Automatic Sinhala News Text Summarizer

Author: Sajini Tennakoon

Abstract

With the present explosion of news circulating the digital space, which consists mostly of unstructured textual data, there is a need to absorb the content of news easily and effectively. While there are many Sinhala news sites out there, no site facilitates recommendation despite the popularity of recommender systems in the current age and day. Therefore, it is effective if the news were presented in a summarized version which tally with the user preferences as well. Our research aims to fill this gap by providing a centralized news platform which recommends news to its users clearly and concisely. The news articles were collected using web scraping and after performing categorization it will be presented in a summarized context. Also, we expect to detect the grey sheep users and to provide separate recommendations to them in order to minimize errors in recommendation. Here the grey sheep users refer to the user group who have special tastes and they may neither agree nor disagree with majority of users. By implementing the proposed system, we hope to provide appropriate solutions to the mentioned requirements and build a user-friendly Sinhala news platform. Considering about the application, manually creating a summary can be time consuming and tedious. The main idea behind building an automatic text summarization is to distinguish the highest significant information from the given content, decrease of the offered text to fewer sentences without leaving the fundamental thoughts of the first content and present it to the end-readers. Implementation of a specific summarizer for Sinhala Language is a major requirement to develop such an application because there is no Sinhala news platform available which presents summarized and categorized news text to the users. The objective is to produce a brief and exact outline of voluminous news messages while focusing on the key thoughts that convey beneficial information without losing the general significance. The research aims to build the summarizer with the use of PyTeaser algorithm. Even though PyTeaser can not be directly used for Sinhala Language, by using language specific modifications, PyTeaser is made available for Sinhala. The logic behind the PyTeaser includes assigning a total score to each sentence based on four features: Title Score, Keyword Frequency, Sentence Length and Sentence Position. Total score is computed by weighting the mentioned features and those weights are constants. Then the sentences with the highest scores are selected to produce the summary. The quality of the summary is evaluated using F Measure, with the use of human-generated summaries which are produced by Sinhala experts. The research focuses to compute the F Measure for all the possible weight combinations made out with original weights of PyTeaser and choose the optimized weight vector which provides the best quality news summary. The summaries for the proposed application are generated using the derived weight set and then those news are expected to recommend to endusers via the recommender system, according to user preferences.