E-ISSN: 2715-2731

Implementasi *Term-Frequency Inverse Document Frequency (TF-IDF)* Untuk Mencari Relevansi Dokumen Berdasarkan *Query*

Moh. Afif Rofiqi,[1,*], Abd Charis Fauzan[2], Afivatu Pratama Agustin[3], Ahmad Agung Saputra[4], Hinayu Diniatul Fahma[5]

Program Studi Ilmu Komputer, Universitas Nahdlatul Ulama Blitar, Indonesia
[1]afifrofiqi9@gmail.com; [2]abdcharis@unublitar.ac.id; [3]afiv.pa@gmail.com; [4]ahmadagungsaputra27@gmail.com; [5]dinia.fahma@gmail.com
* corresponding author

ARTIKEL INFO

Article history
Diterima: 23 Oktober 2019
Direvisi: 15 November 2019
Diterbitkan: 30 Desember 2019

Keywords
Term-Frequency Inverse Document Frequency, Relevansi, Dokumen, *Query*

ABSTRAK

Tujuan dibuatnya penelitian ini adalah untuk mencari relevansi antar beberapa dokumen berupa artikel berita dari beberapa sumber. Metode yang digunakan yaitu metode *Term-Frequency Inverse Document Frequency* karena relevan untuk keakuratan sebuah dokumen. *Term-Frequency Inverse Document Frequency* adalah perhitungan atau pembobotan kata melalui teknik tokenisasi, stopwords, dan stemming, dan frekuensi munculnya kata dalam dokumen yang diberikan menunjukkan pentingnya kata itu di dalam sebuah dokumen. Yang mengunakan data dari artikel berita metode ini melakukan pembobotan kata didalam sebuah dokumen dengan mengalikan nilai TF dan IDF bedasarkan hasil querynya dan dari tiga artikel yang mengasilakan rank score untuk dokumen satu yang berscore 3,90847 dapat disimpulkan bahwa artikel berita pada dokumen satu adalah yang paling relevan dari pada dua artikel lainnya.

PENDAHULUAN

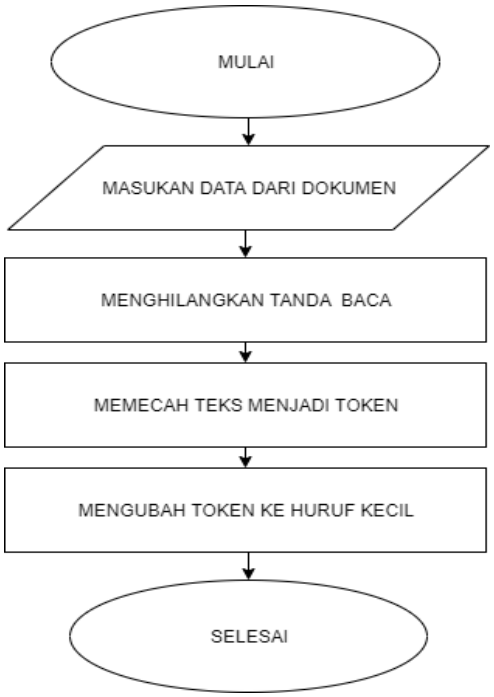
Pentingnyan mencari dokumen saat ini sangat di butuhkan apalagi yang berjumlah banyak maka dari itu dengan menggunakan *metode Term-Frequency Inverse Document Frequency* (*TF-IDF)*, bisa dengan mudah mencari data tersebut dengan tingkat relevan yang tinggi sehingga tidak salah dalam mengambil data dokumen yang di perlukan [1]. Metode yang digunakan mencari relevansi antar beberapa dokumen disini adalah metode *TF-IDF* karena merupakan metode yang cukup mudah dipelajari dan juga mudah diterapkan untuk permasalahan keakuratan sebuah dokumen. *TF-IDF* adalah perhitungan atau pembobotan kata melalui teknik tokenisasi, *stopwords*, dan *stemming*, dan frekuensi munculnya kata dalam dokumen yang diberikan menunjukkan pentingnya kata itu di dalam sebuah dokumen [2]. Metode ini melakukan pembobotan kata didalam sebuah dokumen dengan mengalikan nilai TF dan IDF bedasarkan hasil *query*-nya [3]. Dokumen yang digunakan untuk penelitian ini yaitu menggunakan dokumen dari beberapa artikel berita karena kebayakan berita yang ada tidak sesuai fakta yang ada dan untuk tingkat akurasinya juga cukup di pertayaakan maka dari itu penelitian ini mengambil data dari artikel berita agar mengetahui berita mana yang relevan dengan keadaan saat ini sebagai contoh untuk berita tentang berpindahnya ibu kota Jakarta ke Kalimantan. Pada penelitian sebelumnya metode ini di gunakan dalam pencarian buku dalam sebuah perpustakaan *online* yang dimana *query*nya bedasarkan judul buku. Penelitian terdahulu adalah penelitian untuk Mengukur Akurasi Permintaan pada Repositori menggunakan *User Interface* [4], Penelitian Mengukur Akurasi pada *Query Interface* dengan *Software As A Sevice* (SAAS) [5]. Andayani Implementasi Algoritma TF-IDF pada pengukuran kesamaan dokumen[2].

Kemudian penelitian ini dilakukan untuk mencari tingkat relevansi pada artikel dengan topik pemindahan ibu kota yang menyebar di dunia maya.

METODE

Tokenisasi

Tokenisasi adalah proses untuk membagi atau memecah teks yang dapat berupa kalimat, paragraf atau dokumen, menjadi token atau bagian tertentu seperti kumpulan kata dengan cara menghilangkan tanda baca atau mengubah huruf capital menjadi huruf kecil (*lower case*). Sebagai contoh, tokenisasi dari kalimat "Aku akan berkunjung ke rumahmu. Apa kau ada di rumah?" setelah ditokenisasi menjadi aku-akan-pergi-ke-rumahmu-apa-kau-ada-di-rumah [2][7].

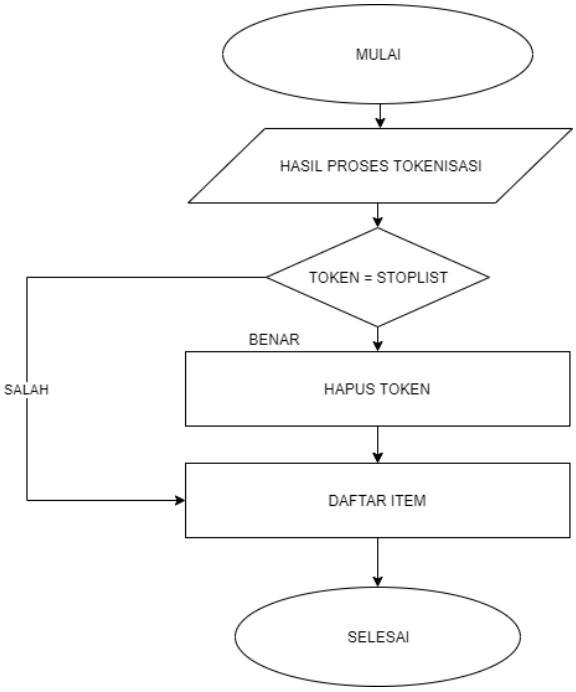


Gambar 1. *Flowcart Tokenisasi*

Gambar 1 adalah *flowcart* dari metode tokenisasi, dimana dokumen dalam database di hilangkan tanda bacanya dan mengubah token menjadi huruf kecil. Pentingnya metode ini karena menjadi awalan dari penerapan metode yang lain.

Stopwords

Stopwords umumnya dimanfaatkan dalam task information retrieval, termasuk oleh *Google*. Contoh *stopwords* untuk bahasa Inggris diantaranya “of”, “the”. Sedangkan untuk bahasa Indonesia diantaranya “yang”, “di”, “ke”. Dalam dunia pemrogramman khusunya di proses klasifikasi data, *stopword* sangat diperlukan yaitu digunakan *stopwords* untuk mengurangi jumlah kata yang harus diproses. Sangat berguna untuk proses *Text Mining*[5] [6]. Gambar 2 menggambarkan proses yang dilakukan saat tahap stopword removal atau filtering, dimana hasil dari proses tokenizing yang dilakukan sebelumnya, akan dicocokan dengan array stoplist yang ada, apabila token yang dicek merupakan stoplist maka token akan dihapus, apabila token bukan termasuk *stoplist* maka token akan dibiarkan tetap ada.



Gambar 2. *Flowcart Stopwords*

Gambar 2 adalah tahapan dari metode *stopword*, yang mana hasil dari metode tokenisasi yang dilakukan sebelumnya, akan dilakukan pencocokan dengan *array stoplist* , apabila token yang dicek termasuk *stoplist* maka token dihapus, jika token tidak termasuk *stoplist* maka token dibiarkan ada.

Stemming

Stemming digunakan untuk memeriksa relevannya beberapa dokumen dibutuhkan proses *stemming*. *Stemming* adalah di mana proses pemetaan kata pada kalimat berimbuhan menjadi kata asli (tanpa kata imbuhan awalan,akhiran,sisipan,kombinasi) yang di jalankan algoritma tertentu.

Contoh :

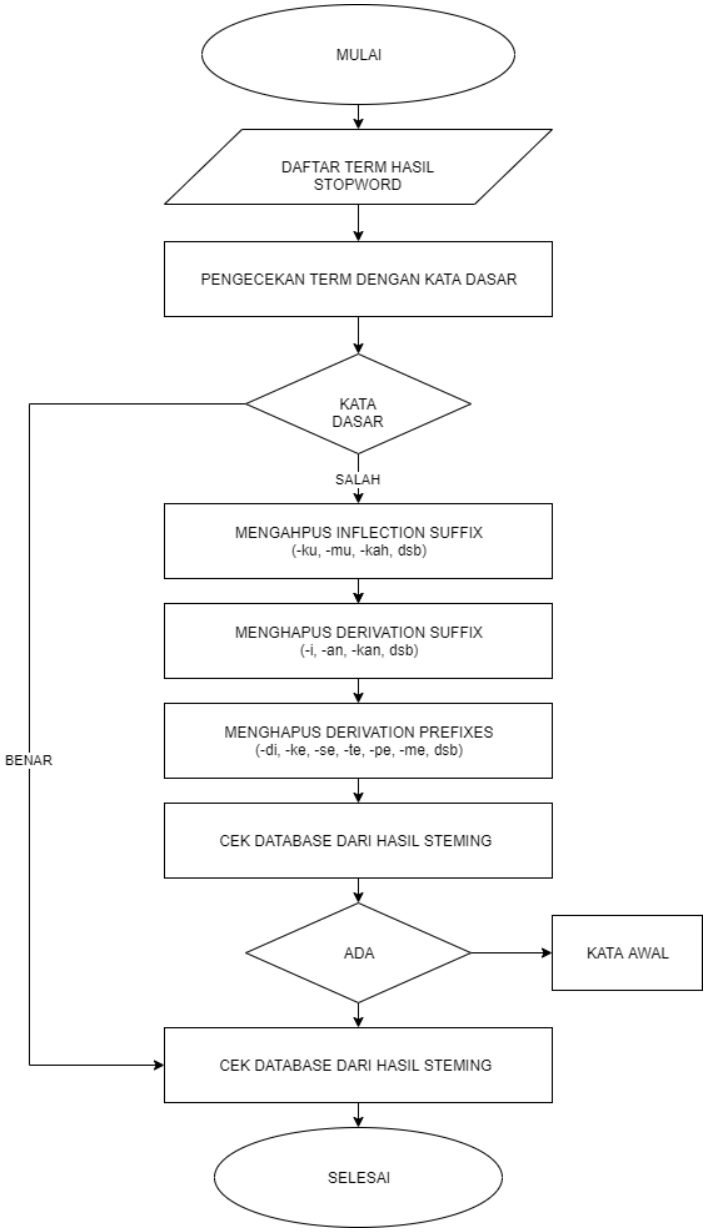
Membetulkan = Betul

Membenarkan = Benar

Ada pula imbuhan pada bahasa Indonesia cukup kompleks, terdiri dari:

- *Prefiks*, imbuhan di depan kata: ber-tiga
- *Suffiks*, imbuhan di akhir kata: makan-an
- *Konfiks*, imbuhan di depan dan di akhir kata: per-ubah-an
- *Infiks*, imbuhan di tengah kata: kemilau.
- Imbuhan dari bahasa asing: *final*-isasi, sosial-isasi
- Aturan perubahan *prefiks*, seperti (me-) menjadi (meng-, mem-, men-, meny-)

Gambar 3 merupakan proses saat tahap *stemming*, yaitu *term* hasil dari proses sebelumnya akan dilakukan pengecekan terhadap database. Apabila t*erm* merupakan kata imbuhan maka akan dilakukan *stemming* dengan melalui 3 tahapan yaitu menghapus *inflection suffix* (seperti -ku, -mu, -kah, dsb), menghapus *derivation suffix* (seperti -i, -an,atau -kan), dan menghapus *derivation prefix* (seperti di-, ke-,se-, dsb). Hasil dari proses *stemming* ini akan dilanjutkan dengan tahap berikutnya untuk dilakukan pembobotan kata menggunakan *algoritma* tf-idf[3][7].



Gambar 3. *Flowcart Stemming*

Perhitungan *Term-Frequency Inverse Document Frequency*

Sebelum melakukan perhitungan *Term-Frequency Inverse Document Frequency* maka harus melakukan proses tokenisasi, *stopwords, stemming*. Setelah proses tersebut selesai peneliti melakukan perhitungan *Term-Frequency Inverse Document Frequency* dengan tahap berikut [8].

- ✓ Menetukan TF (*Therm Frekuensi*) terdiri dari Q, D1, D2, D3
  Q = Query
  D1 = Dokumen 1
  D2 = Dokumen 2
  D3 = Dokumen 3
- ✓ Menghitung df (data frekuensi)
  Dengan rumus : df = Q + D1 + D2 + D3
- ✓ Menghitung D/df

- ✓ Menghitung Idf = log(D/df)
- ✓ Menentukan W (Pembobotan kata) terdiri dari Q, D1, D2, D3
  Q = log(D/df) x Q

PEMBAHASAN

Perhitungan menggunakan metode *Term-Frequency Inverse Document Frequency* disini digunakan untuk mencari relevansi antar beberapa dokumen artikel yang sesuai dengan *query*, adapun *query* yang digunakan yaitu kata Ibu Kota. Alasan digunaknnya query ini karena memang banyak sekali artikel-artikel berita yang sedang membahas permasalahan perpindahan ibu kota Indonesia. Perhitungan relevansi data dapat digunakan sebagai acuan untuk melihat manakah artikel yang sesuai dengan isu perpindahan ibu kota dari Jakarta ke Kalimantan dan yang tidak sesuai [9][10][7]. Berikut beberapa hasil perhitungan dari tiga artikel berita yang di akses pada tanggal 25 November 2019 yang diterbitkan oleh CNN Indonesia dengan judul Presiden Korsel tawarkan kerja sama pindah ibu kota ke Jokowi, KOMPAS.com berjudul Jepang ingin terlibat pemindahan ibu kota, Presiden Jokowi "*Welcome*" dan CNBC Indonesia yang berjudul Aturan Pemindahan Ibu Kota Akan Masuk *Omnibus Law*, tiga artikel tersebut kami ambil karena memang sedang membahas tentang isu perpindahan ibu kota Indonesia dari Jakarta ke Kalimantan.

Tabel 1. hasil proses tokenisasi, *stopword*, *stemming*

Token	tf				df
	Q	D1	D2	D3	
ibu kota	1	8	1	6	15
pindah	0	0	1	1	2
bagai	0	0	1	1	2
Menteri	0	6	0	2	8
Koordinat	0	1	0	0	1
Bidang	0	1	0	2	3
Ekonomi	0	1	3	0	4
Airlangga	0	1	0	0	1
Luar negeri	0	1	0	1	2
Kerja	0	4	6	10	20
Umum	0	1	0	1	2
Rumah	0	1	0	2	3
Rakyat	0	1	0	0	1
Seskab	0	1	0	0	1
Temu	0	2	1	1	4
Usaha	0	4	0	1	5
Jepang	0	11	0	0	11
Kutai	0	1	0	1	2
Kartanegara	0	1	0	0	1
Penajam	0	1	0	0	1
Kalimantan Timur	0	1	1	3	5

Tabel 1 adalah Gambar hasil proses tokenisasi, *stopword*, *stemming*, serta perhitungan jumlah query dan jumlah kata yang muncul disetiap artikel yang digunakan sebagai sempel sedangkan Tabel 2 berisi perhitungan data frekeuensi. Tabel 3 merupakan hasil akhir *Rank Score* dari proses tokenisasi, *stopword*, *stemming* serta perhitungan yang menerapkan metode TF-IDF dari tiga artikel yang telah disebutkan diatas.

Tabel 2. Perhitungan Frekuensi

df	D/df	Idf= log(D/df)	W			
			Q	D1	D2	D3
15	0,2	-0,698970004	-0,69897	-5,59176	-0,69897	-4,19382
2	1,5	0,176091259	0	0	0,17609	0,17609
2	1,5	0,176091259	0	0	0,17609	0,17609
8	0,375	-0,425968732	0	-2,55581	0	-0,85194
1	3	0,477121255	0	0,47712	0	0
3	1	0	0	0	0	0
4	0,75	-0,124938737	0	-0,12494	-0,37482	0
1	3	0,477121255	0	0,47712	0	0
2	1,5	0,176091259	0	0,17609	0	0,17609
20	0,15	-0,823908741	0	-3,29563	-4,94345	-8,23909
2	1,5	0,176091259	0	0,17609	0	0,17609
3	1	0	0	0	0	0
1	3	0,477121255	0	0,47712	0	0
1	3	0,477121255	0	0,47712	0	0
4	0,75	-0,124938737	0	-0,24988	-0,12494	-0,12494
5	0,6	-0,22184875	0	-0,88739	0	-0,22185
11	0,27273	-0,56427143	0	-6,20699	0	0
2	1,5	0,176091259	0	0,17609	0	0,17609
1	3	0,477121255	0	0,47712	0	0
1	3	0,477121255	0	0,47712	0	0
5	0,6	-0,22184875	0	-0,22185	-0,22185	-0,66555

Tabel 3 Rank Score

	15,2762	0,23869	8,59284
Rank Score	**3,90847**	**0,46022**	**2,9262**

KESIMPULAN

Penelitian ini dapat disimpulkan bahwa, dengan metode TF-IDF di atas dapat di ambil garis besarnya yaitu mudahnya mencari data yang relevan dengan metode iTF-IDF dan mudahnya mempelajari bagi para pemula terutama seorang *programing* yang ingin merintis melalui metode ini. Adapun yang harus diperhatikan ialah relevansi dari suatu dokumen tergantung pada nilai *rank* dari setiap dokumen yang ada. Dari tiga artikel yang digunakan untuk penelitian, dapat disimpulkan bahwa artikel berita yang berjudul Presiden Korsel Tawarkan Kerja Sama Pindah Ibu Kota ke Jokowi adalah yang paling relevan dari pada dua artikel lainnya dengan nilai 3,90847.

REFERENSI

[1] R. C. W. Romario Yudo Herlambang, Rekyan Regasari Mardi Putri, "IMPLEMENTASI METODE K-NEAREST NEIGHBOUR DENGAN PEMBOBOTAN," vol. 4, no. 2, pp. 97–103, 2017.

[2] S. Andayani and A. Ryansyah, "Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen," *JuSiTik J. Sist. dan Teknol. Inf. Komun.*, vol. 1, no. 1, p. 53, 2017.

[3] T. D. Ria Melita , Victor Amrizal , Hendra Bayu Suseno, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018.

[4] S. Najah, M. A. Yaqin, L. S. Angreani, and A. C. Fauzan, "The User Interface (UI) Discovery Application to Measure Query Accuracy on Interface Repository," *brilliant*, vol. 3, no. May, pp. 1–6, 2019.

[5] S. Najah, M. A. Yaqin, L. S. Angreani, and A. C. Fauzan, "MENGUKUR AKURASI QUERY PADA INTERFACE REPOSITORY MENGGUNAKAN USER INTERFACE (UI) DISCOVERY

BERBASIS SOFTWARE AS A SERVICE (SAAS)," vol. 4, no. 1, pp. 206–214, 2019.

[6] A. Indranandita, B. Susanto, and A. Rahmat, "Sistem Klasifikasi Dan Pencarian Jurnal Dengan Menggunakan Metode Naive Bayes Dan Vector Space Model," *J. Inform.*, vol. 4, no. 2, 2011.

[7] andi sunyoto Mayeni, Monica , wing wahyu wirnano, "Information Retrieval Dokumen Tesis Untuk," vol. 12, no. 2, pp. 105–115, 2016.

[8] D. Leman, K. Andesa, T. Informasi, M. Komputer, and U. P. Utama, "Implementasi Vector Space Model Untuk Meningkatkan," no. May 2013, pp. 8–15, 2015.

[9] A. Salam, C. Supriyanto, and A. Fahmi, "Integrasi Peringkas Dokumen Otomatis Sebagai Feature Reduction Pada Clustering Dokumen," *Semantik*, vol. 02, no. 01, pp. 145–150, 2012.

[10] R. R. A. Siregar, F. A. Sinaga, and R. Arianto, "Aplikasi Penentuan Dosen Penguji Skripsi Menggunakan Metode TF-IDF dan Vector Space Model," *Comput. J. Comput. Sci. Inf. Syst.*, vol. 1, no. 2, p. 171, 2017.