

Loan Assistant using Text Processing and Natural Language Processing

Poorva Paliwal¹ Pratyush Pare² Priyanshu Sahu³ Kavita Namdeo⁴ Juhi Shrivastava⁵

^{1,2,3,4,5}Department of Computer Science and Engineering

^{1,2,3,4,5}Acropolis Institute of Technology and Research, India

Abstract— This paper aims to analyze and use some of the analytical techniques and tools which can be applied on human language. Nowadays, Natural language processing (NLP) has recently earned much attention for representing, analyzing and modifying text computationally. Its applications are widespread in various fields such as machine translation, detection of spam emails, information extraction, summarization, medical, and question answering etc. The paper analyzes finance related customer queries using text processing and NLP. The output acknowledges the customer about his loan sanctioning status. The goal is to decrease manual processing work in finance sectors and utilize new technologies in efficient way. The end result is to provide customer satisfaction with their experience in dealing with finance facilities.

Keywords: NLP: Natural Language Processing, NLTK: Natural Language Toolkit, NLU: Natural Language Understanding, NLG: Natural Language Generation, TF-IDF: Term frequency-inverse document frequency, ITF: Inverse Term Frequency, ML: Machine Learning

I. INTRODUCTION

Unstructured data, especially text, images and videos contain rich information. However, due to the inherent complexity in processing and analyzing this data, people usually restrain themselves from spending extra time and effort in profound studies of structured datasets to analyze these unstructured sources of data, which can be a potential gold mine [1].

Natural Language Processing (NLP) is all about holding and accessing tools, techniques and algorithms to process and understand natural language-based data, which is usually unstructured like text, speech and others. In this paper, we will be observing tried and tested strategies, techniques and workflows which can be used by practitioners and data scientists to extract useful insights from text data. We will also see some useful and interesting uses of NLP. This paper is all about processing and understanding text data with examples.

II. LITERATURE SURVEY:

The term Text Processing refers to the logic and practice of automation of creation or manipulation of electronic text. The input to natural language processing system will be a simple stream of characters which are Unicode by nature (typically UTF-8). Some processing will be required to change this character stream into a proper sequence of lexical items (words, phrases, and syntactic markers) which can then be used to understand the content in a better way.

The seven phases of text processing are:

1) Sentence Tokenization:

Sentence tokenization (also called sentence segmentation) is the idea of segmenting a string of written language into its component sentences. The concept here is very simple. In

English and any other language, we can split apart the sentences whenever a punctuation mark is seen.

2) Word Tokenization:

Word tokenization (also called word segmentation) is the concept of dividing a string of written language into its component words. In English and many other global languages using some form of Latin alphabet, space can be said as a good approximation of word divider.

3) Text Lemmatization & Stemming:

Text Stemming usually refers to a natural heuristic process that chops off the ends of words in order to derive a stem word. In the hope of achieving the goal correctly most of the time it often includes the removal of derivational affixes. For example, stem word for “beautiful” and “beautifully” is “beauty”. Lemmatization usually refers to doing things properly with the use of a vocabulary, thesaurus and morphological analysis of words, normally targeting to remove inflectional endings only and to provide the base or dictionary form of a word, which is known as the lemma.

4) Stop Words:

Stop words are words which are filtered out before or after processing of text. When machine learning is applied to text, these words can produce a lot of noise. That’s why we need to remove these irrelevant words.

5) Regular Expression:

A regular expression, regex or regexp is a proper sequence of characters that indicates a search pattern. Regular expressions (RE) use the backslash character (\) to point out special forms or to allow special characters to be used without using their special meaning.

6) Bag of Words:

Machine learning algorithms do not work with raw text directly; we need to transform the text into vectors of numbers. This method is called feature extraction [2]. The bag-of-words model is a renowned and simple feature extraction technique used when you work with text. Any information regarding the order or structure of words is discarded. That’s why it is named as a bag of words.

7) TF-IDF:

TF-IDF, short for term frequency-inverse document frequency is a statistical measure used to assess the importance of a word to a document in a collection of works or corpus. The TF-IDF measuring value increases proportionally to the number of times a word appears in the document i.e., frequency of the word but it is offset by the number of documents in the corpus that contain the word.

$$TF(\text{term}) = \frac{\text{Number of times term appears in a document}}{\text{Total number of items in the document}}$$

$$IDF(\text{term}) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term in it}}\right)$$

Finally, combining both the equations:

$$TFIDF(\text{term}) = TF(\text{term}) * IDF(\text{term})$$

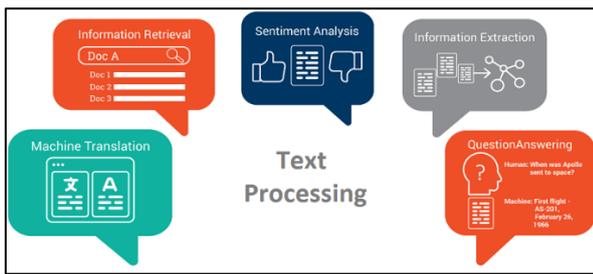


Fig. 1: Elements of Text Processing

III. RESEARCH METHODOLOGY:

A. Natural Language Processing:

Natural Language Processing (NLP) is the basic tract of Artificial Intelligence and Linguistics, completely dedicated to make computers understand the statements or words written in human languages. Natural language processing came into real existence to make easy the user's work and to ease their wish of communicating with the computer in natural language. Since many of the users may be callow or may not be well versed in machine specific language, NLP support those users who do not have enough time to learn new languages or become perfect in it.

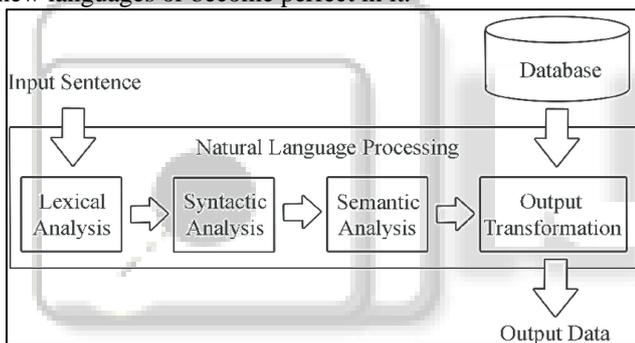


Fig. 2: Working of NLP

B. Benefits of NLP:

The benefits of natural language processing are innumerable. Natural language processing can be applied by companies to improve the efficiency of documentation processes, improve the accuracy of documentation, and identify the most relevant information from huge databases. For example, a hospital might use natural language processing to pull a specific diagnosis from a physician's unstructured notes and assign the treatment.

NLTK (Natural Language Toolkit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to many different work and lexical resources. Also, it contains a stock of text processing libraries for classification, tokenization, stemming, lemmatization, tagging, parsing and semantic reasoning [6].

C. Classification of NLP:

Natural Language Processing majorly can be classified into two parts i.e. Natural Language Understanding (NLU) and Natural Language Generation (NLG) which includes the task to understand and generate the text respectively [4].

NLU is a component of NLP. More accurately, it is a subset of the understanding and comprehension part of NLP. NLG is a kind of technology that simply turns data into Simple-English language.

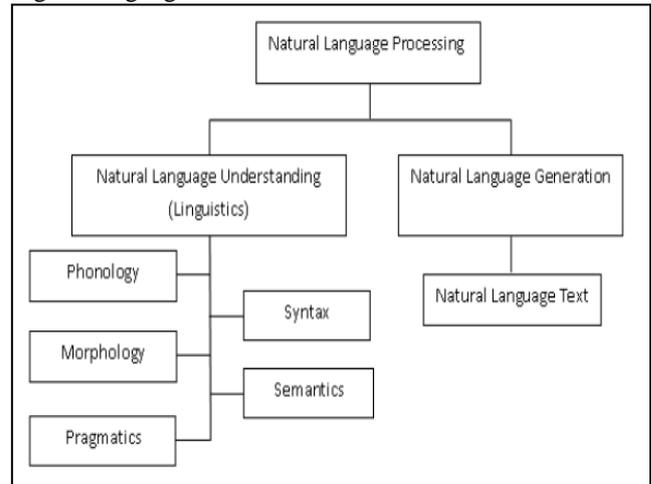


Fig. 3: Classification of NLP [1]

The various important terminologies of Natural Language Processing are: -

- 1) Phonology: Phonology is the part of Linguistics which refers to the proper and systematic arrangement of sound. "Phonology properly is concerned with the functioning, behavior and collective organization of sounds as linguistic items. Phonology consists of semantic use of sound to encode meaning of any language spoken by human.
- 2) Morphology: Morphology which is study of nature of words, are initiated by morphemes. The interpretation of morpheme stays same for all the words, just to understand the meaning humans can break any word into morphemes.
- 3) Lexical: In Lexical phase, both humans and NLP systems, interpret the meaning of individual words.
- 4) Syntactic: This level emphasizes to observe the words in a sentence so as to look over the grammatical structure of the sentence.
- 5) Semantic: Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence.
- 6) Discourse: The discourse phase of NLP engages with units of text longer than a sentence i.e., it does not analyze multi sentence texts as just sequence sentences, which can be explained singly.
- 7) Pragmatic: Pragmatic is concerned with the firm use of language in situations and utilizes central point over and above the central point of the text or matter for understanding the goal.

IV. RESULTS:

This project can handle all the information regarding a person's bank account. It also manages resources which were earlier managed and handled manually. It integrates distinct sections of the organization into consistent manner

so that complex functions can be handled smoothly by any technical or non-technical persons. It aims in automation of finance sector's information, to manage customer account analysis report, to keep record of the transactions done before, consistently update information of all the customers, uploading of digitized records so as to lessen the use of paper, to solve every query of customer regarding his account.

natural-language-processing-for-text-df845750fb63-1-1487a2c8b696

J	A	B	C	D	E	F	G	H	I	J	K	L
1	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Applicant_Income	Co-Applicant_Income	Loan_Amount	Loan_Amount_Term	Credit_History	Property_Area
2	LPO01015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
3	LPO01022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
4	LPO01031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
5	LPO01035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
6	LPO01051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban
7	LPO01054	Male	Yes	0	Not Graduate	Yes	2165	3422	152	360	1	Urban
8	LPO01055	Female	No	1	Not Graduate	No	2226	0	59	360	1	Semiurban
9	LPO01056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural
10	LPO01059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
11	LPO01067	Male	No	0	Not Graduate	No	2400	2400	123	360	1	Semiurban
12	LPO01078	Male	No	0	Not Graduate	No	3091	0	90	360	1	Urban
13	LPO01082	Male	Yes	1	Graduate		2185	1516	162	360	1	Semiurban
14	LPO01083	Male	No	3+	Graduate	No	4166	0	40	180		Urban
15	LPO01094	Male	Yes	2	Graduate		12173	0	166	360	0	Semiurban
16	LPO01096	Female	No	0	Graduate	No	4666	0	124	360	1	Semiurban
17	LPO01099	Male	No	1	Graduate	No	5667	0	131	360	1	Urban
18	LPO01105	Male	Yes	2	Graduate	No	4583	2916	200	360	1	Urban
19	LPO01107	Male	Yes	3+	Graduate	No	3786	333	126	360	1	Semiurban
20	LPO01108	Male	Yes	0	Graduate	No	9226	7916	300	360	1	Urban
21	LPO01115	Male	No	0	Graduate	No	1300	3470	100	180	1	Semiurban
22	LPO01121	Male	Yes	1	Not Graduate	No	1888	1620	48	360	1	Urban
23	LPO01124	Female	No	3+	Not Graduate	No	2083	0	28	180	1	Urban

Fig. 4: Test Cases

V. CONCLUSION:

NLP takes a very important role in new machine human interfaces. When we look at some of the products based on technologies with NLP we can see that they are very advanced and very useful. But various limitations are present. For example, language we speak is highly ambiguous. This makes it very difficult to understand and analyze. Also with many languages spoken all over the world it is very difficult to design a system that is 100% accurate.

These problems get more complicated when we think of different people speaking the same language with different styles. Intelligent systems are being experimented right now. Ideally NLP will influence the development of programming languages, and computer programming will use natural human languages rather than specialized code for development. We will be able to see improved applications of NLP in the near future.

REFERENCES:

- [1] <https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72>
- [2] <https://www.searchtechnologies.com/blog/natural-language-processing-techniques>
- [3] <https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958>
- [4] <https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696>
- [5] <https://towardsdatascience.com/natural-language-processing-nlp-top-10-applications-to-know-b2c80bd428cb>
- [6] <https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part>
<https://towardsdatascience.com/introduction-to->