Abstrak

Dalam kehidupan umat Islam, pemahaman terhadap isi Al-Qur’an sebagai pedoman hidup, adalah hal yang sangat penting. Al-Qur’an sebagai referensi utama umat Islam pada umumnya tertulis dalam bahsa Arab. Untuk melakukan proses teks, salah satu preproses awal adalah tokenisasi. Sehingga semua proses teks mensyaratkan dilakukan tokenisasi terlebih dahulu. Pada bahasa Indoensia dan bahasa Inggris tokenisasi dapat dilakukan dengan sederhana, dimana token dibatasi dengan spasi. Namun pada bahasa Arab, sebagaimana juga pada beberapa bahasa lain seperti bahasa Jepang dan Cina, spasi tidak bisa digunakan sebagai batas token. Pada Tugas Akhir ini, akan dilakukan tokensisasi untuk bahasa Arab dengan kasus untuk Al Qur’an. Sebagai *baseline* adalah *Maximum Matching* (atau disebut juga greedy). Pada maximum matching ini dilakukan proses matching terhadap kamus.

Kata Kunci : *tokenisasi, bahasa arab, Al-Qur’an.*

Abstract

In the Muslims life, the understanding of the Holy Qur'an contents as a life guide is very important. Holy Qur'an as the main Muslims references are written in Arabic language. To process the text, one of the initial preprocessing is tokenization. So all the text processing requires tokenization done beforehand. In Indonesian and English languages, tokenization can simply be done, where tokens are bounded by a space. But in Arabic, as well as in several other languages, such as Japanese and Chinese, spaces can not be used as a token boundaries. In this final project, tokenization will be carried to Arabic in a Holy Qur'an case. By using a Maximum Matching (also called Greedy) as a baseline, the matching process was done against the dictionary.

Keywords: *Tokenization , Arabic , Qur'an*