMPLOYEE SURVEY ANALYSIS (ESA) SCRIPTS

3.1 What's the Requirement of ESA Scripts.

In Today's world of Globalization and competitions, It is the trend which is followed by every company to organize a Engagement and Exit Survey for its employee within the organization to find out the reasons why people wants to join or leave their company. When any person leaves any company, he/she is required to fill an online survey that comprises of various fields which might be the reasons for his leaving the Organization. In that survey, the questions might be in various forms like Check Boxes, Scroll List, Text plain, etc. It is pity easy to record and analysis those questions which involve answering through Checkboxes or Scroll List but situation becomes very hectic for the person who is analyzing that data if the answer is recorded through Text plains or Text Paragraph. When talking about manually reading, the person, who is reading that data, will be required to

go through each and every employees comments to find what were the reasons why they have left the job. Each company comprises of thousands of employees and it is very common in industries that people moves from one organization to another organization. So, keeping the track of all those employees by just manual reading is a tough task.

Figure 1 – A Screen Shot of Employee Exit Survey[1]



Each company spends a lot of money and resources on their employees on their training and growth and therefore, wants to find the reasons why their best employees are leaving them. Thus, we are in an urgent need of something which can help us finding the reasons why any person is leaving his/her organization. Although, there are several tools in the market by some remarkable companies like IBM. But the major point is they all are paid and hence, require a lot of money to invested to purchase them. In comparison with these paid tools, these Python Script are open source and are free of cost. Any organization can also make changes in the scripts according to its requirement. Hence, it is providing us the best reason why to opt for ESA Scripts.

3.2 Functionality of ESA Scripts

ESA Scripts performs following actions as specified below:-

- It corrects all the Spelling Mistakes.
- It corrects all the Repeated Words.
- It performs Lemmatization, Stemming and Tokenization of Data.
- It performs Antonyms and Synonym Operations on words.
- It find out what kind of Verb, Noun or Adjective is used by the Employee.
- It generates Phrases depending upon the type of Grammar Rule one select.
- Removal of Stop Words.
- Encryption and Decryption of Special Stop Words.
- Removal of ASCII Codes.

There are many more important operations which comes under these above specified operations which are discussed later as their roles comes.

3.3 Next Big Step

First of all, Comments of different employees are taken in a single Column of a CSV file and read line wise. Each Comment comprises of different paragraph having different Spelling mistakes, repeated characters in a word and many more mistakes which are required to be removed before we can find out what person meant in his/her reasons for leaving his/her job.

All the files are required to be stored by .py extension and all the important methods or classes are required to be defined in a single library file so that when using those functions and classes we can import them in a go and use them to do whatever we like to do. These methods/classes are defined in library file named as CustomClassLibrary.py and this file is required to be executed at the top before using any of the function or class so that these classes work accordingly whenever they are called in the main script.

There is yet another important thing that we are required to take care of. You must either place all you scripts in the current working directory or you must provide the path where you have placed your scripts. It is highly required and if we do not provide the path of our scripts properly then it will be going to show errors which will return an error that current file do not exist in our directory.

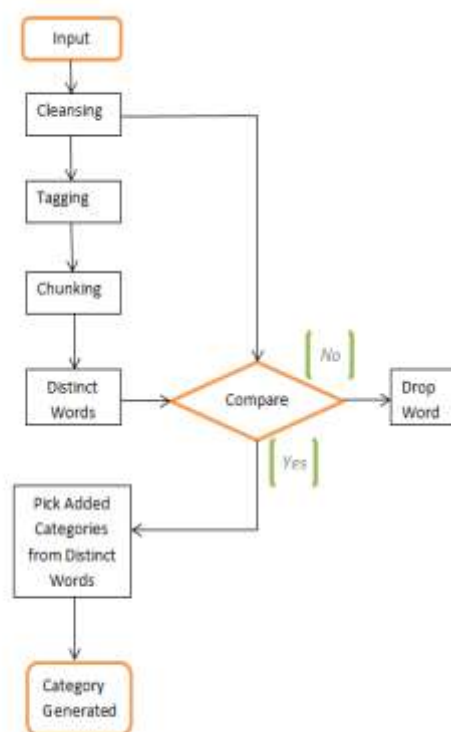


Figure-2 Block Diagram Representing Various Processes to be followed

This Purpose has been divided into 3 Categories which are as follows:

- Cleansing.
- Tagging and Chunking[12].
- Category Generation.

The above described description can be better explained by the figure given below.

3.4 Cleansing

Cleansing, as its name suggests includes the methods which help in cleaning the data which the user has provided. It includes those methods or functions by which one can tokenize data, correct the spellings, remove all the repeated words like if any user wrote 'love' as 'lloovvvee' in a very passionate mode. So they are required to be corrected. There are several Abbreviations that people wrote which are required to be changed to their normal word form. Then there are several stop words in the sentences which do not contribute much to the meaning of that sentence are therefore required to be removed from that sentence. The procedure of this is explained as below.

First of all, we break Paragraph into Sentences and in that procedure some of the words are changed into ASCII Codes which created problem when we further run the process on them and are required to be removed through strip_unicode command. After removing ASCII Codes we tokenize Sentences into words.

Now, explaining each category in detail below.

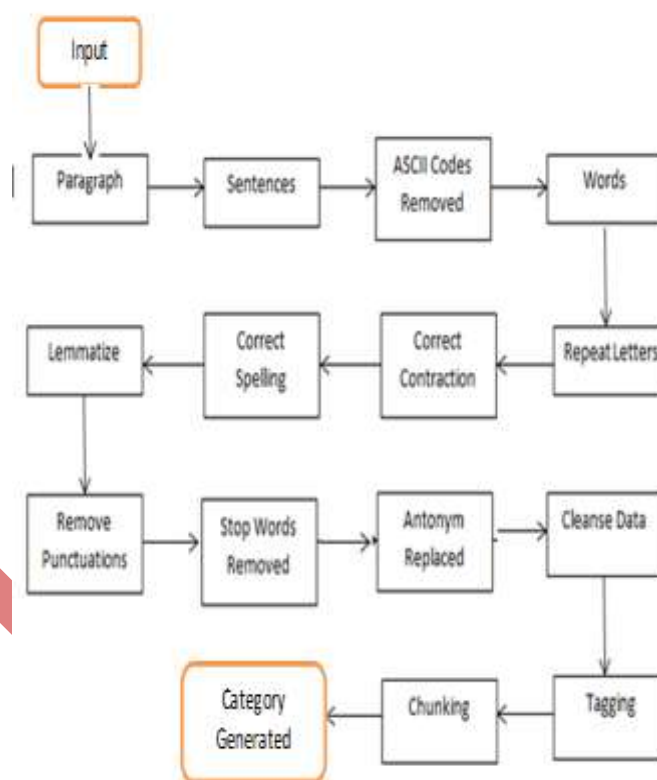


Figure-3 Step wise Explanation for Above Process

These words are processed and all the repeated words like “lloovvvee” are changed to “love” by using repeat replacer function. After that all the short forms or the Abbreviations are changed to their full forms. All the spelling mistakes are required to be corrected before proceeding further. This function is imported using import command and all the methods are required to be defined in our library file named as CustomClassLibrary.py

After correcting all of our spelling mistakes, we lemmatize our word if they are found to be of Noun, Adjective or Verb. For any other category of words, it going to pass the word as it is. After that all the punctuations are removed such as Commas, Exclamation mark, Full Stops, etc.

Here, now we are required to encrypt some of the special words so that they can be used in upcoming process. We will be encrypting some of the words and then removing stop words from that list of words. All those words which do not help in analyzing the sentences like can, could, might, etc are removed from the list of words. Once, Stop words removed, we again decrypt those special words again so that they can be processed now. At this step, we have got the list of words which are going to be passed to do Antonym of words which appears after “not” word.

For Example, “let’s”, “not”, “uglify”, “our”, “code” is changed to “let’s”, “beautify”, “our”, “code”. Thus, we are there with our Cleansed Data.

3.5 Tagging and Chunking

Tagging is a process of allotting different tags to the word in accordance with the part of speech tagging. For this, we have used Classifier based POS tagger[5][10] which is quite a good tagger. When calculated, its efficiency comes out to be over 90% which is quite good. For tagging, we passed the data word wise and find out to which part of speech category it belongs. Either it is a noun or it is a verb or adjective like wise.

We are doing tagging in order to generate tagged word from where we can create a grammar rule so that from them, all the words which comes, forms a meaningful phrase and thus can be wrote in different file.

IV GRAMMAR RULE[11] AND CHUNKING

4.1 Chunk Rule NP

$\{ <RB|DT|NN.*|VB.*>? <VB.*>? <.*>? <JJ.*>? <JJ.*|NN.*>+ \}$

This Chunk Rule can be described as the phrase formed will begin with optional Adverb or Determiner or any kind of Noun or any kind of Verb followed by any kind of optional Verb followed by optional any word followed by any kind of optional Adjective and ending with as many number of any kind of Adjective or any kind of Noun.

4.2 Category Generation

	A	B
1	Word	Category
2	abuse	harassment
3	attack	harassment
4	attitude	harassment
5	baby	personal/family reasons
6	child	personal/family reasons
7	career	lack of growth opportunities/promotion
8	branch	Location issues
9	building	Location issues
10	coffee	stress at work
11	department	work environment
12	dollar	dissatisfaction with salary/benefits
13	economy	dissatisfaction with salary/benefits
14	position	lack of growth opportunities/promotion
15	post	lack of growth opportunities/promotion
16	compassion	stress at work
17	debate	work environment
18	department	work environment

Figure-3 Distinct Categories defined for Chunked words

For Category Generation, we have selected those set of tokenized words which are generated from chunked output. These words are written separately in different file and we manually create category for that. Like if “salary” appears in the file then we have created its category as “salary problem” likewise if “family” appears in

the word then we generated its category as “Personal Issues”. Once this file is created then we compare each and every word of the file and if we find that word in our distinct words file then we are going to generate that category for that word.

Once the category is generated, this category is used to produce the results for the different comments made by user. It is here shown in the figure below.

A	B	C
Input Text	Category	Mapping Word
1. I have been with this company for the very long time. This job could never fulfill my desires for the better career position that's why I am resigning.	Lack of growth opportunities/promotion	career
2. I have been with this company for the very long time. This job could never fulfill my desires for the better career position that's why I am resigning.	Lack of growth opportunities/promotion	position
3. I don't feel very good about what income I was earning here. My income was low. When I shared this with higher management, it didn't bother them. I also wanted to get my shift timings changed as I was not able to spend time with my girlfriend, but my manager did not change my timings.	work environment	girlfriend
4. I don't feel very good about what income I was earning here. My income was low. When I shared this with higher management, it didn't bother them. I also wanted to get my shift timings changed as I was not able to spend time with my girlfriend, but my manager did not change my timings.	disatisfaction with salary/benefits	income
5. I don't feel very good about what income I was earning here. My income was low. When I shared this with higher management, it didn't bother them. I also wanted to get my shift timings changed as I was not able to spend time with my girlfriend, but my manager did not change my timings.	my manager/supervisor	management
6. I don't feel very good about what income I was earning here. My income was low. When I shared this with higher management, it didn't bother them. I also wanted to get my shift timings changed as I was not able to spend time with my girlfriend, but my manager did not change my timings.	work schedule/shift timing	shift
7. My spouse is not well and is hospitalized, so I need to take care of her.	personal/family reasons	spouse
8. I did not see a good work here in this organization. I was expecting some other good work here as per my experience. I am not seeing any learning here. I would like to seek some other opportunity in other company rather than wasting my time here.	Quality of work	experience
9. I did not see a good work here in this organization. I was expecting some other good work here as per my experience. I am not seeing any learning here. I would like to seek some other opportunity in other company rather than wasting my time here.	Lack of coaching/learning	learning
10. I did not see a good work here in this organization. I was expecting some other good work here as per my experience. I am not seeing any learning here. I would like to seek some other opportunity in other company rather than wasting my time here.	Lack of growth opportunities/promotion	opportunity

Figure-4 Categories Generated for different Employees comments

V APPLICATION OF EMPLOYEE SURVEY ANALYSIS (ESA) SCRIPTS

We can do sentimental analysis using this application.

Sentimental Analysis[7]- This is a process of analyzing the sentiments of a person, be it positive, negative or mixed emotions.

We can use the same application for other domains as well like engagement of an employee with the organization.

VI CONCLUSION

This Paper provides a innovative idea which helps in reducing the human efforts as person who is analyzing the data of various employees who had left as now, is not required to go through each and every employee's comments. Thus, by running these scripts we will be able to generate what an employee is talking about, what are the various causes which he found in the company which forced him to resign. Hence, the value of this product goes up when you think analyzing the data of different users of different countries following different languages.

REFERENCES

- [1] <http://123facebooksurveys.com/wp-content/uploads/2011/10/employee-exit-interview-1.png>.
- [2] [http://en.wikipedia.org/wiki/Python_\(programming_language\)](http://en.wikipedia.org/wiki/Python_(programming_language)).
- [3] <https://www.python.org/download/releases/3.4.1/>.
- [4] http://www.nyu.edu/projects/politicsdatalab/workshops/NLTK_Presentation.pdf.
- [5] <https://www.packtpub.com/sites/default/files/3609-chapter-3-creating-custom-corpora.pdf>.

- [6] http://caio.ueberalles.net/ebooksclub.org__Python_Text_Processing_with_NLTK_2_0_Cookbook.pdf.
- [7] <http://fjavieralba.com/basic-sentiment-analysis-with-python.html>.
- [8] <https://www.ics.uci.edu/~pattis/ICS-31/lectures/tokens.pdf>
- [9] <http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>.
- [10] <http://www.monlp.com/2011/11/08/part-of-speech-tags/>
- [11] http://danielnaber.de/language-tool/download/style_and_grammar_checker.pdf.
- [12] <http://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/dongqing-chunking.pdf>.

IJARSE