Nama : Myta Ayu Pratiwi

Nim :15.01.55.0005

**CLUSTERING**

Review Jurnal 1

|  |  |
| --- | --- |
| Judul | Analisis Potensi Daerah Melalui Metode *Document Clustering* Laporan Pelaksanaan Kegiatan  Kuliah Kerja Nyata-Pembelajaran Pemberdayaan Masyarakat |
| Jurnal | Jurnal Edukasi dan Penelitian Informatika (JEPIN) |
| Volume & Hal | Vol. 1 & Hlmn 1-8 |
| Tahun | 2015 |
| Penulis | Dyah Mustikasari, Teguh Bharata Adji, Abdul Kadir |
| Tujuan Peneitian | menambang informasi tentang potensi daerah dari dokumen laporan pelaksanaan kegiatan KKN-PPM menggunakan salah satu metode pada *text mining*, yaitu *document clustering*. |
| Metode Penelitian | Penelitian ini menggunakan metode *clustering* untuk mengelompokkan dokumen dengan dua algoritma, STC dan LINGO. Keduanya dipilih karena algoritma tersebut berbasis frasa. Algortima ini khusus untuk bidang *text mining*.Langkah Penelitian:  Langkah pertama adalah persiapan dokumen, kedua adalah *preprocessing*, ketiga *clustering*, dan yang terakhir tahap evaluasi |
| Objek Penelitian | Berkas (*file*) dokumen yang digunakan dalam penelitian ini berjumlah 546 berkas(Laporan pelaksanaan kegiatan (LPK) KKN-PPM). |
| Hasil Penelitian | algoritma LINGO lebih baik dalam memberikan gambaran tentang potensi daerah dibandingkan algoritma STC. LINGO memunculkan label klaster yang bertema potensi daerah lebih banyak dibanding STC. Dari evaluasi pada penelitian ini, LINGO menghasilkan nilai *F-Measure* 70%, dua kali lebih tinggi daripada STC yang hanya 33%. |
| Kelebihan Penelitian | *Clustering* dilakukan dengan dua pendekatan yaitu, STC dan LINGO, menggunakan Carrot2 Workbench. Penggunaan dua algoritma ini dimaksudkan untuk memperoleh perbandingan algoritma yang memberikan hasil lebih baik dalam penggambaran potensi daerah lokasi KKN-PPM UGM. |
| Kekurangan Penelitian | algoritma LINGO adalah masih menghasilkan klaster "Other Topics". Oleh karena itu, perlu dilakukan penelitian lebih lanjut untuk mengurangi atau menghilangkan klaster “Others Topics” ini. |

Review Jurnal 2

|  |  |
| --- | --- |
| Judul | IMPLEMENTASI TEXT MINING PADA MESIN PENCARIAN TWITTER  UNTUKMENGANALISISTOPIK – TOPIK TERKAIT  “KPK dan JOKOWI” |
| Jurnal | Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS |
| Volume & Hal | Hlmn 1-12 |
| Tahun | 2015 |
| Penulis | Imam Adiyana, Dr. RB. Fajriya Hakim, S.Si., M.Si. |
| Tujuan Peneitian | menganalisis kumpulan data *tweet* dalam *search engine twitter* terkait topik berita KPK dan Jokowi serta meng*cluster*kan kata-kata yang terkait dengan KPK dan Jokowi dalam suatu *cluster* tertentu menggunakan metode *k-means*. |
| Metode Penelitian | **Populasi**  Populasi dari penelitian ini adalah berita/dokumen (teks) pada sosial media *twitter* yang termasuk dalam kategori *microblogging* tentang “kpk” dan “jokowi”. *Tweet-tweet*  pengguna pada *twitter* dapat dijadikan sebagai topik-topik pada media *online*.  **Sampel dan Teknik Pengambilan Sampel**  Sampel yang diambil adalah data teks dari twitter yang diambil dengan permintaan kepada sistem sebanyak 500 *tweet* pada periode waktu tertentu.  **Sumber Data dan Metode Pengumpulan Data**  Data yang digunakan adalah data primer yang dikumpulkan dari media sosial *twitter*. Data diambil dengan cara men*download* dan mengumpulkan data *tweet* mengenai KPK dan Jokowi dari hasil pencarian *search engine twitter* (mesin pencarian *twitter*) dan  menyimpannya menggunakan suatu program dengan *software* “R”.  **Tahapan Analisis Data**  *Data Readings*(Pembacaan data)-> Ekstraksi Fitur (*Tokenizing, Filtering,Stemming, Tagging*)->PembobotanTerm *->*Analisis.  **Pengolahan Data** |
| Objek Penelitian | berita/dokumen (teks) pada sosial media *twitter* tentang “kpk” dan “jokowi”. |
| Hasil Penelitian | hasil *cluster tweet* untuk masing-masing topic (KPK dan Jokowi), dimana jika diperhatikan terdapat kesamaan *cluster tweet* untuk topic KPK dengan topik Jokowi yaitu *cluster* 8 untuk topik KPK dan *cluster* 6 untuk topik Jokowi. Hal tersebut menunjukkan bahwa baik untuk topik KPK maupun Jokowi memiliki hubungan  yang dapat dilihat dari topik utama (ditunjukan dengan kata pertama) pada masing-masing  *cluster tweet*. |
| Kelebihan Penelitian | Pada twitter pengguna dapat menemukan berbagai macam informasi, bahkan ketika pengguna tidak mengetahui akun dari topik yang dicari pengguna dapat menggunakan bantuan mesin pencarian twitter untuk mencarikan  informasi terkait topik yang dicari. Banyaknya kegiatan yang dapat dilakukan menggunakan twitter, maka tidak dipungkiri twitter akan menghasilkan kumpulan data yang besar. |
| Kekurangan Penelitian | Penelitian hanya mengambil kesimpulan data tweet terkait search engine dan itu sebenarnya kurang luas,karen peringkat pertama adalah google |

Review Jurnal 3

|  |  |
| --- | --- |
| Judul | Klastering Dokumen Berita dari Web menggunakan Algoritma Single Pass Clustering |
| Jurnal | Jurnal Teknologi Informasi DINAMIK |
| Volume & Hal | Volume 18 &Hlmn 1-11 |
| Tahun | 2013 |
| Penulis | Herny Februariyanti dan Eri Zuliarso |
| Tujuan Peneitian | memperoleh informasi yang seimbang seperti apa yang diinginkan adalah dengan membaca beberapa dokumen yang membahas topik yang sama,dengan Sistem yang tepat untuk masalah tersebut adalah Sistem Temu Kembali Informasi yang dapat menghasilkan integrasi dari beberapa dokumen elektronik yang berbeda dengan topik bahasan yang sama secara otomatis. |
| Metode Penelitian | 1. Obyek Penelitian  Obyek penelitian dari penelitian ini adalah dokumen teks berupa halaman web http://www.kompas.com.  2.. Data yang diperlukan  Merupakan data yang mendukung dalam penelitian ini meliputi data primer dan data sekunder.  3. Teknik Pengumpulan Data  Pengumpulan data mempunyai tujuan mendapatkan materi – materi yang mempunyai keterkaitan dengan topik penelitian. |
| Objek Penelitian | Dokumen Berita dari Web |
| Hasil Penelitian | menunjukkan bahwa algoritma ini dapat diaplikasikan untuk pengelompokan dokumen berbahasa Indonesia. Pemilihan kata kunci yang tepat akan meningkatkan kualitas temu kembali informasi (information retrieval) pada dokumen. |
| Kelebihan Penelitian | Sistem dapat menampilkan dokumen yang mempunyai kedekatan similaritas dari query yang diinputkan user.  Klaster dapat membantu menemukan dokumen yang ada dalam satu klaster dengan query yang diinputkan user. Klaster dapat membantu mendapatkan dokumen yang relevan. |
| Kekurangan Penelitian | - |

Review Jurnal 4

|  |  |
| --- | --- |
| Judul | ANALISIS KLASTERING LIRIK LAGU INDONESIA |
| Jurnal | Prosiding SINTAK |
| Volume & Hal | Hlm 1-7 |
| Tahun | 2017 |
| Penulis | Afdilah Marjuki, Herny Februariyanti |
| Tujuan Peneitian | pengkategorian lirik lagu Indonesia dengan text mining. |
| Metode Penelitian | 1. K-Means Clustering 2. Hierarchical Clustering 3. Agglomerative Hierarchical Clustering 4. Pemrograman R |
| Objek Penelitian | lirik lagu indonesia. |
| Hasil Penelitian | Lirik lagu yang memiliki tingkat similaritas yang sama akan berada dalam satu klaster. |
| Kelebihan Penelitian | Sistem dapat menampilkan lirik lagu berbahasa Indonesia hasil klastering dengan *hierarchical clustering* dan  *K-Means Clustering*. Sistem yang dibuat berhasil menerapkan *data cleaning* pada teks lirik lagu berbahasa Indonesia. |
| Kekurangan Penelitian | Sistem belum bisa menerapkan stemming nazief-adriani dalam mencari kata dasar teks berbahasa Indonesia  dengan bahasa pemrograman R. |

Review Jurnal 5

|  |  |
| --- | --- |
| Judul | PENGELOMPOKAN TEKS MENGGUNAKAN ALGORITMA CANOPY CLUSTERING |
| Jurnal |  |
| Volume & Hal | Hlmn 1-7 |
| Tahun | 2011 |
| Penulis | Agha Dwi Nugraha¹, Angelina Prima Kurniati², Intan Nurma Yulita³ |
| Tujuan Peneitian | Mengimplementasikan algoritma Canopy Clustering dalam  mengelompokkan dokumen dan Mengetahui dan menganalisis waktu proses peng-cluster menggunakan  Algoritma K-means setelah menggunakan metode Canopy clutering. |
| Metode Penelitian | 1. Studi literatur  Pencarian referensi dan sumber-sumber yang berhubungan dengan teks *clustering*.  2. Tahap pengumpulan data dan *preprocessing* data  Pada tahap ini akan dibangun model berkaitan dengan requirement-requirementyang dibutuhkan pada saat implementasi, mulai dari analisis kebutuhan, desain database, desain aplikasi/*interface.*  3. Tahap perancangan sistem  Pada tahap ini akan dibangun model berkaitan dengan requirementrequirement yang dibutuhkan pada saat implementasi, mulai dari analisis kebutuhan, desain database, desain aplikasi/*interface.*  4. Tahap Implementasi  Pada tahap ini, akan dilakukan implementasi sistem yang mampu  mengklasifikasikan dokumen teks secara otomatis dengan menggunakan metode *K-means dan Canopy Clustering*.  5. Tahap Pengujian Sistem  Pada tahap ini akan dilakukan pengujian program yang telah diimplementasikan dengan melakukan memasukkan dokumen teks lalu melihat hasil klasifikasi.  6. Tahap Analisis Hasil Pengujian  Setelah sistem telah sempurna maka akan dianalisis hasil dari klasifikasi dan kestabilan dari sistem.  7. Tahap Pembuatan Laporan  Pada tahap ini, akan dilakukan penyusunan laporan akhir dan pengumpulan dokumentasi dengan mengikuti kaidah penulisan yang benar dan sesuai dengan ketentuan-ketentuan atau sistematika yang telah ditetapkan oleh institusi. |
| Objek Penelitian | dokumen berita berbahasa Indonesia yang telah mengalami text preprocessing |
| Hasil Penelitian | Nilai parameter T pada *Canopy K-means* berpengaruh  terhadap nilai akurasi dan waktu, pada saat nilai parameter T kecil, akan menghasilkan waktu yang besar dan akurasi yang buruk dibandingkan dengan *K-means* . Untuk nilai parameter T yang besar dihasilkan akurasi yang lebih baik dan waktu yang lebih cepat dibandingkan dengan *K-means.* Akan teteapi apabila nilai T terlalu besar maka akan dihasilkan waktu dan akurasi yang tidak berbeda jauh dengan *K-means.* Maka dibutuhkan nilai parameter T yang tepat untuk medapat kan akurasi dan waktu proses yang lebih baik dan Nilai k pada *Canopy K-means* tidak terlihat perbedaan waktu  yang signifikan dengan *K-means*. |
| Kelebihan Penelitian | Secara keseluruhan *Canopy K-means* mempunyai keunggulan  dalam waktu proses. |
| Kekurangan Penelitian | harus Mencari perhitungan parameter T yang tepat. Hasil yang dihasilkan pun akan lebih baik |

Review Jurnal 6

|  |  |
| --- | --- |
| Judul | CLUSTERING ARTIKEL BERITA BERBAHASA INDONESIA  MENGGUNAKAN UNSUPERVISED FEATURE SELECTION |
| Jurnal | Seminar Nasional Informatika |
| Volume & Hal | Hlmn 1-10 |
| Tahun | 2010 |
| Penulis | Diah Pudi Langgeni1, ZK. Abdurahman Baizal 2 , Yanuar Firdaus A.W.3 |
| Tujuan Peneitian | mengimplementasikan unsupervised feature selection yaitu Document  Frequency (DF) dan Term Contribution (TC) pada clustering berita berbahasa Indonesia |
| Metode Penelitian | Tahap pengumpulan data dan *preprocessing* data  Tahap Implementasi  Tahap Pengujian Sistem  Tahap Analisis Hasil Pengujian |
| Objek Penelitian | mengelompokkan dokumen yang memiliki kesamaan isi.  Clustering dapat diaplikasikan untuk menemukan keterkaitan antar berita |
| Hasil Penelitian | Term Contribution lebih baik daripada Document Frequency yaitu dapat menghasilkan nilai precision dan entropy lebih baik dengan fitur yang lebih sedikit. Hal ini dikarenakan Term Contribution mempertimbangkan frekuensi kemunculan term dan frekuensi dokumen sebuah term, sehingga term yang tetap dipertahankan adalah term yang khas atau bersifat diskriminator, berbeda halnya dengan Document  Frequency yang hanya mempertahankan term – term yang bersifat umum |
| Kelebihan Penelitian | Pada Term Contribution, term yang dianggap baik adalah term yang memiliki nilai term frequency yang besar dan terjadi di sedikit dokumen. |
| Kekurangan Penelitian | Pada Document Frequency, term yang dianggap baik adalah term yang terjadi di banyak dokumen. Padahal yang lebih baik Term Contribution |

Review Jurnal 7

|  |  |
| --- | --- |
| Judul | Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia |
| Jurnal | JURNAL SAINS DAN SENI ITS |
| Volume & Hal | Vol. 6 &Hlmn 1-6 |
| Tahun | 2017 |
| Penulis | Dwi Smaradahana Indra loka, dan Budi Santosa |
| Tujuan Peneitian | penerapan Text Mining untuk melakukan clustering dengan metode K-means pada data tweet Shopee Indonesia untuk mengetahui jenis konten tweet yang banyak dilakukan retweet oleh followers Shopee Indonesia |
| Metode Penelitian | 1. Authentication 2. Pengambilan Data dari Twitter 3. Case Folding 4. Tokenizing 5. Stopword Removal 6. Text Representation 7. Penentuan Jumah Klaster Terbaik 8. Clustering dengan K-means 9. Penentuan Jenis Konten Tweet tiap Klaster 10. Perhitungan Rata-Rata Jumlah Retweet tiap Klaster |
| Objek Penelitian | data tweet Shopee |
| Hasil Penelitian | Penerapan algoritma proses *Text Mining* untuk melakukan *clustering* dengan metode *K-means* pada data *tweet* Shopee Indonesia menghasilkan sejumlah 28 klaster *tweet*.Berdasarkan hasil proses penentuan jenis konten dan perhitungan rata-rata jumlah *retweet* pada tiap klaster, didapatkan bahwa jenis konten pada klaster yang memiliki jumlah *retweet* yang tinggi diantaranya  tentang kuis berhadiah (klaster 4, klaster 20, dan klaster 28), ulang tahun Shopee Indonesia (klaster 11), dan hobi, kuis, dan *lifestyle* (klaster 7). |
| Kelebihan Penelitian | Pelaku bisnis Shopee Indonesia dapat mengetahui jenis konten *tweet* yang banyak dilakukan *retweet* oleh *followers*-nya sehingga dapat menggunakan jenis konten *tweet* tersebut sebagai sarana untuk melakukan *advertising* kepada pengguna Twitter. |
| Kekurangan Penelitian | Penelitian ini hanya berpaku pada banyaknya retweet dan kurang luas,semoga penelitian ini juga mampu menghasilkan analisa kemungkinan yang lain terjadi. |