

DETEKSI TRENDING TOPICS DARI DATA STREAM TWITTER

DENGAN PENDEKATAN SEQUENTIAL K-MEANS

ABSTRAK

Trending topics twitter adalah kumpulan topik tertentu yang banyak dibicarakan dalam *tweet* pengguna. Menemukan *trending topics* bukanlah hal yang mudah karena data yang digunakan sangat besar dan terus mengalir sehingga dibutuhkan media penyimpanan yang besar, pengolahan data yang cukup lama, dan algoritma yang efisien untuk analisis data. Penelitian ini bertujuan untuk merancang sebuah model dan strategi dalam menemukan *trending topics* dari data *stream* di *Twitter*. Pengolahan data secara *streaming* adalah proses dimana data *real-time* diambil dalam rentang waktu tertentu kemudian digunakan untuk mendapatkan sebuah model setelah itu dihapus dan menyediakan ruang untuk data baru. Pendekatan penelitian dilakukan dalam empat tahap, pertama proses pengumpulan data *twitter*. Kedua, praproses data yang terdiri dari praproses teks, pembobotan kata dan seleksi kata. Ketiga, proses analisis data dengan teknik *clustering*, algoritma yang digunakan yaitu *k-means* dan *sequential k-means*. Keempat, pengolahan informasi yang terdiri dari evaluasi *cluster*, deteksi topik, visualisasi *trending topics* dan evaluasi *trending topics*. *Sequential k-means* digunakan karena dapat menerima data masukan secara sekuensial dan *cluster center* dapat di-update. Pengujian terhadap model dilakukan dalam tiga skenario dimana setiap skenario dibedakan antara jumlah data, waktu dan nilai parameter. Setelah itu, evaluasi terhadap hasil *clustering* akan dilakukan dengan menggunakan metode *Dunn Index*. Aplikasi *trending topics twitter* akan dibuat menggunakan bahasa R dan menghasilkan keluaran dalam bentuk histogram.

Kata kunci : *trending topics, data stream, sequential k-means, R*

DETECTION TRENDING TOPICS FROM STREAMING DATA TWITTER USING SEQUENTIAL K-MEANS

ABSTRACT

Trending topics twitter is a collection of specific topics are the most discussed in the tweets. Detection trending topics is not easy because using a streaming big data, so we need a large storage and efficient algorithms for data analysis. This research aims to design a model and strategies to detection trending topics from streaming data Twitter. Processing of streaming data is the process which real-time data taken within a certain time range and then used to obtain a model, after that is data removed and provide a space for a new data. The approach in this research is carried out in four phases, first is data twitter collection. Second, data preprocessing which consisting of text preprocessing, word weighting and word selection. Third, data analysis with clustering techniques using k-means and sequential k-means. Fourth, information processing consisting of cluster evaluation, topic detection, visualization and evaluation of trending topics. Sequential k-means is used because it can receive data sequentially and cluster centers can be updated. Testing of the model is done in three scenarios where each scenario distinguished between the amount of data, time of collection and the value of the parameter. After that, clustering evaluation will be done using Dunn Index. Applications trending topics twitter will be created using the R language and produce an output in the form of a histogram.

Keywords : *trending topics, streaming data, sequential k-means, R*