Natural Language Processing and Python

M/s Purwa Maheshwari
Assistant Professor
ABESIT

Abstract-Natural Language Processing is a subfield of computational linguistics, artificial intelligence and Machine Learning. Since, computers play a great role in transmission and acquisition of information, there is a need to make computers understand natural languages. Technologies based on NLP are gaining widespread acceptance. e.g. Smart phones, other handheld devices are making use of translators, various machine learning approaches for retrieving text written in Chinese or Spanish. Language Processing is emerging to play a central role in this multi-lingual society.

Python is object-oriented, interpreted Language. Python has a very shallow learning curve and its ease of availability online has made its use widespread. This article includes an overview how Python can be used with Natural language Processing to perform simple NLP tasks.

Index Terms— NLP- Natural Language Processing, POS- Part-of-Speech, DIT- Department of Information Technology, nltk- Natural language toolkit, CDAC- Centre for Development and Advance Computing.

——————————   ◆   ——————————

1 INTRODUCTION

Natural Language Processing(NLP) is a field of Computer Science, Artificial Intelligence also called as machine learning and linguistics concerned with the interaction between computers and humans i.e natural languages. In industries as well as academia, there is a need to understand and implement various language and computational linguistics knowledge so that it can be spread worldwide .

Python has a wide range of standard libraries which makes it fit for performing computational and software engineering projects as well . Python is a simple language and in this article we will be able to learn how a small and simple program helps in understanding and analyzing language data. How NLP concepts can be combined with Python in order to deduce the language concepts.

2 LITRETURE SURVEY

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural. languages to perform the desired tasks. Searchable sources available at http://python.org/ and http://www.nltk.org/. Python is simple yet powerful language. It's simple set of commands and libraries makes its use widespread. It has an additional capability of processing linguistic data. Python.org will help you download the latest version of Python for windows. After installing Python, open it and download components of NLTK (natural language toolkit).

2.1 SCOPE

A lot of work has been done in NLP. Reviews of literature on large-scale NLP systems, as well as the various theoretical issues have also appeared in a number of publications example, Jurafsky & Martin, 2000; Manning & Schutze, 1999. Research on NLP is regularly published in a number of conferences such as the annual proceedings of ACL (Association of Computational Linguistics) and its European counterpart EACL, biennial proceedings of the International Conference on Computational Linguistics (COLING).

2.2 TERMS:

2 Before nltk is downloaded, we should be familier with some common terms which are the building blocks of NLP:
3 **Corpus**: large collection of structured set of texts. Text in one language is Monolingual Corpus whereas text in more than one language is termed as Bilingual Corpus.
4 **Lexicon**- Words and their meanings just like a dictionary.
5 **Token**- Entity obtained after splitting up.eg a word if a sentence is tokenized or a sentence if a paragraph is tokenized.
6 **Some basic functions**: *sorted()* gives sorted list of vocabulary items.*len() gives* size of vocabulary. *append()* for adding single atom to list. *index()* for telling the first occurrence of text. *lexical diversity* for repeated calculations on some text avoiding again and again retyping the same formula. *Def* a keyword for defining function. The prompt >>> means Python interpreter is expecting the next command, … prompt indicates that Python expects a code block.
7 Once we have downloaded the *nltk* we have access to the following modules:
8 **Accessing Corpora**- Large set of Text for performing various operations.
9 **Part-of-speech tagging**- Tagging each and every

word according to its part-of-speech such as noun, verb, adjectives, pronoun and so on…

10 **Chunking**- Dividing whole text into small chunks so that operations can be performed easily.

11 **Parsing**- Generating the parse trees for grammars.

12 **Classification-** Grouping the text according to the set to which it belongs. e.g Mango belongs to the group fruit.

2.3 OPERATORS:

2.3.1 a)Relational Operators: Python supports wide range of relational operators for testing the relationship between two values. The are: **<, <=, >, >=, !=, ==** which are pretty much similar to C language. These are also called as Numeric comparison Operator.

b)Word Comparison Operators:

s.startswith(t)- startswith operator tests weather s starts with t.

s.endswith(t)- endswith operator tests weather s ends with t.

s.islower- checks if all characters in s are lowercase.

s.isupper- checks if all characters in s are uppercase

s.isalpha- checks for a non-empty string and all characters in s are alphabetic.

t in s- tests if t is a substring of s.

3 SUCCESS/LIMITATIONS THUS FAR

The most visible results in NLP thus far (last five years) are several commercial systems for database question answering. Enhancements has been made by replacing the fourth generation query languages. Queries and problem solving was dependent on the size of the database, thus limiting the success rate to 80-95%. The success of these systems has depended on the fact that sufficient coverage of the language is possible with relatively simple semantic and discourse models. The semantics are bounded by the semantics of the relations used in databases and the face that words have limited number of meanings in one particular domain. Python has emerged as one of the best object oriented languages in understanding and implementing the linguistic concepts but sky is still too high, a lot of work still needs to be done.

4. Organizations working in the area

There are many organizations , in India as well as abroad which are doing wonders in the area of NLP. Listing some of them are:

1.Natural Language Group at the Information Sciences Institute.

2. University of Edinburgh Natural Language Processing Group.

3. Stanford Natural Language Processing Group

4. CDAC-Centre for development and advance Computing.

5. Natural Language and Information Processing Group at the University of Cambridge.

6. DIT- Department of Information Technology.

This project is associated with the live project “***ANUVADAKSH***”, under **TDIL (**Technology Development for Indian Languages**)**, programme of **DIT.**

It has the objective of developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier, have reached such a platform through its various projects, where it has a potential to generate utility applications, benefiting the masses, which will enable people to access and use IT solutions in their own language.

5. CONCLUSION:

The impact of Natural Language Processing will be greater than the impact of any other microprocessor technology in the last 20 years. Natural Language is becoming one of the most active field among the research areas. It is even attracting many technical youths year by year. This area leads to detailed study of machine learning and artificial intelligence concepts. Python, and its wide set of library along with Natural language tool kit allows many researchers and scholars for moving forward in the area and make new inventions.

6. FUTURE SCOPE:

This paper will give the basic knowledge about what Python is all about and how one can easily hands-on this language without waiting for any sort of outside support. One can easily start working with Python and also use its library with nltk and enjoy this world of computational linguistics.

7. REFERENCES:

[1] Charniak, E. 1993. Statistical Language Learning. Cambridge, MA: MIT Press.

[2] Allen, J. F. 1994. Natural Language Understanding. Redwood City, CA: Benjamin/Cummings.

[3] Winograd, T. 1972. Understanding Natu-

ral Language. New York: Academic Press.

[4] Weizenbaum, J. 1965. ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine. Communications of the ACM, 9 (1): 36-45.

[5] Kenneth W. Church and Patrick Hanks , 1990 , Word association norms, mutual information and lexicography. Computational Linguistics.

[6] David Chiang. 2005.A hierarchical phrase-based model for statistical machine translation.

.