IMPLEMENTASI *TOKENIZING PLUS* PADA SISTEM PENDETEKSI KEMIRIPAN JURNAL SKRIPSI

Paratisa Kharismadita[1], Faisal Rahutomo[2]

Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1]tisa.kharisma@gmail.com, [2]faisal.polinema@gmail.com

Abstrak

Syarat lulus bagi mahasiswa program sarjana, magister dan doktor salah satunya adalah mempublikasikan karya ilmiah. Untuk lulus Sarjana harus menghasilkan jurnal yang terbit pada jurnal ilmiah. Namun banyak sekali kasus plagiarisme atau penjiplakan jurnal yang marak terjadi di Indonesia. Tidak hanya dikalangan mahasiswa program sarjana namun juga terjadi pada beberapa kasus di program magister dan doctoral di beberapa instansi pendidikan. Penerapan sistem pendeteksi kemiripan jurnal tentunya sangat diperlukan untuk mengurangi kasus plagiarisme dikalangan pendidikan. Tahapan yang harus dilalui pada sistem yaitu Tokenizing Plus (membuat library kata berdasarkan KBBI). Tokenizing Plus merupakan proses untuk mendapatkan kata dasar dan kata majemuk yang ada pada KBBI. Metode yang digunakan adalah *Term* Frequency dan Inverse Document Frequency (TF-IDF) dan Cosine Similarity untuk mendapatkan nilai kemiripan. Sistem ini membandingkan keseluruhan dari isi jurnal mulai dari abstrak, judul dan konten.

Kata kunci : Plagiarisme, KBBI, TFIDF, *Cosine Similarity*

1. Pendahuluan

Direktorat Jenderal Pendidikan Tinggi (Ditjen Dikti) Kementerian Pendidikan dan Kebudayaan mengeluarkan surat edaran bernomor 152/E/T/2012 terkait publikasi karya ilmiah. Surat tertanggal 27 Januari 2012 ini ditujukan kepada Rektor/Ketua/Direktur PTN dan PTS seluruh Indonesia. Seperti dimuat dalam laman www.dikti.go.id, surat yang ditandatangani Direktur Jenderal Pendidikan Tinggi Djoko Santoso itu memuat tiga poin yang menjadi syarat lulus bagi mahasiswa program S-1, S-2, dan S-3 untuk mempublikasikan karya ilmiahnya (Wedhaswary, 2012). Untuk lulus program Sarjana harus menghasilkan jurnal yang terbit pada jurnal ilmiah.

Namun banyak sekali kasus plagiarisme atau penjiplakan jurnal yang marak terjadi di Indonesia. Tidak hanya dikalang mahasiswa S-1 namun juga terjadi pada beberapa kasus pada program magister dan doctoral di beberapa instansi pendidikan.

Plagiarisme adalah perbuatan sengaja atau tidak sengaja dalam memperoleh atau mencoba memperoleh kredit atau nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan atau karya ilmiah pihak lain yang diakui sebagai karya ilmiahnya, tanpa menyatakan sumber secara tepat dan memadai (Istiana, 2012).

Dalam hal ini penulis mencoba menerapkan sistem untuk mengidentifikasi kemiripan antara jurnal yang akan diterbitkan dengan jurnal yang sudah diterbitkan pada jurnal ilmiah namun pembandingan masih dalam lingkup program studi Teknik Informatika. Sistem ini mempunyai beberapa tahapan yang harus dilalui yaitu *Tokenizing Plus* (memebuat library kata berdasarkan KBBI), *Filtering* (pembuangan *term* yang tidak mempunyai relevansi), Pembobotan kata (menggunakan metode *Term Frequency* dan *Inverse Document Frequency*) dan Pengukuran kemiripan (menggunakan rumus *Cosine Similarity*). Sistem ini memandingkan kesuluruhan dari isi jurnal mulai dari abstrak, judul dan konten.

2. Tinjauan Pustaka

2.1 Plagiarisme

Plagiarisme adalah kata benda, yang artinya "penjiplakan yang melanggar hak cipta". Tindakan melakukan plagiarisme disebut plagiat, yang berarti "pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat dan sebagainya) sendiri.

2.2 KBBI

Kamus Besar Bahasa Indonesia yang dikenal dengan sebutan KBBI terbit pertama 28 Oktober 1988 saat Pembukaan Kongres V Bahasa Indonesia. Sejak itu kamus tersebut telah menjadi sumber rujukan yang dipercaya baik di kalangan pengguna di dalam maupun di luar negeri. Setiap ada permasalahan tentang kata, KBBI selalu dianggap sebagai jalan keluar penyelesaiannya (Sugono, 2008).

2.3 Tokenizing Plus

Tokenizing Plus merupakan preprocessing pada sistem pendeteksi jurnal skripsi ini. *Tokenizing Plus* merupakan proses converting Kamus Besar Bahasa Indonesia (KBBI) menjadi suatu *libabry* sehingga didapatkan kata majemuk.

2.4 TF-IDF

Metode TF-IDF (Robertson, 2004) merupakan metode untuk menghitung bobot setiap setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, simple dan memiliki hasil yang akurat (Ramos, 2010).

Dalam tahapan perhitungan *term frequency* (tf) menggunakan persamaan

$$\boldsymbol{tf = tf_{ij}} \quad (1)$$

Perhitungan *Inverse Document Frequency* (idf), menggunakan persamaan

$$\boldsymbol{idf_i = log \frac{N}{dfi}} \quad (2)$$

Perhitungan *term frequency inverse document frequency (tfidf)*, menggunakan persamaan

$$\boldsymbol{W_{ij} = tf_i \, . \, idf_i} \quad (3)$$

Dengan W_{ij} adalah bobot dokumen. Bobot dokumen (W_{ij}) dihitung untuk didapatkannya suatu bobot hasil perkalian atau kombinasi antara *term frequency* (tf_i) dan *inverse document frequency* (idf_i).

2.5 Cosine Similarity

Cosine similarity untuk pengukuran kesamaan antara dokumen dan *user query* harus mengakomodasi kata yang yang berarti. *Cosine similarity* masih belum bisa menangani makna semantik teks dengan sempurna.

Berdasarkan vektor kesamaan, kesamaan antara dua vektor dapat didefinisikan sebagai:

$$\boldsymbol{Sim(\vec{q}, \vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\Sigma_{k=1}^{t} w_{qk} x \, w_{dk}}{\sqrt{\Sigma_{k=1}^{t}(w_{qk})^2} . \sqrt{\Sigma_{k=1}^{t}(w_{dk})^2}}} \quad (4)$$

3. Perencanaan

3.1 Deskrispsi Sistem

Aplikasi pendeteksi kemiripan jurnal skripsi di Program Studi Teknik Informatika Politeknik Negeri Malang ini bertujuan untuk menampilkan informasi nilai kemiripan antara jurnal yang akan diunggah dengan jurnal yang sudah ada pada basis data dengan menggunakan ekstraksi Kamus Besar Bahasa Indonesia sebagai penciri dokumen.

3.2 Desain Sistem

Secara sederhana deskripsi umum Sistem yaitu *User* hanya dapat mengguanggah *file* jurnal kemudian sistem memerlukan database dan ekstraksi KBBI untuk mengolah data tersebut sehingga pada akhirnya *user* dapat mengetahui seberapa besar kemiripan jurnal yang diunggah dengan jurnal yang sudah ada pada *database*.

Gambar 1. Desain Sistem

3.3 Flowchart

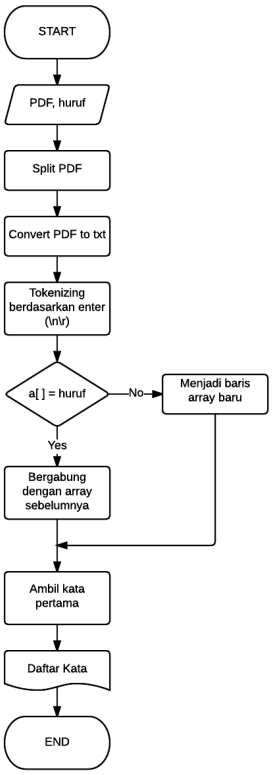
3.3.1 Flowchart Ekstraksi KBBI

Berikut merupakan flowchart alur pada proses mengekstraksi Kamus Besar Bahasa Indonesia (KBBI) yang bernama *tokenizing plus*. Tujuan *tokenizing plus* adalah mendapatkan kata dan gabungan kata (kata majemuk) yang ada pada format KBBI yang sudah diatur oleh tim penyusun sehingga diperlukan beberapa proses untuk mendapatkan kata-kata tersebut.

Gabungan kata atau kelompok kata yang tidak berderivasi di perlakukan sbg sublema. Letaknya langsung di bawah lema yang berkaitan dan disusun berderet ke samping secara berurutan menurut abjad. Unsur pertama gabungan kata itu dicetak dengan tanda hubung ganda (--) (KBBI, 2008). Gabungan kata merupakan kata majemuk yang selalu diungkit pada setiap bahasan karena besarnya pengaruh kata majemuk pada penelitian ini.

dokter *n* sarjana lulusan pendidikan kedokteran yg ahli dalam hal penyakit dan pengobatannya;
-- **anak** dokter yg khusus mengobati penyakit anak-anak; -- **bedah** dokter ahli bedah; -- **gigi** dokter yg mempunyai keahlian dl pengobatan gigi; -- **gula** ahli kimia di laboratorium pabrik gula; -- **hewan** dokter khusus untuk penyakit hewan; -- **Jawa** dokter keluaran sekolah zaman dahulu (Nias, Stovia); -- **jiwa** dokter yg khusus mengobati penyakit jiwa; -- **mata** dokter yg khusus mengobati penyakit mata; -- **spesialis** dokter yg khusus ahli dl satu macam penyakit;

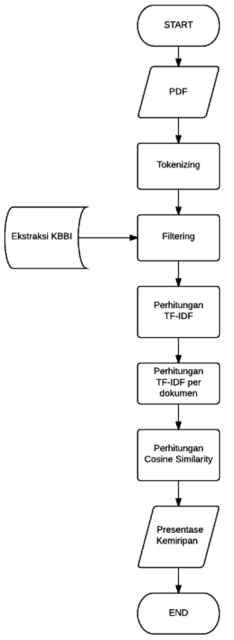
Gambar 2. Format KBBI gabungan kata



Gambar 3. Flowchart Ekstraksi KBBI

3.3.2 Flowchart Alur Sistem

Berikut merupakan flowchart alur sistem mulai dari mengunggah file berupa pdf hingga sistem menampilkan nilai kemiripan dari file pembanding.

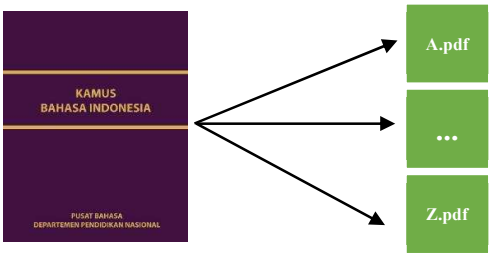


Gambar 4. Flowchart Alur Sistem

4. Uji Coba

4.1 Ekstraksi KBBI

Tokenizing Plus merupakan tahapan untuk mendapatkan kata dasar dan kata majemuk yang ada di Kamus besar Bahasa Indonesia. Ada beberapa proses awal yang harus dilakukan sebelum mulai masuk pada proses kode program. KBBI yang digunakan adalah edisi keempat yang telah disahkan oleh Menteri Pendidikan Nasional yaitu Prof. Dr. Bambang Sudibyo pada 28 Oktober 2008. KBBI edisi keempat mempunyai 1826 halaman dan 91.000 lema yang terdiri dari beberapa jenis susunan kata yang tidak hanya gabungan kata maupun kata dasar saja.



Gambar 5. Pemisahan KBBI

Tahapan awal adalah pemisahan KBBI yang berbentuk pdf ke beberapa 26 pdf berukuran kecil yang berisikan lema berdasarkan urutan abjad. Tujuan memisahkan KBBI adalah untuk mempermudah proses pengambilan kata karena halaman awal KBBI merupakan halaman pembukaan dan tata cara penggunaan. Setelah proses pemisahan pdf selesai, proses selanjutnya adalah pengubahan jenis file yang awalnya berbentuk pdf diubah menjadi berbentuk *plain text* (txt) agar format yang ada sebelumnya tidak berubah.

Gambar 6. *Convert pdf to txt* menggunakan word

Preprocessing telah selesai dilakukan maka selanjutnya adalah masuk pada tahapan pengambilan kata sepeti yang sudah dijelaskan pada bab implementasi dengan menggunakan *rule base* yaitu membuat aturan tersendiri pada pengkodean program agar didapatkan hasil yang diinginkan. Hasil dari ekstraksi KBBI adalah didapatkan 13.148 kata yang terdiri dari kata dasar, kata majemuk dan kata imbuhan.

Dapat diambil kesimpulan bahwa hasil tidak sesuai dengan yang telah dijabarkan oleh KBBI yaitu terdiri dari 91.000 lema dikarenakan penelitian hanya mengambil dua susunan kata yaitu kata dasar dan gabungan kata (kata majemuk).

4.2 TFIDF dan Cosine Similarity

Tabel 1. Hasil Perhitungan *Cosine Similarity*

	doc-1	doc-2	doc-3	doc-4	doc-5	doc-6	doc-7	doc-8	doc-9	doc-10
doc-1		0.32157	0.01361	0,00272	0,35553	0,01299	0,00369	0,05007	0	0,02993
doc-2			0.01021	0,00142	0,23153	0,15182	0,00193	0,04502	0,01392	0,04711
doc-3				0,01002	0,00854	0	0,05531	0,02639	0,01427	0,00902
doc-4					0,07292	0,03487	0,01462	0,00116	0	0,02827
doc-5						0,05781	0,01385	0,00212	0,01727	0,04271
doc-6							0,00093	0,05423	0,00922	0,02512
doc-7								0,00158	0	0,01479
doc-8									0,35062	0,02322
doc-9										0,01187
doc-10										

Pada uji coba sistem pendeteksi kemiripan jurnal menggunakan 10 dokumen dengan diwakili 10 *term* yang memiliki nilai frekuensi tertinggi disetiap dokumen sehingga didapat 100 nilai dan 47 *term*.

TF-IDF (Robertson, 2004) merupakan metode untuk menghitung bobot setiap setiap kata yang paling umum digunakan pada *information retrieval*. TF IDF adalah bobot dokumen (W_{ij}) dihitung untuk didapatkannya suatu bobot hasil perkalian atau kombinasi antara *term frequency* (tf_i) dan *inverse document frequency* (idf_i).

Tabel 1 merupakan hasil pengimplentasian perhitungan TFIDF yang diteruskan dengan perhitungan *Cosine Similarity* untuk menghasilkan nilai kemiripan antara 2 dokumen dengan *range* nilai antara 0 hingga 1. Jika kedua dokumen bernilai 1 maka kedua dokumen tersebut sama.

4.3 Uji Performa

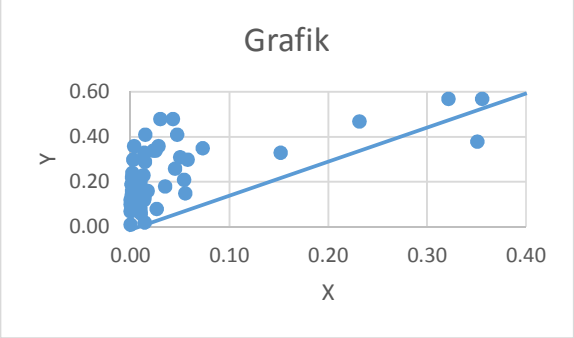
Analisis regresi sebagai kajian terhadap hubungan satu variabel yang disebut sebagai variabel yang diterangkan (*the explained variabel*) dengan satu atau dua variabel yang menerangkan (*the explanatory*). Variabel pertama disebut juga sebagai variabel tergantung dan variabel kedua disebut juga sebagai variabel bebas. Jika variabel bebas lebih dari satu, maka analisis regresi disebut regresi linear berganda. Disebut berganda karena pengaruh beberapa variabel bebas akan dikenakan kepada variabel tergantung..

Tujuan menggunakan analisis regresi ialah membuat estimasi rata-rata dan nilai variabel tergantung dengan didasarkan pada nilai variabel bebas.

Pada regresi linear mempunyai dua rumus yaitu untuk mencari variabel a dan mencari variabel b agar dapat dihasilkan nilai y yang berasal dari persamaan $y = a + bx$. Berikut merupakan rumus mencari a dan b.

$$b = \frac{n\sum xy - (\sum x.\sum y)}{n\sum x^2 - (\sum x)^2} \quad (5)$$

$$a = \frac{(\sum y - (b.\sum x))}{n(\sum x^2) - (\sum x)^2} \quad (6)$$



Gambar 7. Grafik Uji Peforma

5. Penutup

5.1 Kesimpulan

Sistem pendeteksi kemiripan jurnal skripsi dibuat dengan menggunakan metode *Term Frequency and Inverse Document Frequency (TFIDF)* sebagai perhitungan *term* disetiap dokumennya dan *Cosine Similarity* untuk menghitung kemiripan antara 2 jurnal yang menghasilkan nilai 0 jika kedua jurnal sangatlah berbeda dan nilai 1 jika kedua jurnal mempunyai *term* yang sama.

Pada tahapan *preprocessing* dilakukan proses ekstraksi Kamus Besar Bahasa Indonesia (KBBI) yang dinamakan *Tokenizing Plus* untuk didapatkannya kata majemuk yang digunakan pada saat proses *tokenizing* sehingga didapatkan kata majemuk yang dapat dijadikan salah satu acuan untuk proses perhitungan TFIDF dan Cosine Similarity.

5.2 Saran

Pemahaman format tata tulis pada Kamus Besar Bahasa Indonesia (KBBI) harus lebih ditingkatkan agar kata yang didapatkan lebih akurat dan tidak melakukan supervisi secara manual setelah semua *preprocessing* selesai.

Data *term* yang didapatkan disetiap dokumen harus diperbanyak agar nilai kemiripan mendekati dengan nilai sebenarnya.

6. Daftar Pustaka

Ahmad Hatta, A. 2010. "Rancang Bangun Sistem Pengelolaan Dokumen-dokumen Penting Menggunakan Text Mining". *Politeknik Elektronika Negeri Surabaya*

Amin, F. 2012. ", Implementasi *Search Engine* (Mesin Pencari) Menggunakan Metode *Vector Space Model*". *Fakultas Teknologi Informasi Universitas Stikubank Semarang*

Istiana, P dan Purwoko. 2012. *Panduan Anti Plagiarism*. [Online] Tersedia: http://old.lib.ugm.ac.id/exec.php?app=site&act=pandanplagi [3 Januari 2015]

Khairunnisa, N., dkk. 2012. "Aplikasi Pendeteksi Plagiat dengan Menggunakan Metode *Latent Semantic Analysis* (Studi Kasus : Laporan TA PCR)". *Politeknik Caltex Riau, Pekanbaru, Riau*

Prasetyo, B dan Suzana. 2012. "Pengembangan Aplikasi Cerdas Berbasis Intelegensia Buatan Untuk Perbandingan Karya Ilmiah Hasil Penelitian Dalam Upaya Mencegah Plagiasi". *Lokakarya Komputasi dalam Sains dan Teknologi Nuklir Serpong*

Pusat Bahasa, 2008. "Kamus Bahasa Indonesia". *Departemen Pendidikan Nasional Jakarta*.

Salton, G and Buckley C. 1988. "Term-weighting Approaches in Automatic Text Retrieval". *Information Processing and Managemen*t, vol.24, no.5,1988, pp.513−523

Sugono, D. 2008. *Selamat datang di KBBI Daring*. [Online] Tersedia : http://badanbahasa.kemdikbud.go.id/kbbi/ [1 Juni 2015]

Wedhaswary, I.D. 2012. *Syarat Lulus S-1, S-2, S-3: Harus Publikasi Makalah.* [Online] Tersedia : http://edukasi.kompas.com/read/2012/02/03/09280630/Syarat.Lulus.S-1.S-2.S-3.Harus.Publikasi.Makalah [3 Januari 2015]