Features of Initial Processing of Georgian Texts for NLP Tasks

M.Archuadze, M.Tsintsadze, M.Khachidze

maia.archuadze@tsu.ge, magda.tsintsadze@tsu.ge, manana.khachidze@tsu.ge

Department of Computer Sciences Faculty of Exact and Natural Sciences
Iv. Javakhishvili Tbilisi State University
University str., 13, Georgia

Text preprocessing (text component analysis, tagging, morphological and syntactic analysis, information generation, machine translation, etc.) is an integral and essential part of almost all NLP tasks [??]. The success of solving NLP tasks depends on the accuracy of it results. The process is generally considered to be split into two stages of initial processing of text : Sorting and Segmentation. These stages, in turn, include certain actions that require consideration of the various natural features of the text, which relate to the language in which the document is to be produced.

Due to the morphological complexity of Georgian language makes it difficult to apply existing methods and algorithms to solve different NLP tasks without modification and sometimes development of new methods. This paper discusses all the basic steps of initial text processing for NLP tasks and the features that accompany Georgian texts. The proposed Algorithms incorporate these features into the process and form the basis of the software package.

Keywords: Text preprocessing, Stop Words

Referance:

1. Kurdi, M.Z. (2016). Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax, Vol. 1. John
2. Gurusamy, V. and Kannan, S. (2014), 'Preprocessing Techniques for Text Mining', Conference Paper, No. October 2014.
3. M.khachidze, M.Tsintsadze, M. Archuadze, G.Besiashvili. Concept Pattern Based Text Classification System Development for Georgian Text Based Information Retrieval. Baltic J. Modern Computing, Vol. 3 (2015), No. 4, pp. 307–317.
4. M.khachidze, M.Tsintsadze, M. Archuadze. Natural Language Processing (NLP) Based Instrument for Classification of Free Text Medical Records. BioMed Research International. Volume 2016 (2016), Article ID 8313454, 10 pages. http://dx.doi.org/10.1155/2016/8313454