1. Information Engineering

PERANCANGAN SISTEM IDENTIFIKASI DAN BASIS DATA KATA KUNCI MINAT DOSEN UNTUK UJIAN PENDADARAN

TUGAS AKHIR

Diajukan untuk memenuhi sebagian persyaratan

mencapai derajat Sarjana Teknik Industri

Jones Averino

16 06 08729

PROGRAM STUDI TEKNIK INDUSTRI

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ATMA JAYA YOGYAKARTA

YOGYAKARTA

2020

HALAMAN PENGESAHAN

Tugas Akhir Berjudul

PERANCANGAN SISTEM IDENTIFIKASI DAN BASIS DATA KATA KUNCI MINAT DOSEN UNTUK UJIAN PENDADARAN

yang disusun oleh

JONES AVERINO

160608729

dinyatakan telah memenuhi syarat pada tanggal 23 Juli 2020

		Keterangan
Dosen Pembimbing 1	: Anugrah Kusumo Pamosoaji, S.T., M.T.	Telah menyetujui
Dosen Pembimbing 2	: Ririn Diar Astanti, D.Eng.	Telah menyetujui
Tim Penguji		
Penguji 1	: Anugrah Kusumo Pamosoaji, S.T., M.T.	Telah menyetujui
Penguji 2	: Kristanto Agung Nugroho, S.T., M.Sc.	Telah menyetujui
Penguji 3	: Dr. T. Baju Bawono, ST., MT.	Telah menyetujui

Yogyakarta, 23 Juli 2020

Universitas Atma Jaya Yogyakarta

Fakultas Teknologi Industri

Dekan

ttd

Dr. A. Teguh Siswantoro, M.Sc

HALAMAN PENGESAHAN

Tugas Akhir berjudul

PERANCANGAN SISTEM IDENTIFIKASI DAN BASIS DATA KATA KUNCI MINAT DOSEN UNTUK UJIAN PENDADARAN

yang disusun oleh

Jones Averino

16 06 08729

Dinyatakan telah memenuhi syarat pada tanggal 2 Juli 2020

Menyetujui,

Dosen Pembimbing I	Dosen Pembimbing II,
Anugrah Kusumo P., S.T., M.T	Ririn Diar A., S.T., M.MT., D.Eng

Tim Penguji,

Penguji 1,	Penguji 2,
Kristanto Agung N., S.T., M.Sc.	Dr. T. Baju Bawono, ST., MT.

Yogyakarta, 22 Juli 2020

Universitas Atma Jaya Yogyakarta,

Fakultas Teknologi Industri,

Dr. A. Teguh Siswantoro, M.Sc

Pernyataan Originalitas

Commented [KANS1]: Bold

Saya yang bertanda tangan di bawah ini:

Nama : Jones Averino

NPM : 160608729

Dengan ini menyatakan bahwa tugas akhir saya dengan judul “Perancangan Sistem Identifikasi dan Basis Data Kata Kunci Minat Dosen Untuk Ujian Pendadaran” merupakan hasil penelitian saya pada Tahun Akademik 2019/2020 yang bersifat original dan tidak mengandung plagiasi dari karya manapun.

Bilamana di kemudian hari ditemukan ketidak sesuaian dengan pernyataan ini, maka saya bersedia dituntut dan diproses sesuai dengan ketentuan yang berlaku termasuk untuk dicabut gelar Sarjana yang telah diberikan Universitas Atma Jaya Yogyakarta kepada saya.

Demikian pernyataan ini dibuat dengan sesungguhnya dan dengan sebenar-benarnya.

Yogyakarta, 22 Juli 2020

Yang menyatakan,

Jones Averino

Kata Pengantar

Puji dan syukur mahasiswa sampai kepada Tuhan Yang Maha Esa atas perlindungan, rahmat dan kuasanya sehingga mahasiswa dapat menyelesaikan Tugas Akhir yang berjudul “Perancangan Sistem Identifikasi dan Basis Data Kata Kunci Minat Dosen Untuk Ujian Pendadaran”.

Tugas Akhir ini dibuat sebagai salah satu syarat yang dipenuhi untuk mencapai derajat kesarjanaan di Program Studi Teknik Industri, Universitas Atma Jaya Yogyakarta.

Dalam proses kerja praktek maupun pengerjaan laporannya, mahasiswa ingin mengucapkan rasa terima kasih yang sedalam-dalamnya kepada semua pihak yang telah membantu mahasiswa:

1. Bapak Anugrah Kusumo Pamosoaji, S.T., M.T dan Ibu Ririn Diar Astanti, S.T., M.MT., D.Eng sebagai dosen pembimbing mahasiswa saat penyelesaian Tugas Akhir.
2. Kedua orang tua, adik dan segenap keluarga yang selalu senantiasa mendukung dan memberikan semangat selama kegiatan kerja praktek
3. Audrey Tejawijaya yang terus membantu dan memberikan tambahan semangat selama pengerjaan Tugas Akhir.
4. Teman-teman yang telah membantu pada saat pembuatan Tugas Akhir.
5. Semua pihak lain yang tidak bisa disebutkan namanya satu per satu yang juga telah membantu mahasiswa menyelesaikan kegiatan dan laporan kerja praktek

.

Mahasiswa menyadari bahwa laporan ini jauh dari kata sempurna namun senantiasa berharap agar dapat beguna dan bermanfaat bagi untuk seluruh pihak terkait dan yang membutuhkan.

Yogyakarta, 22 Juli 2020

Penulis

DAFTAR ISI

Commented [KANS2]: Bold

DAFTAR GAMBAR

Commented [KANS3]: Bold

DAFTAR TABEL

DAFTAR LAMPIRAN

Commented [KANS5]: Bold

PERANCANGAN SISTEM IDENTIFIKASI DAN BASIS DATA KATA KUNCI MINAT DOSEN UNTUK UJIAN PENDADARAN

Jones Averino

16 06 08729

Intisari

Di Program Studi Teknik Industri Universitas Atma Jaya Yogyakarta, setiap mahasiswa perlu menyelesaikan skripsi yang akan diujikan dalam ujian pendadaran. Dalam proses ujian, mahasiswa akan diuji oleh dosen pembimbingnya serta dua dosen penguji. Pemilihan penguji untuk ujian pendadaran saat ini masih dilaksanakan dengan manual. Proses tersebut belum menerapkan penggunaan sistem basis data untuk mempermudah pencarian. Sehingga dengan cara ini berisiko terjadi pemilihan dosen penguji yang area of interestnya tidak sesuai dengan topik pengujian.

Pembuatan basis data ini akan menggunakan metode text mining yang mengintegrasikan beberapa teknik *natural language processing*. Data yang digunakan untuk *text mining* berasal dari ujian pendadaran yang telah dilakukan oleh civitas TI UAJY. Basis data yang dibuat dari hasil dari *text mining* diharapkan dapat mempermudah proses pemilihan dosen untuk menguji ujian pendadaran.

Kata kunci: *text mining*, ujian pendadaran, program studi teknik industri, universitas atma jaya yogyakarta, *natural language processing,* basis data, pemilihan dosen.

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Proses ujian pendadaran adalah tahapan terakhir dalam kegiatan mata kuliah tugas akhir di Program Studi Teknik Industri Universitas Atma Jaya Yogyakarta (PSTI UAJY). Ujian pendadaran atau yang biasa disebut dengan skripsi menjadi ujian terakhir dan terberat bagi seorang mahasiswa TI UAJY yang harus mempersiapkan hasil tugas akhirnya dan mempresentasikannya di hadapan tiga hingga empat orang dosen.

Pendaftaran ujian pendadaran sendiri dilakukan oleh mahasiswa di dalam kantor PSTI UAJY, yaitu dilakukan oleh ketua PSTI UAJY. Oleh karena sifat hasil tugas akhir dari mahasiswa yang terspesialisasi ke dalam satu topik maka dosen yang akan menguji presentasi tersebut diprioritaskan merupakan dosen yang memiliki bidang minat atau *area of interest* di bidang tersebut.

Commented [j6]: Revisi mahasiswa saja

Salah satu masalah dalam proses pendaftaran ujian pendadaran adalah pencocokan topik skripsi dengan *area of interest* dosen yang ada. Walaupun semua dosen memahami ketiga belas *body of knowledge of industrial engineering* menurut Institute of Industrial and System Engineers (IISE) (PSTI UAJY, 2020), tentunya dosen akan dapat lebih paham dengan topik yang masuk dengan *area of interest*-nya.

Proses yang ada selama ini dilakukan dengan secara manual mengidentifikasi topik yang ada di dalam judul skripsi dan kemudian secara manual mencari dosen yang dapat menguji untuk topik tersebiut. Cara ini tentu berisiko terjadi pemilihan *area of interest* dosen penguji yang kurang sesuai dengan topik yang diuji. *Area of interest* dosen juga perlu dihafal oleh ketua agar dapat menentukan dosen yang akan menguji ujian tersebut dengan cepat.

Apabila ketua program studi berganti ataupun sedang berhalangan, maka orang yang akan menggantikan tugas ketua program studi harus kembali mempelajari topik-topik mata kuliah yang ada dan mencocokannya dengan judul skripsi yang ada. Hal ini tentu sulit karena judul-judul skripsi selain rumit juga mengandung *keyword* atau kata kunci yang perlu diidentifikasi. Oleh sebab itu seseorang memerlukan pengetahuan yang cukup luas untuk setiap topik yang ada. Ini akan

tetap menjadi tantangan walaupun seseorang telah cukup lama berada di dalam dunia akademis teknik industri.

Kesulitan-kesulitan dalam proses pengidentifikasian *keyword* dengan topik skripsi dan proses mencocokkan topik skripsi tersebut dengan *area of interest* dosen yang ada yang masih menggunakan proses manual dengan program Microsoft Excel membuat proses tahap awal pendaftaran ujian pendadaran cukup sulit untuk dicocokkan secara tepat yang mengakibatkan kekeliruan pada proses pemilihan dosen. Kekeliruan ini mengakibatkan dosen yang dipilih untuk menguji ujian pendadaran kurang sesuai dengan *area of interest*-nya, sehingga mempengaruhi kualitas sidang pendadaran.

Oleh karena itu, diperlukan pembuatan profil dosen yang baru untuk membantu hal tersebut. Profil tersebut dapat dibangun dengan melakukan *text mining* sebagai alat untuk membangun basis datanya.

1.2. Perumusan Masalah

Permasalahan yang akan dicari solusinya berdasarkan latar belakang yang telah dikemukakan adalah kesulitan dalam mencocokkan *area of interest* dosen dengan topik pendadaran yang akan diujikan, terutama saat mencari dosen yang *area of interestnya* berada di topik tersebut sedang berhalangan untuk menguji ujian pendadaran. Kemudian terdapat juga kesulitan pada proses identifikasi topik pendadaran berdasarkan *keyword-keyword* yang terdapat pada judul.

1.3. Tujuan Penelitian

Tujuan yang dicari dalam penelitian ini adalah untuk melakukan *text mining* demi mendapatkan *keyword* dari judul ujian pendadaran yang telah dilakukan.

1.4. Batasan Masalah

Batasan yang ditentukan dalam pelaksanaan penelitian adalah:

a. Objek penelitian adalah kantor Prodi Teknik Industri di UAJY.
b. *Basis data* yang dibuat akan diisi berdasarkan data judul pendadaran semester genap 2017/2018, semester genap 2018/2019 dan semester gasal 2019/2020.

Commented [KANS7]: 2017/2018 revisi juga yang lain

c. Dosen yang datanya dianalisis adalah dosen yang setidaknya telah dua kali terlibat dalam ujian pendadaran dan masih berstatus sebagai dosen tetap FTI UAJY.

BAB 2
TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Terdapat beberapa cara atau metode yang dapat dilakukan untuk mengidentifikasi *area of interest* dosen. Salah satu metode yang paling populer adalah *text mining*. *Text mining* telah banyak digunakan dalam berbagai aplikasi dalam kehidupan manusia mulai dari biomedis hingga keamanan.*Text mining*, atau biasa disebut dengan *text analytics,* adalah proses untuk mendapatkan informasi dalam suatu bagian ataupun kumpulan teks yang terdapat dalam suatu dokumen yang dilakukan secara otomatis melalui bantuan suatu komputer (Prilianti & Wijaya, 2014).

Text mining adalah proses analisis teks yang muncul dalam sebuah dokumen untuk menemukan dan menangkap informasi semantik untuk disimpan dalam sebuah sistem yang bernama *Knowledge Organization Structure (KOS)*. Tujuan akhirnya adalah memungkinkan pencarian pengetahuan melalui akses teks atau visual untuk digunakan pada berbagai aplikasi yang signifikan secara. Dengan demikian melalui *text mining* maka informasi-informasi yang berguna dari rangkaian teks atau kalimat dalam dokumen-dokumen, seperti *keyword*, dapat ditarik (Liddy, 2000). Tujuan dari analisis serta pengolahan *text mining* ini adalah untuk mengambil dan melihat informasi yang dapat berguna dari sebuah sumber kumpulan data teks dengan melakukan pengidentifikasian dan pengeskplorasian suatu pola yang unik ataupun menarik. Dalam melakukan *text mining*, sumber kumpulan data yang dapat dipakai adalah kumpulan dokumen yang tidak memiliki struktur dan diperlukan pengelompokkan agar dapat mendapatkan informasi yang sejenis (Santoso & Wijaya, 2016), Proses *text mining* memiliki tujuan untuk mencari dan mengidentifikasi informasi yang memiliki nilai guna dari kumpulan teks yang terdapat dalam dokumen. Kumpulan data teks yang dijadikan sumber tersebut tidak memiliki format yang terstruktur atau setidaknya termasuk dalam format semi-terstruktur (Wiguna, 2011).

Pada zaman sekarang, diperlukan adanya metode atau sistem yang digunakan untuk mengatur dan mengidentifikasi dokumen tersebut agar dapat dilakukan pencarian informasi yang releven sesuai dengan kebutuhan. Hal ini dikarenakan informasi yang berupa dokumen yang telah berbentuk digital telah tumbuh dengan

kecepatan yang cukup pesat. Pertumbuhan ini seiring berjalannya waktu terus berjalan dan jumlahnya menjadi semakin tinggi (Hariadi dkk, 2009).

Kata kunci yang dibuat oleh penulis jurnal sangat umum dijumpai dalam artikel penelitian dan tujuannya adalah untuk memberikan suatu gambaran tentang artikel tersebut untuk pembaca. Judul, abstrak, dan kata-kata kunci, dapat menyampaikan ide-ide pokok suatu jurnal, tetapi tidak selalu ada dalam dokumen. Kemudian, judul makalah dapat disimpulkan sebagai kalimat yang paling penting dalam proses ekstrasi kata-kata kunci dalam suatu dokumen (Bhowmik, 2008).

Dalam prakteknya, sebagian besar informasi yang tersedia terdapat dalam bentuk teks yang tidak terstruktur (atau lebih tepatnya dalam bentuk terstruktur secara implisit). Terdapat teknik-teknik khusus yang mengolah data tekstual yang kemudian memerlukan sebuah proses untukengekstraksi informasi dari kumpulan teks tersebut. Teknik-teknik ini dikumpulkan untuk menjadi *text mining*, dan digunakan untuk menemukan dan menggunakan struktur implisit (misalnya struktur gramatikal) dari sebuah teks. Untuk menjalankan tugasnya*, text mining* dapat menggunakan dan mengintegrasikan beberapa teknik *natural language processing* (yang digunakan misalnya untuk memproses ulang data yang berbentuk teks (Besacon & Rajman, 1998).

Istilah *Natural Language Processing (NLP)* atau pemrosesan bahasa alami mencakup serangkaian teknik untuk membuat, memanipulasi, dan menganalisis bahasa alami atau bahasa manusia secara otomatis. Meskipun sebagian besar teknik NLP diambil dari ilmu linguistik dan kecerdasan buatan, NLP juga dipengaruhi oleh bidang-bidang yang relatif baru seperti pembelajaran mesin atau *machine learning*, statistik komputasi, dan ilmu kognitif.

Untuk melakukan analisis NLP, bisa menggunakan beberapa bahasa pemrograman, seperti python. Meskipun python sudah memiliki sebagian besar fitur yang diperlukan untuk melakukan proses NLP yang sederhana, namun program ini masih mempunyai fitur yang cukup untuk melakukan sebagian besar tugas-tugas standar NLP. Di sinilah *Natural Language Tool Kit (NLTK)* atau toolkit bahasa alami diperlukan. NLTK adalah kumpulan modul dan korpora, yang dirilis dengan lisensi *open-source*, dengan ini maka NLP dan NLTK dapat dimungkinan untuk dipelajari oleh siswa dan digunakan untuk melakukan penelitian NLP (Madnani, 2007).

Teknik *text mining* didekasikan untuk mengekstrasi informasi dari data teks yang tidak terstruktur dan *Natural Language Processing (NLP)* bisa digunakan sebagai alat untuk memperbaiki prosedur ekstrasi informasi.

Dua cara dasar untuk melakukan pengindeksan teks secara penuh adalah dengan menggunakan *term frequency* (perhitungan berapa kali kemunculan suatu istilah di sebuah dokumen) dan *inverse document frequency* (pentingnya suatu istilah berbanding terbalik dengan banyaknya istilah itu muncul di dokumen-dokumen). Kata-kata yang sering muncul dapat disimpan di dalam *stop list* dan dihapus dari teks sebelum sebuah indeks dibuat (contohnya untuk kata “di” dan “dan”) (Hulth, 2003).

Dalam sebuah studi eksperimen, berdasarkan hasil metrik yang dilakukan, *term frequency* lebih unggul dalam kumpulan fitur data yang ukurannya kecil (di bawah 1000), terutama untuk ukuran di bawah 200). Hasil obervasi menunjukkan bahwa akumulasi informasi dalam data terkumpul dalam kecepatan yang lebih tinggi walaupun memiliki penyebaran yang cukup tinggi antar kelas data (Azam & Yao, 2012).

2.2. Dasar Teori

2.2.1. Text Mining

Text Mining adalah sebuah bidang baru yang mencakup metode penelitian dan *software* baru yang digunakan dalam bidang akademis dan juga oleh perusahaan-perusahaan dan badan pemerintah untuk mencari informasi dalam data teks. Peneliti telah menggunakan *text mining* di proyek-proyek besar untuk memprediksi semuanya mulai dari pasar saham dan juga kejadian demonstrasi atau protes politik. *Text Mining* juga sering digunakan dalam riset *marketing* dan aplikasi bisnis lainnya baik dalam bidang pemerintahan ataupun pertahanan.

Dalam beberapa tahun terakhir, *text mining* juga mulai digunakan dalam *social science*, yang meliputi penggunaan di biand antropologi, komunikasi, ekonomi, edukasi, sains politik, psikologi, dan sosiologi.

T*ext mining* biasanya dalam pengambilan atau ekstrasi informasi (metode untuk mendapatkan teks), dalam metode stastitik lanjutan, yang seringkali melibatkan *Natural Language Processing (NLP)* yang digunakan untuk *part-of-speech tagging*dan *syntactic parsing*. *Text mining* juga seringkali melibatkan *Named Entity Recognizition (NER)*, suatu proses yang menggunakan penggunaan teknik

statistik untuk mengidentifikasi fitur teks bernama seperti orang, organisasi, dan nama tempat (Ignatow & Mihalcea, 2018).

Proses untuk melakukan ekstrasi data pada suatu struktur teks data dinamakan sebagai *text extraction*. *Text extraction* adalah suatu bagian dari *text mining* yang merupakan teknik analisis teks yang mengekstrasi bagian-bagian data yang spesifik seperti *keywords*, nama entitas, alamat, email, dan sebagainya. Dengan menggunakan *text extraction*, maka suatu perusahaan dapat menghindari kesusahan yang dialami pada saat menyortir data secara manual untuk mengambil informasi-informasi penting yang diinginkan.

Text extraction dapat dibagi menjadi 3 jenis, yaitu *keyword extraction*, named *entity recognizition*, dan *feature extraction*. *Keyword extraction* merupakan proses untuk mengambil kata-kata kunci relevan (*keywords*) dari sebuah teks yang dapat digunkan untuk meringkas isi dari konten tersebut. Dengan menggunakan ekstrator *keyword* maka seseorang dapat melakukan proses pengindeksan data yang dapat dengan mudah dicari, meringkas isi dari sebuah teks, membuat t*ag clouds*, dan sebagainya. *Named Entity Recognzition* (NER) memungkinkan seseorang untuk mengidentifikasi dan mengekstrasi nama dari perusahaan, organisasi, atau orang dari sebuah teks. *Feature extraction* membantu mengidentifikasi karakteristik tertentu dari sebuah produk atau jasa dalam sebuah set data. Contohnya adalah pada saat melakukan analisis deskripsi produk, maka seseorang dapat dengan mudah mengekstrasi fitur-fitur seperti warna, merk, model dan lain sebagainya (Cuoto, 2019).

2.2.2. Natural Language Processing

Natural Language Processing, atau biasa disingkat NLP, merupakan manipulasi dari bahasa alami manusia, seperti teks atau percakapan, oleh perangkat lunak komputer (Brownlee, 2019).

Istilah *'Natural Language Processing'* (*NLP*) biasanya digunakan untuk menggambarkan sebuah fungsi dari komponen perangkat lunak atau perangkat keras dalam sistem komputer yang menganalisis atau mensintesiskan bahasa lisan atau tulisan. Kata *'natural'*, atau alami, maksudnya adalah untuk membedakan perkataan dan tulisan manusia dengan bahasa yang lebih formal, seperti notasi matematika atau logika, atau bahasa-bahasa komputer, seperti Java, LISP, dan C ++. Sebenarnya, *'Natural Language Understanding' (NLU)* itu dikaitkan dengan tujuan yang lebih ambisius dari memiliki sistem komputer yang

benar-benar memahami bahasa alami sebagai kekuatan manusia. Dalam proses text mining, NLP membantu dalam hal '*ranked' retrieval'* atau pengambilan berperingkat yang berarti bahwa pendataan atau pengindeksan dengan juga mengurutkan bobot sebuah kata (Jackson & Moulinier, 2002).

Secara umum ada dua cara untuk melakukan proses tersebut dua jenis metode, yaitu term frequency atau biasa dikenal dengan *Bag of Words (BoW)* dan *TF-IDF (Term Frequency-Inverse Document Frequency).*

A. Bag of Words

Model BoW adalah salah satu cara untuk mengekstraksi fitur-fitur dari sebuah teks untuk digunakan dalam pemodelan, seperti untuk algoritma pembelajaran mesin (machine learning).

Hanya istilah yang paling sering muncul yang akan dilihat. Sebelum menerapkan kriteria tersebut, kita perlu menghapus stopwords, atau kata-kata yang tidak memiliki substansi bagi isi dokumen, seperti kata sambung. *Term frequency* telah banyak digunakan digunakan sebagai kriteria seleksi global (Jackson & Moulinier, 2002).

BoW merupakan representasi teks yang menggambarkan kemunculan kata dalam dokumen. Hal ini melibatkan dua hal, yaitu kosakata kata-kata yang dikenal dan sebuah cara untuk mengukur kemunculan kata-kata yang dikenal (Brownlee, 2017)

BoW mencantumkan kata-kata yang kemudian dipasangkan dengan jumlah kata tersebut dalam sebuah teks atau kumpulan teks. Dalam tabel tersebut terdapat tempat kata dan dokumen yang secara efektif disimpan dalam bentuk vektor atau matriks (Nicholson, 2019)

Tabel 2.1. Contoh Penerapan *Bag of Words* (Zhou, 2019)

Commented [j8]: Revisi Tabel dan Citasi

Dokumen	the	cat	sat	in	hat	with
The cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1
Total	5	3	2	1	2	1

B. Metode TF-IDF

Metode TF-IDF (*Term Frequency-Inverse Documenty Frequency*) adalah ukuran statistik yang mengevaluasi seberapa relevan suatu kata yang terdapat di sebuah dokumen di dalam kumpulan beberapa dilakukan dokumen. Hal ini dilakukan dengan mengalikan dua metrik: berapa kali sebuah kata muncul dalam sebuah dokumen, dan frekuensi dokumen terbalik dari kata tersebut di seluruh kumpulan dokumen.TF-IDF merupakan pengembangan dari model BoW. Metode ini memiliki banyak kegunaan, terutama dalam menganalisis teks secara otomatis, dan juga untuk mendapatkan kata-kata dalam algoritma *machine learning* untuk *Natural Language Processing (NLP).*

Berbeda dengan BoW, setelah mendapatkan munculnya frekuensi setiap kata dalam kumpulan dokumen, maka frekuensi tersebut akan dinormalisasi terhadap banyaknya dokumen tempat kata tersebut muncul. (Bengfort dkk, 2018).

TF-IDF diciptakan untuk proses pencarian dokumen dan pengambilan informasi. Metode ini bekerja dengan meningkatkan secara proporsional skor akhir sebuah kata dengan berapa kali sebuah kata muncul dalam dokumen, tetapi juga diimbangi dengan jumlah dokumen yang mengandung kata tersebut. Jadi, kata-kata yang umum di setiap dokumen, seperti ini, apa, dan jika, akan memiliki peringkat yang rendah karena meskipun mereka mungkin muncul berkali-kali dalam sebuah dokumen, karena mereka memiliki arti yang penting bagi dokumen tersebut. (Stenacella, 2019).

Adapun rumus untuk menghitung skor TF-IDF adalah sebagai berikut:

Untuk kata i di dalam dokumen j

$$w_{i,j} = tf_{i,j}\ x\ \log(\frac{N}{df_i}) \quad (1)$$

$Tf_{i,j}$ = jumlah kemunculan kata i di dalam dokumen j

df_i = Total dokumen yang memiliki istilah i

N = total dokumen keseluruhan

Misalkan di sebuah dokumen yang berisikan 100 kata, terdapat kata kucing yang muncul 3 kali. *Term Frequency* untuk 'kucing' adalah (3/100) = 0,03. Sekarang, asumsikan kita memiliki 10 juta dokumen dan kata 'kucing' muncul dalam seribu dokumen ini. Kemudian, *Inverse Document Frequency* dihitung dengan log

(10.000.000 / 1.000) = 4. Dengan demikian, bobot Tf-idf adalah hasil dari jumlah ini: 0,03 * 4 = 0,12. (Manning dkk, 2008).

BAB 6
KESIMPULAN DAN SARAN

6.1. Kesimpulan

Dari hasil pengerjaan text mining, terdapat beberapa hal yang dapat disimpulkan adalah penelitian ini telah mengidentifikasi kata-kata kunci atau keyword yang sering muncul dalam ujian pendadaran yang diuji oleh setiap dosen dan mendapatkan gambaran arah *area of interest*-nya.

6.2. Saran

Saran untuk penelitian selanjutnya adalah:

1. Penelitian lebih lanjut dapat dilakukan dengan menggunakan tambahan data dari aplikasi ujian pendadaran yang baru, yaitu SIPETA.
2. Penelitian berikutnya dapat juga memasukkan BoK dari setiap skripsi yang telah melewati ujian pendadaran.
3. Perlu adanya modul *stopword* khusus yang dapat mengolah data lebih lanjut dan membuang kata-kata yang tidak relevan dalam menggambarkan *area of interest* setiap dosen.

DAFTAR PUSTAKA

Azam, N., & Yao, J.T. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 39 (2012), 4760–4768.

Bengfort, B. dkk. (2018). *Applied Text Analysis With Python*. Sebastopol: O'Reilly Media, Inc.

Besancon, R. & Rajman, M. (1998). Text Mining: Natural Language techniques and Text Mining applications. *Advances in Data Science and Classification*, 1997, 473-480.

Bhowik, R. (2008). Keyword Extraction from Abstracts and Titles, IEEE Southeastcon, 2008, 610-617.

Brownlee, J. (2017, September 17). *What Is Natural Language Processing?* Diakses pada tanggal 10 Mei 2020 dari https://machinelearningmastery.com/natural-language-processing/.

Brownlee, J. (2017, Oktober 9). *A Gentle Introduction to the Bag-of-Words Model*. Diakses pada tanggal 10 Mei 2020 dari https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Cuoto, Javier. (2019). *Text Mining: The Beginner Guide*. Diakses pada tanggal 5 Oktober 2019 dari https://monkeylearn.com/text-mining/.

Hariadi, M. dkk. (2009). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naive Bayes. Seminar Nasional Electrical, Informatics, and It's Educations, 2009, 71-74.

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Proceedings of the 2003 *Conference on Empirical Methods in Natural Language Processing,* 2003, 216-233.

Ignatow, G. & Mihalcea, G. (2018). *An Introduction to Text Mining Research Design Data Collection and Analysis*, Hundred Oaks: SAGE.

Jackson, P. & Moulinier, I. (2002). *Natural language processing for online applications: Text retrieval, extraction and categorization*. Amsterdam: John Benjamins Publishing Company.

Liddy, E.D. (2000). Text Mining. *Bulletin of the American Society for Information Science - October/November*, 2000, 13-14.

Madnani, N. (2007). *Getting started on natural language processing with Python.* XRDS: Crossroads, The ACM Magazine for Students, 13 (4), 5-15.

Manning, C. dkk. (2008). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Nicholson, C. (2019). *A Beginner's Guide to Bag of Words & TF-IDF*. Diakses pada tanggal 7 Juni 2020 dari https://pathmind.com/wiki/bagofwords-tf-idf.

Prilianti, K.R. & Wijaya, H, (2014). Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering. Jurnal Cybernatika, 2(1), 1-6.

PSTI UAJY. (2020). Buku Pedoman Pelaksanaan dan Penulisan Laporan Tugas Akhir. Yogyakarta: Program Studi Teknik Industri Universitas Atma Jaya Yogyakarta.

Santoso, H.A. & Wijaya, A.P. (2016). Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government. Journal of Applied Intelligent System, 1(1), 48-55

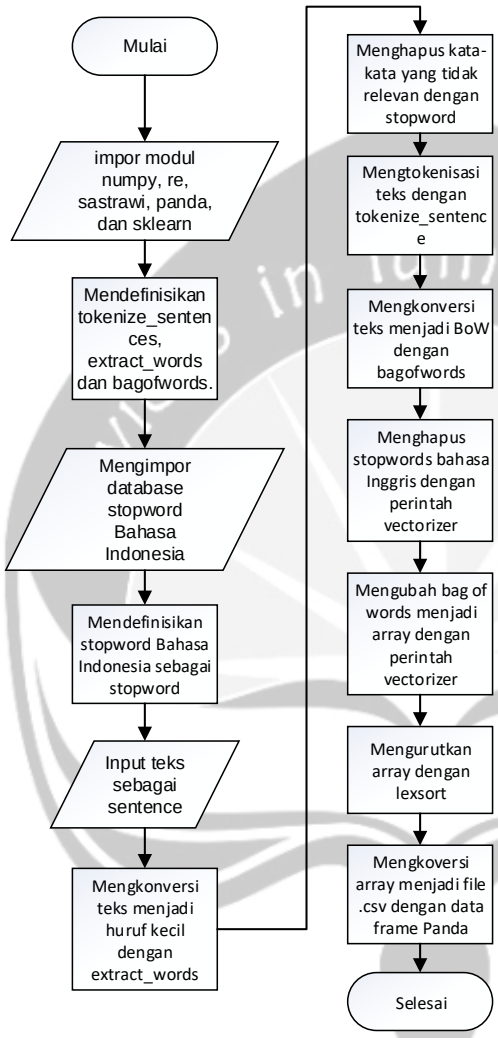
Stenacella, B. (2019). *What is TF-IDF?*. Diakses pada tanggal 21 September 2019 dari https://monkeylearn.com/what-is-tf-idf/.

Universitas Atma Jaya Yogyakarta. (2020). Teknik Industri – Universitas Atma Jaya Yogyakarta. Diakses pada 18 Juni 2020 dari http://fti.uajy.ac.id/industri/#:~:text=Program%20Studi%20Teknik%20Industri%20Universitas,sebagian%20besar%20bekerja%20pada%20sektor

Wiguna, I. (2011). Aplikasi Katalog Online untuk Pencarian Konten Buku dengan Metode Text Mining pada Perpustakaan STIKOM Surabaya. Penerbit Sekolah Tinggi Manajemen Informatika & Teknik Computer Surabaya

Zhou, V. (2019). *A Simple Explanation of the Bag-of-Words Model*. Diakses pada tanggal 28 Juli 2020 dari https://towardsdatascience.com/a-simple-explanation-of-the-bag-of-words-model-b88fc4f4971)

Lampiran



Lampiran 1. Diagram Alir Program *Text Mining*