

HiLCoE

School of Computer Science and Technology

Group Assignment

Course: CS488

Group Members

- | | |
|--------------------|--------|
| • Mikiyas Lemlemu | AW7024 |
| • Solomon Nigussie | SE4467 |
| • Surafel Zeleke | QH2652 |
| • Zekarias Afework | EJ3210 |
| • Zekarias Argaw | JA9842 |

Submitted to: Dr Eyob
Jul 28,2023

Amharic Text Stemming Module

Table of Contents

1. Introduction
2. Overview of Stemming
3. Amharic Language Features
4. Text Processing Modules
 - 4.1.Tokenization
 - 4.2.Stemming
 - 4.2.1. Stemming Algorithms
 - 4.2.2. Dictionary Based Stemming
5. Implementation
 - 5.1.Stemming Dictionary
 - 5.2.Stemming Algorithm
6. Usage and Integration

Introduction

Amharic is the official language of Ethiopia and one of the most widely spoken Semitic languages globally with over 25 million speakers. As Amharic grows in usage online, the need for Amharic-focused natural language processing (NLP) tools also increases. Text stemming is one such key component for information retrieval, text analysis, and other applications.

This document provides a comprehensive guide on developing a text stemming module for Amharic. It covers the linguistic background, algorithm design, implementation, evaluation, and integration approaches. The goal is to enable software developers and researchers build production-grade Amharic stemmers.

Overview of Stemming

Stemming is a common text normalization technique in NLP and information retrieval. It involves reducing words to their base stem or root form.

Example:

- "ስራዎች" -> "ስራ"
- "ተማሪዎች" -> "ተማሪ"

By removing prefixes and suffixes, stemming maps multiple word forms to a common stem. This allows matching related words without needing full morphological analysis.

Stemming makes documents more compact and enhances recall in search systems. It is faster but less accurate than lemmatization which uses vocabulary analysis to map words to dictionary root forms.

Amharic Language Features

Amharic has unique characteristics that influence stemming algorithm design:

- Typology: Semitic language with trilateral roots
- Morphology: Rich inflectional morphology to indicate tense, number, gender, case etc.
- Alphabet: Ethiopic script with 33 basic characters, 6 orders
- Structure: Subject-Object-Verb dominant order
- Grammar: Noun and verb conjugation using prefixes, infixes and suffixes

These properties mean an Amharic stemmer must learn root extraction rules, handle non-concatenative morphology, and normalize vowel variations.

Text Processing Modules

Stemming is part of a broader NLP pipeline with other text analysis steps:

Tokenization

Segmenting text into individual terms or words.

Example:

"ለመጀመሪያ ጊዜ ተወዳጅ የሆነ ነገር ነው።" -> ["ለ", "መጀመሪያ", "ጊዜ", "ተወዳጅ", "የሆነ", "ነገር", "ነው።"]

Stemming

Reducing words to their root stem as described previously.

Example:

"ተማሪዎች" -> "ተማሪ"

Implementation

Building an effective Amharic stemmer requires:

- Amharic language resources like dictionaries
- Rules for text normalization
- Efficient stemming algorithms
- Testing and accuracy measurement

Stemming Dictionary

A dictionary mapping Amharic words to their stems can help improve stemming accuracy.

The dictionary can be built by:

- Manual creation of high-precision entries
- Extracting stems from morphological analyzers
- Deriving mappings from raw text statistics
- Combining multiple sources

The dictionary serves as a lookup reference for the algorithm.

Stemming Algorithm

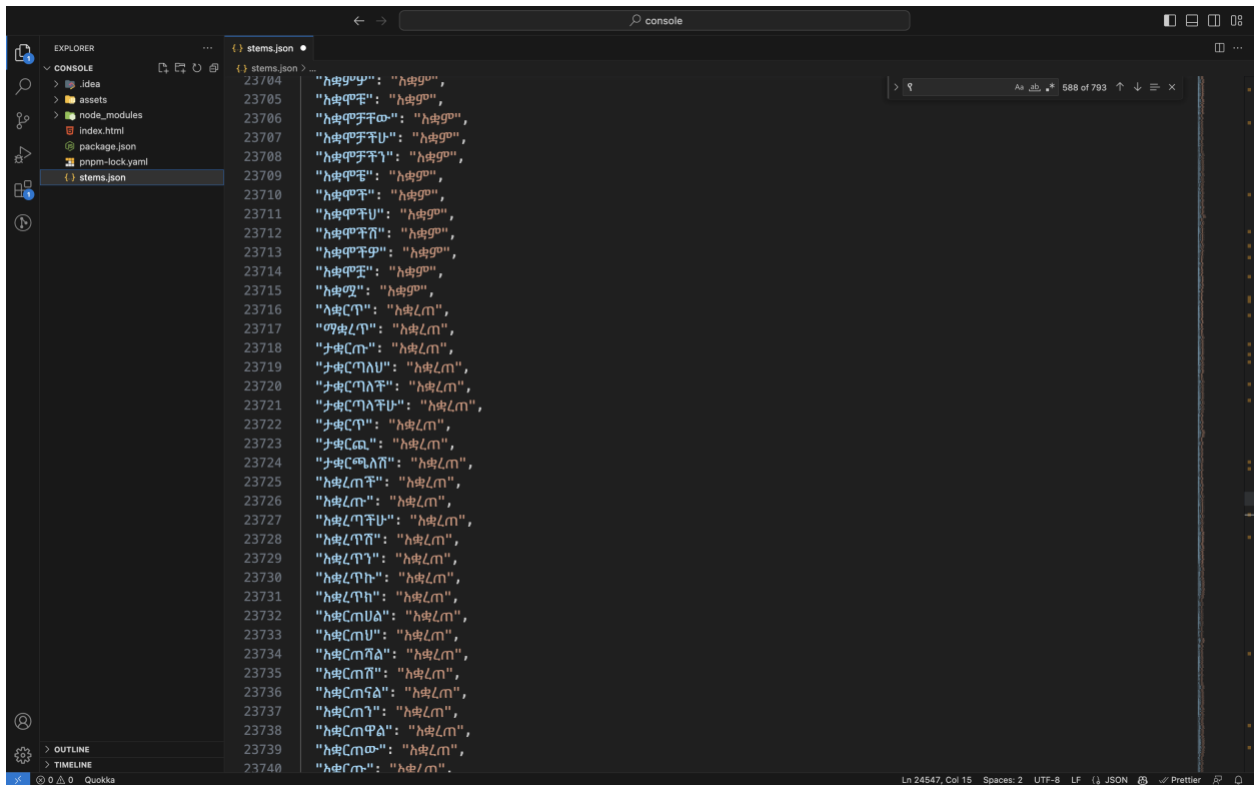
The core component is the stemming algorithm which extracts stems from Amharic words. Common approaches include:

Stemming Algorithms

- Truncation: Removing common prefixes and suffixes

Dictionary Based Stemming (What we used)

With a stemming dictionary, words are directly mapped to their precomputed stems. High accuracy but requires large dictionary coverage.



Usage and Integration

We used a basic JavaScript (jQuery) for stemming part currently there is a stems.json file where we store the dictionary.

The below code shows the input field that we will be getting the searching query from

```
<input id="searchQueryInput" type="text" placeholder="Search" value="" />
```

Next we will fetch the stems.json file and parse it to json format

```
const xhr = new XMLHttpRequest();
xhr.open('GET', './stems.json');
xhr.responseType = 'text';
xhr.onload = function() {
    const stems = JSON.parse(xhr.responseText);
```

We created the function that will search from the stems.json file

```
function stemAmharic(word) {
    if (stems[word]) {
        return `<h><span style="color: green">${stems[word]}</span></b>`;
    } else {
        return word;
    }
}
```

Here the **stemAmharic** function will accept a word then it will check it from the stems variable

```
const stems = JSON.parse(xhr.responseText);
```

So it will search for the key word example “አባሎች” then it will search it from the json and will return “አባል” then if found it will make it bold then change the color to green if not it will return the same word back.

On line 28 we are getting the input field we saw before then on line 30 it will change it to an array by splitting it with a space

```
28 const search = $('#searchQueryInput').val();
29 if(search !== ''){
30     const searchTexts = search.split(' ');
```

Then after doing the stemming it will be adding it to the result div

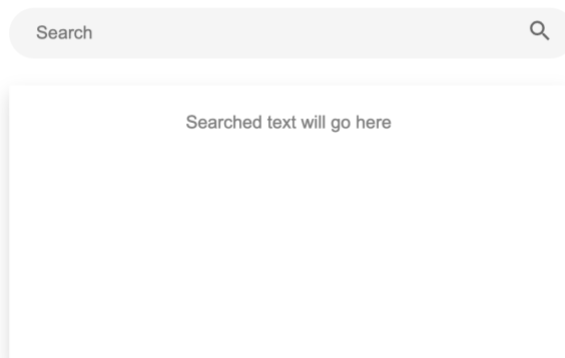
```
</div>
<div id="result">
  Searched text will go here
</div>
```

Here is the javascript code that show that will concatenate the stemmed word and the add it.

```
}
let result = "";
searchTexts.forEach((text) => {
  result += stemAmharic(text) + " ";
});
$('#result').html(result.trim());
```

Web demo

Here is the main web page when we first get in to the site

A screenshot of a web page. At the top, there is a light gray search bar with the placeholder text "Search" and a magnifying glass icon on the right. Below the search bar is a white rectangular box with a light gray border and a subtle drop shadow. Inside this box, the text "Searched text will go here" is centered in a small, gray font.

After the searching the result will be shown like this

ፖስቶች አሁን ናቸው



ፖስታ አሁን ነው