2023 HPCC Systems Community Summit

College	RV College of Engineering ®, Bengaluru
Department	Computer Science and Engineering
Student	Adarsh U
Email	adarshu.cs21@rvce.edu.in
USN	1RV21CS011
Ph.no	+91 9606858375
Project Title	Enhance The English NLP Dictionary using HPCC NLP++ plugins

Proposed Project Synopsis

Introduction:

NLP is a growing field that enables computers to understand and generate human language. It has diverse applications across industries, such as customer service, sentiment analysis, virtual assistants, machine translation, information gathering, and more. NLP++ is a dedicated computer language for building natural language text analyzers, which combines different methods and a hierarchical knowledge base. It allows programmers to emulate human text understanding processes. The VisualText IDE is a developer's environment that utilizes NLP++ and Conceptual Grammar to create text analyzers. However, the current English dictionary struggles to keep up with the evolving language, resulting in a less comprehensive coverage of new words and phrases. This proposal compares NLP++ with other programming languages and showcases its capabilities in natural language tasks.

Objective:

The objective of this proposal is to showcase the superiority of NLP++ over popular programming languages in the realm of natural language processing (NLP) by presenting a comprehensive comparison. The proposal aims to highlight the unique capabilities and advanced features of NLP++ in effectively analyzing and comprehending human language. By demonstrating its versatility and efficiency, the proposal seeks to position NLP++ as the preferred choice for NLP applications in various domains. Additionally, the proposal aims to address the limitations of traditional English dictionaries in keeping pace with the evolving language, emphasizing the need for cutting-edge NLP technologies. Through extensive experimentation and evaluation, the proposal endeavors to provide researchers, developers, and practitioners with compelling evidence of NLP++'s superiority, with the ultimate goal of driving wider adoption and advancement in the field of natural language processing.

Methodology:

The systematic approach to enhance the English language data from the English Wiktionary. The webpages from the English Wiktionary were carefully downloaded and transformed into a consolidated .xml file. Through the utilization of Python scripts, the large dataset was efficiently processed, extracting individual word entries and arranging them in alphabetical order. Rigorous data cleaning techniques were applied to ensure the accuracy and integrity of the information. To implement the obtained enhancement resources effectively, cutting-edge tools like VisualText and Visual Studio Code were utilized. Through extensive testing and training using carefully selected words, the Knowledge Base (KB) was systematically built to facilitate the comprehensive understanding of the English language. The KB, powered by the Conceptual Grammar (CG) Knowledge Base Management System in NLP++, provided a robust framework for storing linguistic, conceptual, and domain knowledge. To further analyze the wikitext entries, a series of well-crafted analyzers were developed, extracting valuable information such as tags, headers, and parts of speech. This systematic and innovative methodology, combining cutting-edge technology and linguistic analysis, ensures the accurate and efficient extraction of valuable insights from the English Wiktionary, setting this project apart as a top contender in the competition.