INTISARI

Implementasi Teknik Web Scraping Pada Proses Topic Modelling Portal Berita

Oleh

Pramudya Rian Dewantoro
08/265296/PA/11863

Seiring berkembang pesatnya teknologi, menyebabkan penyebaran informasi yang begitu cepat. Perkembangan teknologi ini mendorong manusia untuk bisa mendapatkan informasi lebih cepat khususnya berita. Berita menampilkan berbagai macam topik. Dengan pesatnya arus berita, diperlukan metode yang lebih cepat dan efisien untuk mendapatkan topik.

Skripsi ini dibuat untuk mengetahui topik-topik pada portal berita. Metode yang digunakan yaitu dengan mengimplementasikan *web scraping* pada proses t*opic modeling.* Teknik scraping mengambil artikel dari portal berita dan kemudian diolah menggunakan topic modeling. Topic modeling yang digunakan adalah Latent Dirichlet Allocation (LDA). LDA adalah model probabilitas generative dari koleksi data diskrit seperti kumpulan-kumpulan teks. Pada proses scraping, perancangan sistem dilakukan dengan identifikasi kelas tag HTML. Tag HTML yang digunakan yaitu tag yang mengapit judul, isi dan tanggal berita untuk kemudian dibuatkan template *scraping*. Data yang diperoleh kemudian diolah dengan pemodelan LDA, sehingga dapat diketahui topik-topik yang sering muncul dari portal berita nasional.

Sistem ini dibuat dengan menggunakan bahasa pemrograman Python 2.7.11 dengan modul-modul pendukungnya. Sistem ini bisa memproses *web scraping* dari portal berita Kompas dan kemudian disimpan ke file CSV. Berita Kompas diambil dalam rentang 7 hari. Dokumen CSV yang diperoleh kemudian diolah menggunakan modul python LDA untuk diperoleh *topic modelling*. Hasil keluaran berupa topik pada portal berita yang paling sering muncul.

Kata Kunci : Web Scraping, Topic Modeling, LDA, Text Processing

ABSTRACT

Web Scraping Implementation For News Portal Topic Modelling Process

By

Pramudya Rian Dewantoro
08/265296/PA/11863

Information technology is currently growing fast. Human need to develop technology to obtain information, especially news. Online news is spread faster than conventional news. News portal is one of the media for human to reach any information. Each news portal contain any topics that was helped out to know news topics. For news that spread faster with the various topics, required a method to gain the specific topic.

The method which is implemented in this study is the Web Scraping on topic modeling process. This sort of technique grabs the article from the news portal then processed using topic modeling. Latent Dirichlet Allocation (LDA) is a method that use for topic modeling process. LDA a generative probabilistic model for collections of discrete data such as text corpora. The system design is conducted by identifying the tag HTML class. The HTML Tags that used are the tags that are included within the news title, news content and the news date. The grabbed tags then filled into the scraping template to get its data to be collected. The Data that is obtained then will be processed by the LDA modeling, hence the user will be acknowledged about the topics that frequently appeared in the national news portal.

This system is built using the Phyton Programming version 2.711 by utilizing the supporting libraries. This system can process the web scraping from Kompas news portal and store it into CSV files. Kompas news article is taken in 7 days span. The obtained CSV document then will be processed using the LDA Phyton Libraries to get the modeling topic. The result will be in the most frequent topic form on the news portal

Keyword : Web Scraping, Topic Modeling, LDA, Text Processing