Underthesea 1.3.5 - Text Normalization

Vũ Anh

August 2022

1 Introduction

Text normalization is a fundamental task in Vietnamese natural language processing. In this study, we focus on building a simple module for this task. We propose the method of using a set of rules to solve this problem. Then compare the effectiveness with two popular tools in Vietnamese. The results of the study are quite positive.

2 Vietnamese Text Normalization Task

Normalization tasks:

- Punctuation standardization (luý → lúy, cưú → cứu)

3 Methods

Method: Build 640 rules for mapping syllables

4 Results

For evaluation, we comparison our normalize words with two popular tools in Vietnamese viet_text_tools and VietnameseTextNormalizer.
Comparison with viet_text_tools

# differences	1326
# non miss spell	722
# miss spell	604

Comparison with VietnameseTextNormalizer
TO BE DONE

5 Discussion & Conclusions