Jordan University of Science and Technology

Automatic Arabic Text Categorisation: A Comprehensive Comparative Study

Authors: Ismail Hmeidi Mahmoud Al-Ayyoub Nawaf A. Abdulla Abdalrahman A. Almodawar Raddad Abooraig Nizar A. Mahyoub

Abstract: Text Categorization or Classification (TC) is concerned with placing text documents in their proper category according to their contents. Due to the various applications of TC and the large volume of text documents uploaded on the Internet daily, the need for such an automated method stems from the difficulty and tedium of doing such a process manually. The usefulness of TC is manifested in different fields and needs. For instance, the ability of automatically classifying an article or an email into its right class (Arts, Economics, Politics, Sports, etc.) would be very appreciated by individual users as well as companies. This paper is concerned with TC of Arabic articles. It contains a comparison of the five best known algorithms for TC. It also studies the effects of utilizing different Arabic stemmers (light and root-based stemmers) on the effectiveness of these classifiers. Furthermore, a comparison between different data mining software tools (Weka and RapidMiner) is presented. The results illustrate the strong accuracy provided by the SVM classifier, especially when used with the light10 stemmer. This outcome can be used in future as a baseline to compare with other unexplored classifiers and Arabic stemmers.