Centroid-Based Language Identification Using Letter Feature Set

Hidayet Takcı[1] and İbrahim Soğukpınar[2]
[1]Gebze Institute of Technology 41400 Gebze /Kocaeli-Turkey
htakci@bilmuh.gyte.edu.tr
[2]Gebze Institute of Technology 41400 Gebze /Kocaeli-Turkey
ispinar@bilmuh.gyte.edu.tr

Abstract. In recent years, an unexpected amount of growth of the text documents volume has been observed on the internet, intranet, in digital libraries and newsgroups. To obtain useful information and meaningful patterns from these documents, a great many researchers known under the term "text mining" have been carried out. Among them text categorization is to be mentioned that covers the problem of classifying documents relative to their similarities. One of techniques applied in this area is called centroid-based document classification method. All researchers on text categorization use the notion of frequency somehow or other. In this study, letter frequencies (LF) have been used for text categorization. By making use of letter frequencies information, the centroid-based document classification has been carried out. An experiment has been done on language detection for text documents. Its results allow propose that the letter-based text categorization should be done prior to term based text categorization.

1. Introduction

With the current spread of worldwide access, the volume of available texts increased written in different languages. Automated treatment of these texts that anyway requires natural language processing, necessitate a preliminary identification of the language used [2]. Language identification problem can be seen as a specific instance of the more general problem of an item classification through its attributes [3].

Language identification is one of the text categorization applications [8]. In language identification study languages will be pre-defined categories. So, we can identify a language by using a text categorization algorithm. As far as centroid-based document classification is one of techniques of text classification algorithm can be applied with the purpose of language identification. It has a linear time complexity, and it is easy for use [7].

Generally, in language identification studies, short words or common words [9], n-grams [4, 11], unique letter combinations [10] etc. are used as feature set. Usage of too many features is a disadvantage for fast language identification processes. Instead, we could alternatively use letter feature sets. In a study, it has been mentioned that, letters could be used for characterization of documents [5]. In addition to, sometimes, word based identification techniques are not easily applied to Japanese and Chinese.