Web Crawling Untuk Pembangunan Korpus Bahasa-Bahasa Daerah Indonesia = Building Corpora for Indonesian Regional Languages by Web Crawling

Alif Iqbal Hazairin, author

Abstrak

Bahasa daerah adalah bahasa yang digunakan sebagai penghubung pada masyarakat suatu daerah atau suatu kelompok masyarakat tertentu di samping bahasa utama, yaitu bahasa Indonesia. Keragaman bahasa daerah di Indonesia merupakan kekayaan budaya yang harus dipertahankan sepanjang zaman. Sayangnya, penggunaan bahasa daerah yang berkurang serta minimnya perhatian masyarakat pada digitalisasi bahasa daerah membuat bahasa daerah semakin terpinggirkan. Tak terkecuali pada bidang NLP, belum ada perkembangan signifikan dalam puluhan tahun terakhir yang melibatkan bahasa daerah sebagai subjek penelitian. Oleh karena itu, penelitian ini mencoba memberikan salah satu cara untuk meningkatkan kembali pelibatan bahasa daerah dalam penelitian khususnya NLP. Penelitian ini mencoba membangun korpus teks untuk sebanyak mungkin bahasa daerah di Indonesia menggunakan metode web crawling. Sistem melakukan crawling untuk mengumpulkan web berbahasa daerah sebanyak-banyaknya dan kontennya diambil dengan melakukan web scraping. Teks hasil scraping selanjutnya dinormalisasikan dan dilakukan language identification pada tiap kalimatnya. Kalimat dengan bahasa mayor seperti Indonesia dan Inggris dibuang, dan kalimat yang berbahasa daerah dipertahankan. Hasilnya adalah korpus teks untuk ratusan bahasa daerah di Indonesia. Harapannya hasil penelitian ini dapat menjadi batu loncatan penelitian bahasa daerah NLP di Indonesia selanjutnya.

.....Regional languages are languages used as a means of communication within a specific region or community, in addition to the main language, which is Indonesian. The diversity of regional languages in Indonesia is a cultural wealth that should be preserved throughout time. Unfortunately, the diminishing use of regional languages and the lack of attention given by society to the digitization of these languages have led to their marginalization. This holds true even in the field of Natural Language Processing (NLP), where there has been little significant development involving regional languages as research subjects in recent decades. Therefore, this study aims to provide a method to re-engage regional languages, particularly in NLP research. The research attempts to build a text corpus for as many regional languages in Indonesia as possible using web crawling methods. The system will crawl the web to collect regional language websites and extract their content through web scraping. The scraped texts will then undergo a normalization process and language identification process for each sentence. Sentences in major languages such as Indonesian and English will be discarded, while sentences in regional languages will be retained. The outcome of this research will be a text corpus for hundreds of regional languages in Indonesia. The hope is that the results of this study can serve as a stepping stone for the next NLP research on regional languages in Indonesia.