DESIGN OF NATURAL LANGUAGE PROCESSING BASED WEB PAGE WITH BROWSER PAD

CHANDRA SEKHAR KOPPIREDDY[1], G. VIJAYA KUMAR[2] and R. KRISHNA ROHITH[3]

[1,2]Asst. Professor, [3]Student

Department of Computer

Science and Engineering

Pragati Engineering College

Surampalem, East Godavari, Kakinada

E-mail: chandrasekhar.k@pragati.ac.in

  vijay9908914010@gmail.com

  rohith5453@gmail.com

Abstract

Web scraping is a technique used to extract data from websites and it is the pillar of information retrieval in a world wide web which is the ever growing one. There are two main ways of extracting data from a website: Static and Dynamic Scraping. Static scraping requires input beyond the target website because the user needs to inspect HTML content of the target and find certain patterns in the templates and are then used to extract data. Dynamic scraping is a very broad topic and it is tackled from many different angles: tree-based, natural language processing, machine learning etc. This paper proposes a solution for those who are amenable to flourish a web page. They need not be a technical or a non-technical developer. They just need to have an idea of developing an algorithm. The algorithm can be of any kind. Using natural language processing (NLP) tool kit and algorithm detection tool kit, the software built will take the input algorithm from the user as voice command, processes it and generates the output in the browser.

I. Introduction

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. It is the technology used to aid system to

understand the human's natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.

Static scraping requires input beyond the target website because the user needs to inspect the HTML content of the target and find certain patterns in the templates that are then used to extract data. Static scraping is also very vulnerable to changes in the template of the web page Natural Language Processing is the driving force behind the following common applications: Language translation applications such as Google Translate Word Processors such as Microsoft Word and Grammarly that employ NLP to check grammatical accuracy of texts.

Interactive Voice Response (IVR) applications used in call centres to respond to certain users' requests. Personal assistant applications such as OK Google, Siri, Cortana, and Alexa.

In order to build the software for the proposed idea, we need to develop a notepad and a browser using python qt and python modules and combine them to develop a browser pad application. To this, combine the NLP tool kit and build the final application that takes voice command as the input and generates the output web page from it.

A. Proposed Method:

Firstly, we need to develop a notepad and a browser using python qt and python functions. Then develop a code that combines them to form a browser pad. To this, import NLP tool kit for processing the voice commands from the user. The algorithm must be trained in such a way that it can understand all the components of web page designing say HTML tags and CSS syntactical components. This can be achieved using Machine Learning.

(Note: The voice commands can be either technical or non-technical but make sure that it is understandable by the machine)

B. Evaluation:

The field of NLP is divided into 2 categories:

1. Natural Language Understanding (NLU): Natural language understanding is a smaller part of natural language processing. Once the

language has been broken down, it's time for the program to understand, find meaning, and even perform sentiment analysis. The program breaks language down into digestible bits that are easier to understand. It does that by analyzing the text semantically and syntactically.

2. Natural Language Generation (NLG): The input representation provided to an NLG system may be symbolic (for example, an expert system knowledge base) or numeric (for example, a database containing stock market prices) but it is generally nonlinguistic in nature. Early work in the field relied on the use of hand-crafted knowledge sources, which sometimes meant that the representations used embodied unspoken assumptions. More recent work has been able to take advantage of representations created for other purposes; using these as the input to the generation process reinforces the realization that the elements of the underlying representation may not correspond in a straightforward way to words and sentences.
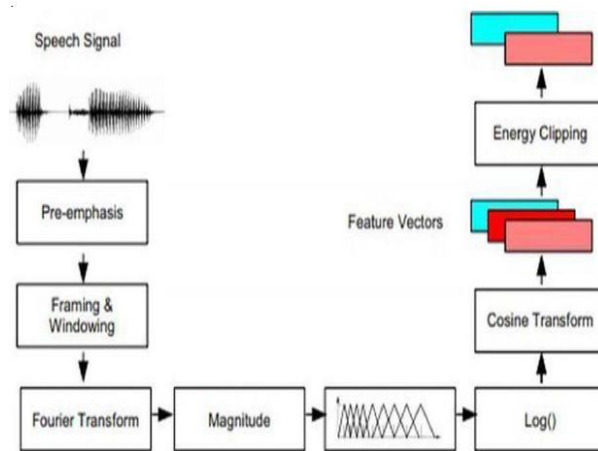
II. Algorithm

Step 1. Develop an application for notebook and program utilizing python qt and python modules.

Step 2. For this application build the code.

Step 3. Next we have to introduce characteristic language preparing toolbox in this segment.

Step 4. After introducing the client need to provide orders to build the

code.

Step 5. The voice order must be in type of specialized or non-specialized calculation.

Step 6. Using calculation recognition toolbox, the voice order of the calculation by the client can be recognized and changed over into the machine justifiable language.

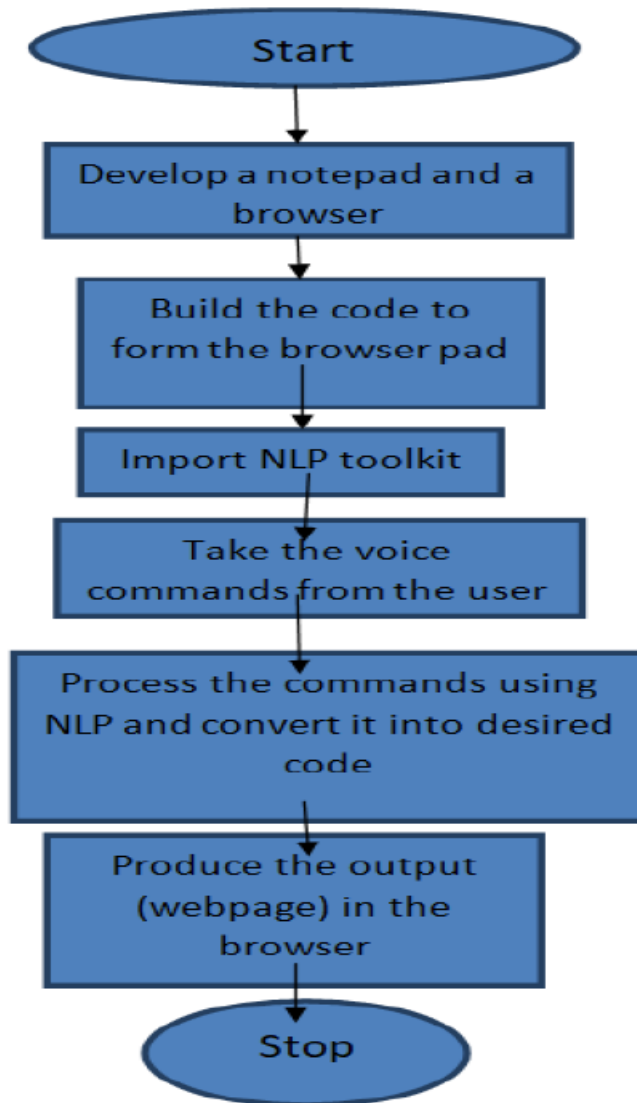**Step 7.** After that transformation the regarded yield should be shown in program

Areas of Natural Language Processing:

- Morphology

- Semantics

- Grammar and parsing (syntactic analysis)

- Semantics

- Pragmatics

- Discourse/Dialogue

- Spoken Language Understanding

Areas of Speech Recognition:

- Signal Processing

- Phonetics

III. Flowchart

```
                    ┌─────────────┐
                    │    Start    │
                    └──────┬──────┘
                           │
                           ▼
              ┌────────────────────────┐
              │ Develop a notepad and a│
              │        browser         │
              └───────────┬────────────┘
                          │
                          ▼
              ┌────────────────────────┐
              │   Build the code to    │
              │  form the browser pad  │
              └───────────┬────────────┘
                          │
                          ▼
              ┌────────────────────────┐
              │   Import NLP toolkit    │
              └───────────┬────────────┘
                          │
                          ▼
              ┌────────────────────────┐
              │    Take the voice      │
              │ commands from the user │
              └───────────┬────────────┘
                          │
                          ▼
              ┌────────────────────────┐
              │ Process the commands using│
              │ NLP and convert it into desired│
              │          code          │
              └───────────┬────────────┘
                          │
                          ▼
              ┌────────────────────────┐
              │   Produce the output   │
              │    (webpage) in the    │
              │        browser         │
              └───────────┬────────────┘
                          │
                          ▼
                    ┌─────────────┐
                    │    Stop     │
                    └─────────────┘
```

Rules to be followed:

The algorithm must be clear and in a proper manner It should not be in a pseudo code format It can be either a technical algorithm or a non-technical algorithm No special character should be included No number should be

included If the technical algorithm is given to the system it should be in a clear manner.

Example: If we need a head tag in the code, then we need to give a voice command as open html tag, open head tag etc.

If a non-technical user is giving a voice command then it must be a language with proper sentence formation and machine understandable.

Example: They can give the model of the web page need in the format of voice command.

IV. Conclusion

This concept of designing webpages using NLP enhances the standards of developing and will be very useful for any people who require a webpage of their own just by knowing the proper usage and guidelines of the application developed. It makes work more easy, comfortable and also time efficient.

References

[1]   J. Alarte, D. Insa, J. Silva and S. Tamarit, Web template extraction based on hyperlink analysis. In: Escobar, S. (ed.): XIV Jornadas sobreProgramación Y Lenguajes, PROLE 2014,Revised Selected Papers EPTCS, Google Scholar 173 (2015), 16-26.

[2]   Q. Liu, M. Shao, L. Wu, G. Zhao and G. Fan, Main content extraction from web pages based on node characteristics. J. Comput. Sci. Eng. Cross Ref Google Scholar 11(2) (2017), 39-48.

[3]   E. Ferrara, P. DeMeob, G. Fiumarac and R. F. Baumgartnerd, Web Data Extraction, Applications and Techniques: A Survey. arXiv:1207.0246v4 [cs.IR], 10 June (2014).

[4]   Open NLP Documentation. https://opennlp.apache.org/docs/.Accessed 2 Nov 2019 [5].https://becominghuman.ai/a-simple-introduction-to-natural-language-processing- ea66a1747b32.

[6]   Bharati, Akshar, Chaitanya, Vineet, Kulkarni, P. Amba, and Sangal, Rajeev. Anusaaraka: Machine Translation in stages. Vivek, A Quarterly in Artificial Intelligence, 10(3), NCST, Banglore. India (1997), 22-25.

[7]   Kommaluri Vijayanand, Sirajul Islam Choudhury and Pranab Ratna, (2002).

[8]   VAASAANU BAADA - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts. Language Engineering Conference. Hyderabad, India. Source:http://portal.acm.org/citat ion.cfm ?id=788716

[9]   Preeti Dubey, The study and development of Macine translation from Hindi language to Dogri language: An important tool to bridge the digital divide. Thesis submitted in the Department of Computer Science & IT, University of Jammu, (2014).

[10] R. M. K. Sinha, R. Jain and A. Jain, Translation from English to Indian languages: ANGLABHARTI Approach. In proceedings of Symposium on Translation Support System STRANS 2001.February 15-17, IIT Kanpur, India (2001), 167-172.