International Journal of Research in Engineering, Science and Management
Volume-3, Issue-5, May-2020
www.ijresm.com | ISSN (Online): 2581-5792

Kannada Shabhdakosha

Kavyashree R. Bhat[1*], Saritha Shetty[2]
[1]*Student, Department of MCA, NMAM Institute of Technology, Karkala, India*
[2]*Assistant Professor, Department of MCA, NMAM Institute of Technology, Karkala, India*
*Corresponding author: kavyashreerbhat631@gmail.com

***Abstract*: One of the first things required for natural language processing (NLP) tasks is a corpus. In linguistics and NLP, corpus (literally Latin for body) refers to a collection of texts. Corpora are the knowledge base in corpus linguistics. This is a Natural language processing based project which mainly concentrates on creating a corpus of Kannada words. The Kannada Corpus is a language corpus made up of texts collected from the Internet. The project mainly aims to extract different Kannada words from newspapers and pre-process it in different phases and creating a large collection of Kannada words in a text file and making that file available for future work on any Kannada related project like annotation algorithm.**

***Keywords*: Corpus, Linguistics, NLP, NLTK.**

1. Introduction

The key objective of this paper is to create a free corpus for Kannada language named as "Kannada Shabhdakosha" which can be used for Kannada research work in future.

"Kannada corpus" is the collection of Kannada words. This is basically a Natural Language Processing on Kannada language. The Kannada words are collected from different articles of Kannada newspapers present on the Internet. The selection of newspaper is fully a choice of user. User can select any newspaper and copy the link in provided textbox and can start extracting the content from the selected newspaper. The words collected from the newspaper are pre-processed with several steps for removal of unwanted characters. Further the words are compared with the list of stop words, once the words are compared all the stop words found are eliminated.

This corpus contains only the unique words. This project is developed by using Python programming language for both backend and frontend. The extraction of words is compassed from newspapers using a built in package available in Python.

2. Related Work

Creating an effective Kannada corpus is a challenge because there are no built in tool available to process the Kannada language.

Parameswarappa, S., and V. N. Narayana work on "Kannada word sense disambiguation for machine translation." Kannada Corpus tool, a suite of Perl (Program Extraction and Reporting Language) programs implementing an iterative procedure to build Kannada corpora from the web. The procedure requires is, first a set of "seed" words list is built and later a set of "seed" URLs (Uniform Resource Locator) containing documents in the Kannada language is collected by sending queries to commercial search engines (Google and Yahoo) [1].

"Brown corpus" by sketch engine (full name Brown University Standard Corpus of Present-Day American English) was the first text corpus of American English. This corpus consists of 1 million words (500 samples of 2000+ words of each) running text of edited English.

"The Kannada Web Corpus" by sketch engine – is a language corpus made up of texts collected from the Internet.

3. Methodology

Python language has rich set of packages to work in the field of Natural Language Processing. Some of the methods and packages used are:

A. NLP

Natural Language Processing (NLP) is a branch of AI that helps computers to understand, interpret and manipulate human language. NLP helps developers to organize and structure knowledge to perform tasks like translation, summarization, named entity recognition, relationship extraction, speech recognition, topic segmentation, etc.

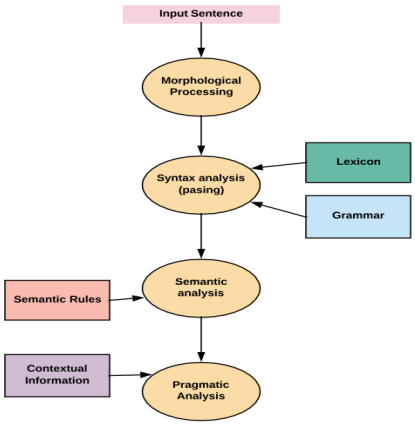

Fig. 1. Flow chart for NLP

B. NLTK

NLTK stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Tokenization,

Stemming, Lemmatization, Punctuation, Character count, word count is some of these packages which will be discussed in this tutorial.

C. *Tokenization*

Tokenization is the process by which big quantity of text is divided into smaller parts called tokens. Natural language processing is used for building applications such as Text classification, intelligent Chabot, sentimental analysis, language translation, etc. It becomes vital to understand the pattern in the text to achieve the above-stated purpose. These tokens are very useful for finding such patterns as well as is considered as a base step for stemming and lemmatization.

4. Implementation

Implementation is the stage of the project where the theoretical design is turned to a working system. Implementing Kannada Shabhdakosha has following steps:

A. *Extraction*

In this the user can select the required newspaper and by clicking on the button to visit the page the newspaper main page is opened and user can select the link and copy the link back in textbox of the application then user finally click the button to start the extract process for Kannada words

B. *Pre-processing*

This is to pre-process the content received from the newspaper article in this phase all the unnecessary characters are removed and only the Kannada words are preserved for further process.

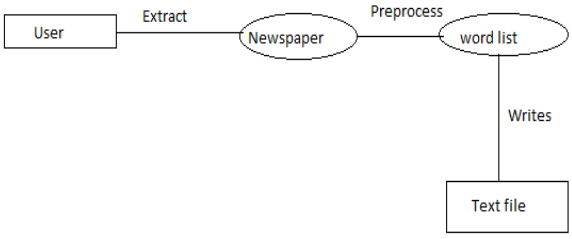

Fig. 2. Flow chart for Kannada Shabhdakosha

C. *Pre-processing*

This is a main phase of the Kannada corpus here all the words received from the preprocess phase are checked for the uniqueness only the unique words taken for further process and again the words are checked for the stops words all the stop words present in the list are removed and then the file are stored in the file.

D. *Display*

The goal of this phase to provide count and list of words based on the user needs like words start from different category.

5. Result

After successfully completing the implementation it is possible to extract the words from different Kannada newspaper and stored the unique word in a text file.

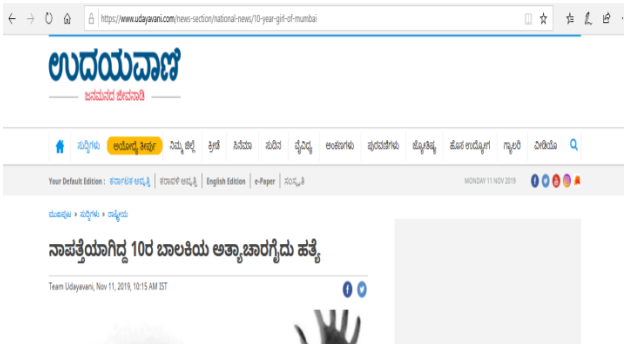

Fig. 3. This is to copy the link of newspaper

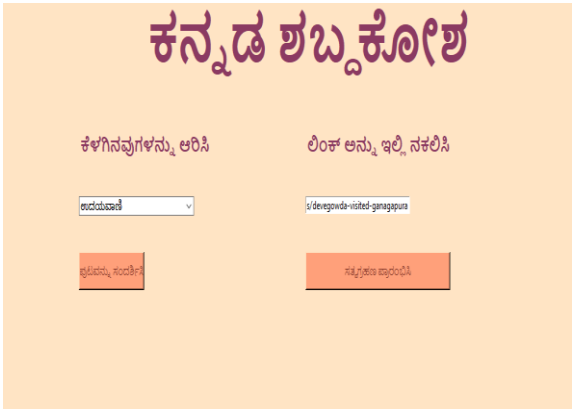

Fig. 4. This is to select newspaper and to copy the link from newspaper to page and to start the extraction of words from newspaper

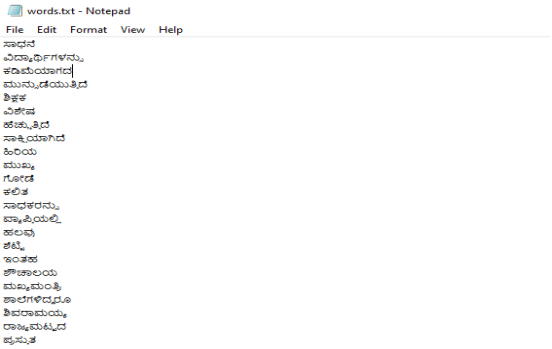

Fig. 5. Output of words which are retrieved from newspaper article

6. Conclusion and Future work

We have looked at the processing of the corpus creation methodologies with areas such as extraction of words from newspapers, elimination of stop words, storing the rest of the unique words in a word file and also we have displayed the count of words based on some required area of interest.

In the near future we can consider the corpus creation methodologies to extend beyond the areas such as extracting from various Kannada articles, annotation like part-of-speech

tagging etc. And also we can make more automation in extraction and storing of words.

References

[1] Parameswarappa, S., V. N. Narayana, and G. N. Bharathi. "A novel approach to build Kannada web Corpus." *2012 International Conference on Computer Communication and Informatics*. IEEE, 2012.

[2] BR, Shambhavi, and P. Ramakanth Kumar. "Kannada part-of-speech tagging with probabilistic classifiers." *international journal of computer applications* 48.17 (2012): 26-30.

[3] Parakh, Mona, N. Rajesha, and M. Ramya. "Sentence boundary disambiguation in Kannada texts." *Language in India, www. Language in India. com, Special Volume: Problems of Parsing in Indian Languages* (2011): 17-19.

[4] Shridhara, M. V., et al. "Development of Kannada speech corpus for prosodically guided phonetic search engine." 2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE). IEEE, 2013.

[5] Parameswarappa, S., and V. N. Narayana. "Kannada word sense disambiguation for machine translation." *International Journal of Computer Applications* 34.10 (2011).