# From Floyd to SPFA

Jiahao Li

**Abstract**

Tokenization is the process of dividing Chinese sentence romanization into tokens that represent the romanizations of each Chinese character. This paper proposes a statistical algorithm based on dynamic programming to address the tokenization problem in the design of Chinese, and more specifically, Cantonese input methods. It is an algorithm with $\Theta(n^3)$ time complexity, in which $n$ is length of input romanization sequence. In the end of the paper, the accuracy of the algorithm is evaluated based on several Chinese language corpora.

## References