p-ISSN : 2443-2210
e-ISSN : 2443-2229

Jurnal Teknik Informatika dan Sistem Informasi
Volume 6 Nomor 3 Desember 2020

Pemanfaatan *Latent Semantic Indexing* untuk Mengukur Potensi Kerjasama Jurnal Ilmiah Lintas Universitas

http://dx.doi.org/10.28932/jutisi.v6i3.2894

Edward Hanafi Fernando[#1], Hapnes Toba ✉[*2]

[#] *Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi*
Universitas Kristen Maranatha
Jl. Prof. drg. Surya Sumantri No.65 Bandung
[1]1879004@maranatha.ac.id
[2]hapnestoba@it.maranatha.edu

Abstract— **This paper presents a cooperation recommendation strategy between higher education institutions. The recommendation is based on the contents of journals published in a university journal portal. As a case study, we concentrate our approach for the journals with information technology themes. All journals from 10 reputed universities will be compared by using keywords and the contents of the journal themselves. A partnering recommendation list is built by utilizing Latent Semantic Indexing (LSI). LSI technique is used to reduce the curse of dimensionality from the original dataset and to generate topical analysis from all journals as a semantic representation for each journal. Topic modeling is used to calculate the categorical similarity in the dataset of each university journal and a search query. After all categorical similarities have been calculated, an average value of journal topics coherence is used to construct the final recommendation of partner candidates. This approach ensures that the final recommendation is based on the interest of each university rather than the frequencies of matched keywords in each journal.**

Keywords—**Document Similarity; Latent Semantic Indexing; Recommendation System; Topic Modeling; Web Crawling**

I. PENDAHULUAN

Menemukan mitra untuk bekerja sama dalam kerja sama jurnal ilmiah merupakan salah satu tantangan dalam dunia pendidikan tinggi. Salah satu pencarian mitra yang biasa dilakukan adalah mencari seorang *reviewer* (mitra bestari) atau dalam hal pertukaran makalah. Semakin banyaknya jumlah makalah yang ada memberikan sebuah tantangan dalam mendapatkan calon mitra yang cocok. Untuk mencapai tujuan tersebut pencarian mitra haruslah terukur dengan baik dan diupayakan sejalan dengan ciri khas dari sebuah universitas. Sebagai tolak ukur dalam menentukan mitra yang baik, profil jurnal ilmiah universitas dapat digunakan untuk dijadikan kecocokan.

Dewasa ini sebuah universitas memiliki halaman web portal jurnal yang di dalamnya terdapat informasi mengenai apa saja makalah yang dipublikasi pada universitas tersebut. Melalui informasi tersebut dapat diambil beberapa karakteristik yang dapat digunakan untuk menentukan kriteria mitra yang sepadan. Namun informasi yang ada di internet terkadang tersedia terlalu banyak dan tidak terstruktur sehingga dapat menyita waktu untuk membandingkan setiap informasi agar dapat menghasilkan sebuah keputusan yang baik.

Pada makalah ini diujicobakan data jurnal ilmiah yang terdapat pada portal berbagai universitas bereputasi di Indonesia yang tertera pada laman *web Science and Technology Index* (SINTA). Data jurnal yang digunakan adalah: judul, abstrak serta penulis makalah. Perbandingan yang dilakukan bukanlah perbandingan langsung terhadap kemunculan kata pada data yang sudah ditarik, melainkan data akan diolah untuk mengerucutkan topik yang sering dibahas pada penerbitan jurnal universitas tersebut. Untuk menghasilkan topik tersebut digunakan pendekatan *Latent Semantic Indexing* (LSI) dari data tekstual jurnal universitas yang sudah diekstrak. Dengan membandingkan kemunculan topik dalam setiap koleksi jurnal universitas, maka hasil rekomendasi yang dibangun dapat dianggap mencirikan bidang keilmuan yang menjadi kekhasan sebuah universitas atau fakultas.

Rangkaian percobaan di dalam makalah ini membahas bagaimana proses pengambilan data tekstual berupa dokumen jurnal yang dibutuhkan, untuk kemudian akan diproses agar dapat dijadikan sebuah rekomendasi. Rekomendasi dibentuk dengan mencari kemiripan antar dokumen dengan menggunakan LSI. Dengan LSI, kesamaan tidak akan dibentuk hanya berdasarkan pencocokan kata antara data dengan dengan kueri namun berdasarkan pengelompokkan kata yang muncul bersamaan (*word co-ocurrences*). LSI akan digunakan pula untuk membangun topik-topik dari data dokumen jurnal kemudian akan dibandingkan dengan kata kunci atau data dokumen jurnal yang diambil terpisah, dan dari sana rekomendasi kerja sama akan diambil.

p-ISSN : 2443-2210
e-ISSN : 2443-2229

II. KAJIAN LITERATUR

A. Web Scraping dan *Web Crawling*

Ekstraksi data akan dilakukan terhadap *web* portal jurnal universitas. Jumlah sumber data yang akan diekstrak relatif banyak, oleh karena itu proses ekstraksi data secara manual akan memakan waktu cukup banyak. Oleh karena itu Teknik *web scraping* dan *web crawling* akan digunakan untuk mempermudah dan memperpendek proses ekstraksi data. Pada makalah ini pustaka *Scrapy* untuk Python akan akan digunakan dalam melakukan proses *web scraping* dan *web crawling*.

Web scraping adalah teknik untuk mengekstrak informasi dari berbagai macam dokumen web secara otomatis [1]. Teknik *web scraping* adalah sebuah program pintar atau *web script* yang mengumpukan konten dari sebuah halaman *web*, dan kemudian menyimpanyan ke dalam format terstruktur di dalam sistem lokal untuk analisis lanjutan [1]. Umumnya, *web crawler* mengindikasikan kemampuan sebuah program untuk dapat menavigasi halaman-halaman *web* secara mandiri, bahkan tanpa tujuan yang didefinisikan dengan terperinci, tanpa batas penjelajahan yang ditawarkan sebuah situs [2].

Web crawler melintasi halaman-halaman *web* untuk diekstraksi oleh *web scrapin*g. Saat pengambilan data dari halaman *web*, cukup sulit untuk menentukan halaman *web* yang relevan. Dengan *web crawler*, pengunjungan halaman *web* dapat menggunakan *local search algorithm* yang disediakan oleh pustaka *Scrapy* untuk membatasi *Uniform Resource Locator* (URL) antara halaman *web* [3]. Contoh pemanfaatan mekanisme *local search algorithm* yang disediakan oleh pustaka *Scrapy* dalam proses *scraping* dapat dilihat dalam Algoritma 1.

```
class ProfileScrapper(CrawlSpider):
    name = 'maranatha_jurnal'
    allowed_domains = ['maranatha.edu']
    deny_domains = []
    start_urls = ['https://journal.maranatha.edu/i
ndex.php']
    base_url = 'https://journal.maranatha.edu/inde
x.php'
    rules = [Rule(LinkExtractor(
        allow=[],
        deny=[],
        deny domains = []),
        callback='parse_filter_book',
        follow=True)]
```

Algoritma 1. Pengaturan untuk membatasi proses *crawling* dalam *Scrapy*

Isi di dalam variabel *'name'* adalah nama dari proses yang akan dipanggil oleh pustaka Scrapy untuk menjalankan proses. Variabel '*allowed_domain'* adalah pengaturan yang dibutuhkan untuk menentukan domain yang akan di ambil datanya. Variabel '*start_urls'* adalah pengaturan yang diakukan untuk menentukan titik awal URL penarikan data. Variabel '*base_url*' adalah pengaturan yang dilakukan untuk menentukan pembatasan URL yang akan ditarik datanya. Dengan '*base_url*' setiap alamat yang memiliki nilai selain dari yang ditentukan akan diabaikan dalam proses penarikan data. Variabel '*callback*' adalah pengaturan untuk menentukan fungsi yang akan diambil. Pada kasus ini fungsi yang akan diambil adalah fungsi untuk melakukan penarikan data pada URL.

```
def parse_filter_book(self, response):
    #ambil nama url yang di crawl
     url = response.request.url
     #filter url yang akan di crawl
     if 'article' in url and 'pdf' not in url &&'/
         0' not in url:
         #proses crawl page
```

Algoritma 2. Fungsi untuk melakukan penarikan data dari halaman *web*

Pada kode di dalam Algoritma 2, dilakukan beberapa filtrasi pada URL. Sebelum penarikan dilakukan terhadap URL untuk melihat apakah pada URL tersebut mengandung kata kunci '*article*' dan 'pdf' dan '/0'. Apabila terdapat kata kunci yang sudah ditentukan pada URL, maka proses penarikan tidak akan dilakukan. Filtrasi ini berguna untuk membatasi URL yang memiliki pola tertentu, sehingga waktu proses penarikan data dapat menjadi lebih efisien karena langsung menuju pada target datanya.

B. Latent Semantic Indexing (LSI)

Dalam penelitian [4], LSI digunakan untuk memberikan rekomendasi mata kuliah yang akan diambil oleh siswa. Di dalam makalah tersebut, dikatakan bahwa penggunaan LSI menjadi salah satu bagian yang dapat memberikan hasil yang lebih masuk akal dibandingkan metode lainya. Pada makalah [5], LSI digunakan dalam menentukan sebuah rekomendasi bagi pengguna *Twitter* berdasarkan *posting* pengguna yang berhubungan dengan film. Pada makalah [6], LSI digunakan dalam perbandingan temu balik informasi terhadap *Vector Space Model*. Hasil dari makalah tersebut menyatakan bahwa LSI unggul dari sisi skalabilitas dan performa waktu untuk memproses data dalam jumlah besar, serta sangat berpotensi untuk menghasilkan model klasterisasi data tekstual.

Oleh karena itulah, dalam penelitian yang dieksplorasi pada makalah ini, LSI akan digunakan untuk membangun sebuah rekomendasi bagi kerja sama jurnal ilmiah antar universitas. Rekomendasi akan dibangun berdasarkan jurnal-jurnal yang ada pada *web* portal jurnal tiap universitas. Dengan memanfaatkan LSI, sebuah model topik akan dibangun guna menghasilkan himpunan topik. Dari topik-topik tersebut kemudian akan dihitung kedekatannya terhadap setiap jurnal. Kemudian, dari nilai kedekatan tersebut akan dibangun sebuah rekomendasi dengan menghitung nilai agregasi dari kemunculan topik.

LSI adalah metode temu balik informasi yang menyusun informasi menjadi struktur semantik dengan memanfaatkan asosiasi berdimensi tinggi dari kata-kata yang tersirat dengan sebuah objek di dalam sebuah *dataset*. Struktur yang dihasilkan mencerminkan pola asosiatif utama dalam data, sementara beberapa variasi yang lebih kecil diabaikan. Hal ini memungkinkan penarikan hasil berdasarkan makna tersembunyi daripada hanya pencocokan kata kunci. LSI menyaring informasi berdasarkan preferensi pengguna. Melalui pola kemunculan kata, LSI dapat menyimpulkan struktur hubungan artikel dengan kata-kata [7]. Pada pemrosesan LSI, dokumen direpresentasikan sebagai matriks [8], dengan baris merupakan sebuah kata dan kolomnya adalah dokumen atau unit yang lebih kecil berupa bagian-bagian dari dokumen.

Setiap elemen pada matriks LSI mengandung frekuensi kemunculan kata pada setiap bagian dokumen [8]. Matriks tersebut direduksi dengan menggunakan metode *Singular Value Decomposition* (SVD) untuk menyaring gangguan, terutama kata-kata dengan tingkat kemunculan rendah, yang ditemukan di dalam dokumen dan mempertahankan atribut yang paling relevan dari data yang sudah diberikan [9], [10]. Kedekatan antar dokumen kemudian dapat dihitung berdasarkan entitas dalam ruang dimensi yang sudah direduksi [11]. Sebagai contoh adalah proses mencari kesamaan antara 1 kueri (Q) dengan 3 dokumen (D1, D2, dan D3) berikut ini:

Q: "gold silver truck"
D1: "Shipment of gold damaged in a fire."
D2: "Delivery of silver arrived in a silver truck."
D3: "Shipment of gold arrived in a truck."

Proses LSI berawal dengan membangun sebuah matriks A berdasarkan Tabel I dengan indeks *i, j* dengan *i* adalah jumlah kemunculan kata pada dokumen *j* [10].

TABEL *I*
TABEL KATA – DOKUMEN

	D1	D2	D3
a	1	1	1
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
in	1	1	1
of	1	1	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

Dari matriks A kemudian dilakukan proses komputasi SVD, yang digunakan untuk melakukan dekomposisi terhadap kata pada matriks dokumen menjadi tiga matriks: T, sebuah matriks dengan dimensi sejumlah dokumen; S, sebuah matriks berisi nilai singular dengan dimensi sejumlah dokumen, dan D, juga sebuah matriks dengan dimensi sejumlah dokumen [12]. Matriks asli bisa didapatkan kembali dengan melakukan perkalian matriks TSD^T.

Pada LSI matriks T, S, D direduksi menjadi *k* dimensi [12], yang berukuran lebih kecil dari jumlah dokumen. Tujuan reduksi dimensi adalah untuk mengurangi gangguan pada ruang laten, yang menghasilkan struktur relasi kata yang lebih kaya yang menunjukan latensi semantik terhadap dokumen [12].

$$T = \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2626 & 0.3749 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3749 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix}$$

$$S = \begin{bmatrix} 4.0898 & 0.000 & 0.000 \\ 0.000 & 2.3616 & 0.000 \\ 0.000 & 0.000 & 1.273 \end{bmatrix}$$

$$D = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix}$$

$$D^T = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

LSI bergantung pada parameter *k* untuk mereduksi dimensi. Kueri dibandingkan terhadap vektor dokumen yang sudah direduksi, diukur dengan nilai singular melalui kesamaan kosinus. Kemudian dari matriks baru tersebut diambil nilai singular dengan *k* terbesar. Penentuan nilai *k* dapat ditentukan tanpa aturan tertentu. Misalnya ditentukan nilai $k = 2$. Dengan demikian, matriks A yang baru untuk Tabel I, akan menjadi $A_2 = T_2S_2D_2^T$

$$T_2 = \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3749 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3749 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 4.0898 & 0.000 \\ 0.000 & 2.3616 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} -0.4945 & 0.6492 \\ -0.6458 & -0.7194 \\ -0.5817 & 0.2469 \end{bmatrix}$$

$${D_2}^T = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

Dari matriks tersebut kemudian didapatkan koordinat vektor 2 dimensi. Baris dari D menyimpan nilai *eigenvector*.

Koordinat berikut ini adalah koordinat dari masing-masing dokumen pada Tabel I.

d1 = (-0.4945, 0.6492)
d2 = (-0.6458, -0.7194)
d3 = (-0.5817, 0.2469)

Untuk melakukan perhitungan kemiripan antara sebuah kueri dengan dokumen, kueri tersebut perlu ditransformasi untuk mendapatkan koordinat vektor pada ruang dimensi yang sudah direduksi. Misalnya untuk kueri '*gold silver truck'* pada Tabel I, nilai vektor kueri adalah sebagai berikut, dengan kalkulasi melalui formula (1).

$$q = q^T T_2 S_2^{-1} \quad \text{... (1)}$$

$$q^T = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1]$$

$$T_2 = \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3749 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3749 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix}$$

$$S_2^{-1} = \begin{bmatrix} \frac{1}{4.0898} & 0.000 \\ 0.000 & \frac{1}{2.3616} \end{bmatrix}$$

$$q = [-0.2140 \quad 0.1821]$$

Hasil kemiripan dapat digunakan untuk mengurutkan dokumen dengan mengaplikasikan *cosine similarity* antara kueri dengan dokumen dengan menggunakan formula (2).

$$sim(q, d) = \frac{q \cdot d}{|q||d|} \quad \text{... (2)}$$

Dengan perhitungan pada formula (2), maka didapatkan sim(q, d_1) = -0.0541, sim(q, d_2) = 0.9910, sim(q, d_3) = 0.4478. Dari hasil tersebut dapat dilihat bahwa dokumen nomor 2 memiliki tingkat kemiripan yang paling tinggi.

Pada makalah ini LSI digunakan untuk membangun rekomendasi bagi universitas dalam memilih mitra jurnal ilmiah. Model topik dengan LSI akan dilakukan untuk membangun topik-topik, dengan *topik coherence* sebagai acuan jumlah topik yang akan dibentuk.

Topic Modeling dengan LSI

Tujuan utama penggunaan LSI adalah membangun sebuah model topik yang dapat merepresentasikan sebuah dokumen. Model topik mempelajari kumpulan kata-kata dari korpora yang besar tanpa adanya supervisi. Berdasarkan pada kata-kata yang digunakan di dalam dokumen, model mengumpulkan relasi tingkatan topik dengan mengasumsikan bahwa sebuah dokumen mengandung himpunan kecil dari ringkasan topik. Setelah topik tersebut dipelajari, topik tersebut dapat berkorelasi dengan konsep yang dipahami manusia.

Dengan metode tanpa supervisi ini, informasi semantik yang berguna untuk berbagai kebutuhan bergantung pada mengidentifikasi topik atau konsep yang unik seperti distribusi semantik, induksi indra kata, dan pengambilan informasi dapat diekstrak [13].

Topic Coherence

Ketika menggunakan model topik, perlu diperhitungkan sejauh mana topik yang dipelajari cocok dengan penilaian manusia dan dapat membedakan antara ide-ide di dalamnya. Untuk itulah, perlu dipertimbangkan beberapa pertanyaan sebagai kunci yaitu [13]:

- Berapa banyak topik yang harus dipelajari?
- Berapa banyak topik yang dianggap berguna?
- Bagaimana topik tersebut saling berhubungan dengan tes semantik?
- Sebaik apa topik tersebut mengidentifikasi dokumen yang sejenis?

Pada proses LSI dapat ditentukan jumlah topik yang dihasilkan. Setiap kata yang memiliki nilai yang tinggi dianggap kata yang paling sering muncul. Kata kata yang memiliki probabilitas ini biasanya adalah top-10 atau 15 dan digunakan untuk menginterpretasikan dan secara semantik memberi label pada sebuah topik [14]. Pada perangkat *Gensim* dapat ditentukan jumlah topik yang akan dibuat ketika LSI berhasil dibuat.

Penentuan nilai jumlah topik yang akan dibentuk menjadi salah satu hal yang penting dalam membentuk sebuah model LSI. Jika nilai yang ditentukan terlalu kecil dikhawatirkan memberikan topik yang terlalu luas, sedangkan jika nilai yang ditentukan terlalu besar topik tidak dapat ditafsirkan dengan baik [14]. Proses pengukuran dari data untuk memberikan kualitas topik yang baik terkadang berkorelasi negatif terhadap interpretabilitas manusia, sehingga prediksi topik terkadang menjadi kurang koheren dari perspektif manusia [14]. Pembentukan topik ini merupakan hal penting ketika pembentukan topik dilakukan terhadap koleksi dokumen atau untuk mendapatkan sebuah tren pada sebuah penielitian. Untuk itu sebuah pengukuran koheren dibuat untuk dapat secara otomatis mengukur dan menampilkan ukuran koheren sebuah topik.

Sebuah topik dikatakan koheren apabila hampir semua kata, atau kata yang memiliki nilai tertinggi saling berhubungan. Untuk dapat mendapatkan hasil koherensi tersebut, pengukuran terhadap *topic coherence* dilakukan dengan eksplorasi secara sistematis dan empiris dan dalam jumlah yang besar terhadap urutan kepentingan topik yang diakukan oleh manusia [15]. Sebagai tambahan pengukuran koheren didapatkan dengan menggabungkan elemen dasar yang ada [14]. Pada *Gensim* dapat diukur nilai koherensi dari topik dengan menggunakan 3 jenis pengukuran, yaitu: C_V, UMass, UCI.

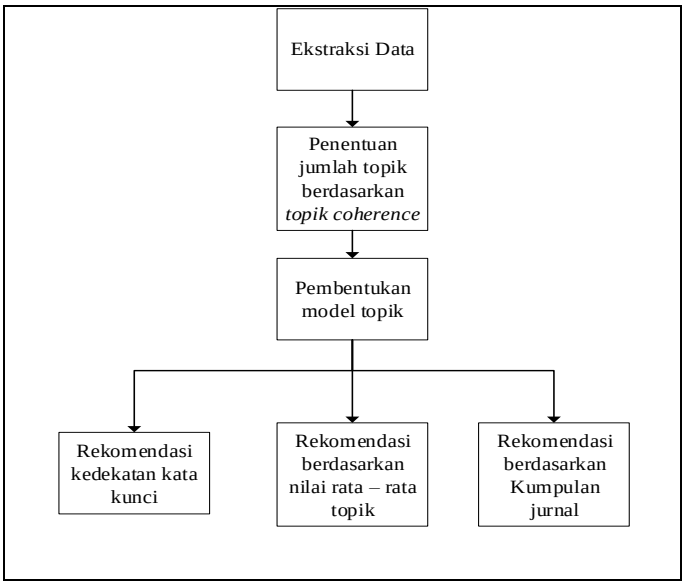
Pengukuran C_V berdasarkan pada 4 bagian, yaitu:

- Segmentasi data menjadi pasangan kata.
- Kalkulasi kata atau probabilitas pasangan kata.
- Kalkulasi dari konfirmasi ukuran yang menghitung sekuat apa sebuah set kata mendukung set kata lainya.
- Agregasi dari pengukuran konfirmasi individu menjadi nilai koheren secara menyeluruh.

Pengukuran UCI dan UMass mengkalkulasi koherensi sebuah topik sebagai jumlah dari nilai pasangan himpunan kata-kata yang membentuk suatu topik tertentu. Pengukuran UCI mendefinisikan nilai pasangan kata sebagai *Pointwise Mutual Information* (PMI) antara dua kata. Probabilitas kata dihitung dengan menghitng frekuensi kemunculan kata terhadap korpus eksternal [13]. Pada derajat tertentu, metrik ini dapat dianggap sebagai perbandingan eksternal terhadap evaluasi semantik yang sudah diketahui. Pengukuran UMass mendefinisikan nilai sebagai dasar pada kemunculuan kata pada sebuah dokumen. Perhitungan dilakukan terhadap korpus asli yang digunakan untuk membentuk model topik. Hal ini ditujukan untuk mengkonfirmasi bahwa sebuah model menggunakan data yang dikenali oleh *corpus* [13].

III. METODOLOGI

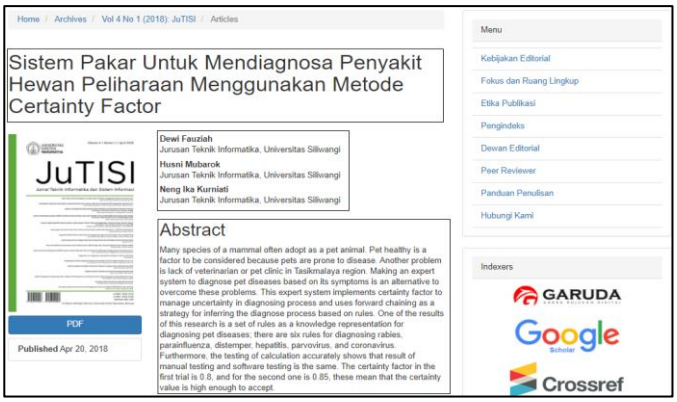
Dalam riset ini proses diawali dengan ekstraksi data dari portal jurnal untuk tiap universitas. Setelah data dikumpulkan, kemudian penentuan koherensi topik dilakukan untuk mendapatkan jumlah topik yang baik sebagai pembentuk model topik. Proses kemudian dilanjutkan dengan membentuk model topik. Dari model topik tersebut kemudian dibentuk rekomendasi dengan 3 cara. Gambar 1 menggambarkan garis besar dari proses yang dilakukan pada makalah ini.



Gambar 1. Proses pembentukan rekomendasi

Sumber data pada makalah ini diambil dari berbagai situs jurnal ilmiah dengan afiliasi dalam lingkup Fakultas Teknologi Informasi atau sejenisnya. Terdapat sepuluh universitas yang dipilih untuk disandingkan dengan Jurnal JuTISI yang diterbitkan oleh Fakultas Teknologi Informasi Universitas Kristen Maranatha. Sepuluh universitas dipilih berdasarkan urutan pada laman *web Science and Technology Index* (SINTA), yaitu: http://sinta.ristekbrin.go.id/. Lima universitas negeri dan lima universitas swasta, yang menerbitkan jurnal terkait bidang teknologi informasi, dipilih berdasarkan urutan peringkat teratas pada *web* tersebut.

Secara spesifik sumber data yang diambil adalah abstrak, judul, dan nama penulis dari makalah-makalah yang ada pada portal jurnal fakultas tersebut seperti yang ditunjukan pada Gambar 2. Data jurnal yang diambil tidak terpatok pada rentang tanggal tertentu, yang artinya data jurnal yang diambil adalah seluruh jurnal yang terdapat pada portal jurnal tiap universitas. Pemilihan data abstrak, judul didasari bahwa judul dan abstrak dianggap dapat mewakili isi dari makalah. *Keyword* pada abstrak ikut dalam proses ekstraksi dengan anggapan bahwa *keyword* merupakan bagian dari abstrak. Data penulis makalah dimasukan dalam proses ekstraksi data dengan tujuan agar pada hasil akhir akan didapatkan juga secara spesifik penulis mana yang dapat direkomendasikan sebagai reviewer. Data akan diambil dengan memanfaatkan pustaka *Scrapy*. Proses pengambilan data dilakukan secara otomatis untuk setiap jurnal portal.



Gambar 2. Contoh halaman jurnal dan target pengambilan data

Data yang diambil akan disimpan pada *file* teks terpisah berdasarkan universitas, nama penulis, dan judul dari jurnal. Setiap data akan disimpan pada folder terpisah berdasarkan universitas masing-masing jurnal. Format *file* saat penghimpunan data adalah: '[(nama penulis);(judul jurnal)].txt'. Data akan dibagi menjadi dua yaitu data yang akan dijadikan pembanding serta data yang akan menjadi target pembanding.

Data yang dijadikan pembanding pada makalah ini adalah data-data jurnal yang akan dijadikan pembanding terhadap data target. Data target pembanding merupakan kueri atau kata kunci, selain itu data target pembanding juga dapat berupa kumpulan jurnal dari satu universitas yang akan dicarikan rekomendasi kerja sama dengan universitas lainya.

A. Ekstraksi Data

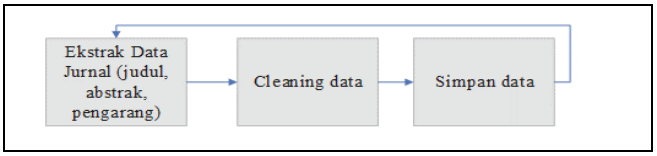
Himpunan data pada penelitian ini,diambil langsung dari portal berbagai jurnal universitas. Untuk pengambilan data, metode yang digunakan adalah ekstraksi data dengan menggunakan teknik *web scraping* dan *web crawling*. *Scrapy* adalah sebuah pustaka untuk melakukan *crawling*

p-ISSN : 2443-2210
e-ISSN : 2443-2229

pada situs *web* dan mengekstrak struktur data yang dapat digunakan secara luas, misalnya untuk *data mining*, pemrosesan informasi dan pengarsipan sejarah [16]. Dengan memanfaatkan *Scrapy*, maka data judul, abstrak, dan penulis jurnal dapat diekstrak dari halaman jurnal.

Data jurnal yang diambil dari *web* portal resmi sebuah universitas dianggap telah melewati saringan oleh universitas tersebut dan dianggap layak untuk dipublikasikan. Oleh karena itu setiap jurnal yang terdapat pada portal jurnal tersebut dianggap dapat mewakili topik-topik yang dianggap menarik pada universitas tersebut.

Pengumpulan data dilakukan terhadap portal jurnal ilmiah dari universitas: Universitas Indonesia (UI), Institu Teknologi Bandung (ITB), Universitas Gadjah Mada (UGM), Universitas Diponegoro (UNDIP), Universitas Pendidikan Indonesia (UPI), Universitas Telkom (Telkom), Universitas Bina Nusantara (BINUS), Universitas Pendidikan Ganesha (UNDIKSHA), Universitas Negeri Medan (Unimed), Universitas Katolik Widya Mandala Surabaya (UKWMS), dan Universitas Kristen Maranatha (UKM) sebagai pembanding. Sebagai batasan, data yang diambil terbatas pada judul jurnal, abstrak jurnal, dan penulis jurnal. Data diambil dengan metode ekstraksi dengan menggunakan Algoritma 3 terhadap halaman *web* jurnal. Gambar 3 menggambarkan alur dari proses ekstraksi data yang dilakukan dari halaman web portal jurnal universitas.



Gambar 3. Alur proses ekstraksi data

```
class ProfileScrapper(CrawlSpider):
    setting allowed_domains, start_urls,
    base urls, rule, callback
    def parse_filter_book(self, response):
        url = response.request.url
        if 'article' in url && 'pdf' not in url &
        & '/0' not in url:
      Ambil konten halaman berdasarkan xpath.
      Filter dan pre proses konten.
      Simpan dalam file dengan tipe txt
```

Algoritma 3. Algoritma untuk melakukan ekstraksi data jurnal dari sebuah web portal jurnal

Ekstraksi dimulai dari URL yang terdaftar di dalam parameter '*start_url*', kemudian dilanjutkan pada setiap *link* yang ada pada URL tersebut. Proses ini berulang hingga tidak ada lagi *link* yang sesuai dengan kriteria yang diberikan. Kriteria *link* yang paling utama yang digunakan adalah kriteria '*allowed_domain*' dan '*base_url*' dimana penarikan ekstraksi hanya dilakukan apabila *link* memiliki domain yang sesuai dengan '*allowed_domain*' dan memiliki '*base_url*' yang sudah ditentukan.

Selain itu filtrasi tambahan di lakukan pada fungsi '*parse_filter_book'*. Pada fungsi tersebut filtrasi dilakukan terhadap URL untuk menentukan pola yang diinginkan. Hal ini dilakukan agar mengurangi pemrosesan dari *link* yang didapatkan selama proses berlangsung. Kemudian, proses filtrasi dilakukan terhadap konten yang ditampilkan pada Gambar 2, meliputi antara lain:

- menghilangkan *tag html* yang mungkin ikut terbawa ketika proses *scraping*.
- Menghilangkan kata-kata yang tidak diinginkan atau dianggap tidak relevan.

Tabel II menunjukan detil mengenai hasil ekstraksi data yang sudah dilakukan.

TABEL II
DETIL HASIL EKSTRAKSI DATA

Nama universitas	Jumlah makalah jurnal	Rata-rata besar data (per *file* dalam satuan KB)
Universitas Bina Nusantara (BINUS)	568	3.537
Universitas Pendidikan Ganesha (UNDIKSHA)	323	3,931
Institut Teknologi Bandung (ITB)	290	3,896
Universitas Diponegoro (UNDIP)	268	2,881
Universitas Negeri Medan (UNIMED)	239	4
Universitas Kristen Marana-tha (UKM)	197	3,776
Universitas Gadjah Mada (UGM)	197	5,228
Universitas Indonesia (UI)	189	3,555
Telkom University (TELKOM)	184	3,978
Universitas Katolik Widya Mandala Surabaya (UKWMS)	175	4
Universitas Pendidikan Indonesia (UPI)	29	3,862

B. Data Preprocessing

Untuk mengurangi gangguan terhadap pemroses data dilakukan beberapa pra-pemrosesan data, yaitu:

- *Stemming* untuk mengembalikan kata kepada kata asalnya, contohnya perekenomian menjadi ekonomi, pertumbuhan menjadi tumbuh, membanggakan menjadi bangga
- Menghapus kata sambung seperti: yang, agar, supaya, jika, jikalau dan lain lain
- Merubah semua huruf besar menjadi huruf kecil
- Menghapus tanda baca
- Menghapus *stop words*

Berikut adalah contoh preproses data yang dilakukan pada makalah ini. Pada kumpulan teks: *"
 Perencanaan </br> 2020 Produksi dan Perbaikan Tata Letak di PT*

Berkat Anugrah Alam Cemerlang PT Berkat Anugrah Alam Cemerlang adalah sebuah industri pembuatan Air Minum dalam Kemasan dengan Merk Fikaro dan Puas. Dalam dunia industri, tata letak pabrik yang terencana dengan baik akan ikut menentukan efisiensi dan kesuksesan kerja. Tata letak pabrik yang ada di PT Berkat Anugrah Alam Cemerlang beberapa kali mengalami perubahan, sehingga terjadi pemisahan ruang produksi, pemindahaan gudang produk jadi, beberapa penataan luas areal kurang optimal, dan jarak pemindahan bahan menjadi lebih panjang. Oleh karena itu pengaturan kembali departemen-departemen perlu dilakukan untuk mengurangi biaya-biaya yang ditimbulkan. Untuk merancang tata letak pabrik pada penelitian ini menggunakan Systematic Layout Planning (SLP), di mana untuk penyusunan tata letak metode khusus yang digunakan adalah Algoritma CORELAP. Dari hasil perhitungan Algoritma CORELAP didapatkan dua alternatif tata letak yang kemudian dicari penggunaan energi listrik yang paling minimum. Untuk memberikan dukungan perencanaan tata letak pabrik, maka dalam penelitian ini juga disertai perhitungan biaya energi dan biaya investasi perpipaan", akan dilakukan pra-pemrosesan data agar dapat diterima pada proses LSI.

Hasil dari pra-pemrosesan kalimat tersebut adalah: *"['rencana', 'produksi', 'tata', 'letak', 'pt', 'berkat', 'anugrah', 'alam', 'cemerlang', 'pt', 'berkat', 'anugrah', 'alam', 'cemerlang', 'buah', 'industri', 'air', 'minum', 'kemas', 'merk', 'fikaro', 'puas', 'dunia', 'industri', 'tata', 'letak', 'pabrik', 'rencana', 'efisiensi', 'sukses', 'tata', 'letak', 'pabrik', 'pt', 'berkat', 'anugrah', 'alam', 'cemerlang', 'alami', 'ubah', 'pisah', 'ruang', 'produksi', 'pemindahaan', 'gudang', 'produk', 'tata', 'luas', 'areal', 'optimal', 'jarak', 'pindah', 'bahan', 'atur', 'departemen', 'biaya', 'timbul', 'rancang', 'tata', 'letak', 'pabrik', 'teliti', 'systematic', 'layout', 'planning', 'slp', 'susun', 'tata', 'letak', 'metode', 'algoritma', 'corelap', 'darihasil', 'hitung', 'algoritma', 'corelap', 'alternatif', 'tata', 'letak', 'cari', 'energi', 'listrik', 'minimum', 'dukung', 'rencana', 'tata', 'letak', 'pabrik', 'teliti', 'hitung', 'biaya', 'energi', 'biaya', 'investasi', 'pipa']"*. Dapat dilihat bahwa kalimat dirubah secara keseluruhan menjadi token-token kata dan setiap kata menjadi huruf kecil dan dikembalikan kepada bentuk kata dasarnya.

C. Pemodelan Topik

Setelah pra-pemrosesan data selesai dilakukan, kemudian data tersebut akan diolah untuk menghasilkan model topik. Untuk memproses data tersebut digunakan metode LSI. Sebelum dapat membentuk sebuah model topik, dokumen-dokumen diubah menjadi vektor agar dapat diproses secara matematis. Proses pemodelan topik melibatkan dua tahap, yaitu: pembentukan *dictionary & corpus bow*, dilanjutkan dengan pembentukan model topik.

Pembentukan Dictionary & Corpus BOW

Corpus berisi koleksi dokumen yang memiliki dua tujuan, yaitu: sebagai input untuk *training* model, dan untuk mengekstrak topik setelah model topik selesai dibuat [17]. *Corpus* kemudian diubah ke dalam bentuk *dictionary* yang berisi setiap kata yang ada pada *corpus*. *Corpus* yang sudah dibentuk ke dalam bentuk dictionary tersebut kemudian dibentuk menjadi model *bag-of-words* (*bow*) [17], dan setiap kata dalam *dictionary* akan memiliki kode identitas unik. *Dictionary* adalah kumpulan kata-kata yang ada pada *dataset* dan sudah diberikan identitas unik, sedangkan *corpus bow* adalah sebuah *corpus* data yang bentuknya sudah ditransformasi menjadi jumlah kemunculan kata pada masing-masing dokumen. Untuk mendapatkan *dictionary* dan *corpus bow* tersebut, pada *dataset* yang sudah dikumpulkan akan dilakukan pra-pemrosesan, yaitu membentuk *dataset* menjadi sebuah *dictionary*, dan dari *dictionary* tersebut kemudian *corpus bow* akan dibentuk. Berikut adalah contoh potongan dari *dictionary* dan *bow*:

'aafam': 7939,
'aashto': 8131,
'abad': 5774,
'abadi': 3605,
'abai': 8835,
'abandon': 9134,
'abaqus': 6038,
'abaya': 6413,
'abbreviated': 8196,
'abbreviation': 15491,
'abcd': 12655,
'abdi': 14509,
'abduction': 12906,
'abductive': 12907,
'abfi': 2663,
'ability': 2313,
'abilitydengan': 13643,
'abilityrendah': 13644,
'abilityterhadap': 13645

[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1)],
[(2, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1)],

Pembentukan Model Topik

Dalam pembentukan model topik, perlu ditentukan berapa jumlah topik yang harus dibangun agar topik-topik tersebut optimal. Hal ini dimaksudkan agar setiap topik yang dibentuk adalah topik unik dan tidak terjadi redundansi pada setiap topik yang dibuat. Untuk pengukuran *topic coherence*, pada makalah ini digunakan pengukuran metrik UCI, metrik CV dan metrik UMASS. Hasil pengukuran akan bersifat sebagai rekomendasi. Pada prosesnya apabila dirasa diperlukan maka jumlah optimal akan disesuaikan untuk bisa mendapatkan nilai rekomendasi yang diperlukan.

```
def compute_coherence_values(dictionary, corpus,
texts, stop, start=2, step=3):
    for num_topics in range(start, stop, step):
  Bentuk model lsi
  Hitung koherensi topik model
return model_list, coherence_values
```

Algoritma 4. Algoritma untuk perhitungan *topic coherence*

Setelah mendapatkan rekomendasi *topic coherence* sesuai Algoritma 4, maka model topik kemudian akan dibentuk dengan jumlah topik yang sudah didapat dari proses perhitungan *topic coherence* tersebut. Kemudian model LSI akan diaplikasikan terhadap *corpus* untuk mendapatkan nilai kedekatan topik terhadap setiap dokumen jurnal pada *dataset*.

D. Perhitungan similarity untuk rekomendasi

Nilai *similarity* dapat sangat bervariasi antara satu universitas dengan lainnya, maka diperlukan adanya perhitungan nilai rata-rata. Dengan menggunakan metode agregasi, maka dari berbagai kata kunci pada hasil metode pertama dan kedua, akan didapatkan nilai akhir yang mewakili kedekatan secara komprehensif. Proses perhitungan untuk menghasilkan rekomendasi dilakukan dengan tiga cara.

Cara pertama, perhitungan kemiripan dilakukan terhadap dokumen jurnal pada *dataset* dengan kata kunci kueri. Proses perhitungan kemiripan ini didasarkan pada kedekatan antara kueri pada setiap jurnal yang ada pada dataset. Proses perhitungan diawali dengan mengubah bentuk kueri yang sudah diberikan terlebih dahulu (*pre-defined*) menjadi *dictionary* dan *bow*, kemudian ditransformasi menjadi vektor LSI dengan cara yang sama dengan yang dilakukan pada *dataset* jurnal. Setelah itu, antara kueri dan *dataset* dihitunglah similaritasnya. Universitas yang memiliki nilai tertinggi dari perhitungan tersebut dianggap cocok untuk dijadikan mitra berdasarkan kueri yang diberikan. Sebagai contoh, apabila diinginkan untuk mencari mitra universitas yang memiliki minat pada topik *machine learning*, maka kueri '*machine learning*' akan dimasukan. Kemudian kueri tersebut ditransformasi dengan model topik yang sudah dibuat dan hasilnya dibandingkan terhadap seluruh *dataset* jurnal yang sudah diambil dari masing-masing universitas. Hasil yang memiliki kemiripan terhadap kueri tersebut dianggap yang paling cocok untuk untuk dijadikan mitra dalam hal yang berhubungan dengan machine learning.

Cara kedua, perhitungan kemiripan dilakukan dengan melihat pada nilai rata-rata topik di setiap universitas. Setelah topik berhasil dibentuk, topik kemudian akan diaplikasikan terhadap setiap jurnal yang ada pada *dataset* sehingga akan didapatkan seberapa besar nilai setiap topik terhadap setiap jurnal. Setelah itu setiap jurnal akan di ambil rata-ratanya berdasarkan universitas. Dari nilai tersebut kemudian akan diambil rekomendasi berdasarkan nilai rata-rata topik.

Cara ketiga, proses perhitungan dilakukan dengan membandingkan kumpulan jurnal sebuah universitas terhadap *corpus*. Proses ini hampir sama dengan proses perhitungan terhadap kueri kata kunci, bedanya kata kunci diganti dengan kumpulan jurnal pada universitas sebagai kueri.

IV. EVALUASI

Hasil akhir dari proses LSI adalah reduksi kolom kata yang jumlah reduksinya ditentukan sejak awal proses berlangsung. Jumlah kolom yang direduksi ini disebut sebagai topik. Topik ini adalah sekumpulan kata-kata yang memiliki bobot nilai yang menentukan seberapa penting kata tersebut pada sebuah topik. Melalui kata-kata dalam topik tersebut dapat ditelusuri minat utama pembahasan dalam sebuah jurnal. Berikut adalah contoh dari dua kelompok topik LSI:

[(0, '0.565"ajar" + 0.262*"medium" + 0.250*"learning" + 0.221*"siswa" + 0.169*"interaktif" + 0.141*"hasil" + 0.141*"kembang" + 0.123*"student" + 0.115*"multimedia" + 0.113*"teliti"'),*
(1, '-0.344"ajar" + 0.219*"information" + 0.182*"company" + 0.147*"application" + -0.137*"siswa" + 0.132*"business" + 0.128*"method" + 0.123*"data" + 0.120*"process" + 0.119*"service"')]*

Pada contoh topik '0' yang dimunculkan di atas, dapat ditelaah bahwa topik tersebut mengenai pengajaran mengenai multimedia yang interaktif, atau mengenai perkembangan siswa terhadap pengajaran melalui multimedia yang interaktif. Sedangkan pada topik '1', secara intuisi dapat ditelaah bahwa topik tersebut adalah mengenai pelajaran bagi siswa mengenai aplikasi proses bisnis pada sebuah perusahaan.

Topic coherence merupakan salah satu hal yang penting dalam pembentukan model LSI. *Topic coherence* dapat digunakan untuk menentukan berapa jumlah topik yang disarankan. Di dalam LSI, hal ini adalah nilai dimensi *k*. Topik di sini adalah kumpulan dari kata-kata yang memiliki nilai yang tinggi pada proses pembentuk LSI. Jumlah topik yang terlalu kecil memberi kemungkinan bahwa topik yang dihasilkan tidak merepresentasikan karakteristik jurnal pada *dataset*. Sedangkan jumlah topik yang terlalu besar dapat menimbulkan redundansi antar topik sehingga mungkin ada topik yang sama dan jumlah topik yang dihasilkan dapat mempengaruhi hasil perhitungan similaritas yang dihasilkan.

Sebagai contoh, misalkan jumlah kelompok topik divariasikan untuk kumpulan dokumen di bawah ini dengan nilai jumlah kelompok topik 2, 4, dan 6:

[(0, '0.565"ajar" + 0.262*"medium" + 0.250*"learning" + 0.221*"siswa" + 0.169*"interaktif" + 0.141*"hasil" + 0.141*"kembang" + 0.123*"student" + 0.115*"multimedia" + 0.113*"teliti"'),*
(1, '-0.344"ajar" + 0.219*"information" + 0.182*"company" + 0.147*"application" + -0.137*"siswa" + 0.132*"business" + 0.128*"method" + 0.123*"data" + 0.120*"process" + 0.119*"service"')]*

Topik-topik di atas adalah hasil pembentukan topik dengan 2 kelompok topik.

[(0, '0.565"ajar" + 0.262*"medium" + 0.250*"learning" + 0.221*"siswa" + 0.169*"interaktif" + 0.141*"hasil" + 0.141*"kembang" + 0.123*"student" + 0.115*"multimedia" + 0.113*"teliti"'),*
(1, '0.343"ajar" + -0.219*"information" + -0.182*"company" + -0.146*"application" + 0.137*"siswa"*

+ -0.132"business" + -0.128*"method" + -0.123*"data" + -0.120*"process" + -0.119*"service"'),*
(2, '0.217"sistem" + -0.177*"learning" + 0.160*"nilai" + 0.154*"metode" + -0.140*"ajar" + 0.132*"prose" + 0.130*"informasi" + 0.123*"teliti" + 0.122*"aplikasi" + 0.117*"hasil"'),*
(3, '0.314"image" + -0.277*"company" + -0.211*"information" + -0.203*"business" + 0.161*"algorithm" + 0.155*"feature" + 0.136*"accuracy" + 0.126*"learning" + 0.121*"classification" + -0.119*"management"')]*

Topik-topik di atas adalah hasil pembentukan topik dengan 4 kelompok topik.

[(0, '0.565"ajar" + 0.262*"medium" + 0.250*"learning" + 0.221*"siswa" + 0.169*"interaktif" + 0.141*"hasil" + 0.141*"kembang" + 0.123*"student" + 0.115*"multimedia" + 0.113*"teliti"'),*
(1, '-0.344"ajar" + 0.219*"information" + 0.181*"company" + 0.147*"application" + -0.137*"siswa" + 0.132*"business" + 0.128*"method" + 0.123*"data" + 0.120*"process" + 0.119*"service"'),*
(2, '0.217"sistem" + -0.177*"learning" + 0.161*"nilai" + 0.155*"metode" + -0.141*"ajar" + 0.131*"prose" + 0.130*"informasi" + 0.123*"teliti" + 0.121*"aplikasi" + 0.117*"hasil"'),*
(3, '0.314"image" + -0.274*"company" + -0.210*"information" + -0.202*"business" + 0.161*"algorithm" + 0.156*"feature" + 0.136*"accuracy" + 0.126*"learning" + 0.121*"classification" + 0.120*"network"'),*
(4, '0.515"learning" + 0.312*"student" + -0.254*"ajar" + -0.204*"image" + -0.203*"company" + -0.159*"siswa" + 0.141*"game" + 0.134*"robot" + -0.126*"business" + 0.096*"university"'),*
(5, '-0.559"robot" + -0.236*"game" + 0.211*"learning" + -0.163*"mobile" + -0.128*"sensor" + 0.127*"student" + -0.127*"application" + 0.110*"nilai" + -0.108*"image" + -0.104*"android"')]*

Topik-topik di atas adalah hasil pembentukan topik dengan 6 kelompok topik. Pada setiap pembentukan topik, dengan jumlah kelompok yang berbeda, dapat terlihat urutan kata yang berbeda pula.

V. DISKUSI HASIL

A. Persiapan Data

Pada tahap awal penelitian ini, dilakukan ekstraksi data dari halaman web portal jurnal universitas. *Scrapy* akan digunakan untuk secara otomatis menarik data yang sudah dijadikan target di setiap halaman makalah dan secara otomatis berpindah ke halaman lainya setelah 1 halaman selesai diekstraksi. Setelah proses ekstraksi data selesai, disusunlah *corpus* dokumen dari data tersebut. Proses pembentukan *corpus* dilakukan secara otomatis. Setiap dokumen di-*load* satu per satu dan disimpan ke dalam *dataframe*. Tabel III menunjukan contoh hasil *load* dokumen jurnal. Semua data dalam tabel data tersebut kemudian dibentuk menjadi *corpus* dokumen. Proses pembentukan *corpus* melibatkan proses tokenisasi pada setiap isi *file*, pembentukan *dictionary* dari data, dan pembentukan *bag-of-words* (*bow*) berdasarkan data dan *dictionary*. Gambar 4, 5, dan 6 merupakan contoh hasil dari proses tersebut.

```
0  [deteksi, tanda, vital, pasien, rumah, sakit, ...
1  [klasifikasi, motif, batik, ciri, wavelet, tra...
2  [seleksi, citra, dasar, ciri, algoritma, thres...
3  [application, mean, algorithm, index, indonesi...
4  [playing, game, feasible, greedy, strategy, re...
5  [analisis, seleksi, citra, manfaat, konsep, cb...
6  [teknik, ambil, putus, multi, kriteria, metode...
7  [implementation, reconfiguration, robot, opera...
8  [capai, usaha, customer, relationship, managem...
9  [study, factor, affecting, software, applicati...
```

Gambar 4. Hasil tokenizing dokumen

```
'access': 666,
'accessable': 987,
'accessed': 831,
'accessfood': 11328,
'accessibility': 745,
'accessible': 3054,
'accessing': 4551,
'accessory': 3448,
'accident': 6260,
'accidentally': 9178,
```

Gambar 5. Hasil Pembentukan dictionary dokumen

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1)]
[(2, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1)]
```

Gambar 6. Hasil pembentukan bow dari dokumen

Setelah *corpus*, *dictionary*, dan *bow* terbentuk, dilakukanlah pengukuran *topic coherence* yang akan dibuat dengan memanfaatkan metode LSI. Pengukuran akan dilakukan dengan 3 metode pengukuran yaitu c_uci, c_v dan u_mass. Proses dilakukan dengan pertama-tama membentuk model topik kemudian di ukur dengan menggunakan fungsi '*CoherenceModel'* yang disediakan oleh pustaka *Gensim*.

Proses tersebut akan diulang dengan parameter jumlah topik yang berbeda di setiap perulangan, dan kemudian hasil akan ditampilkan dalam bentuk grafik bergaris. Setelah mendapatkan jumlah topik yang optimal, kemudian ditentukanlah kueri yang akan menjadi dasar untuk menghitung kesamaan antar universitas berdasarkan dokumen jurnal. Penentuan kueri terbagi menjadi dua yaitu: sebuah kata kunci, atau berdasarkan kumpulan jurnal dari sebuah universitas.

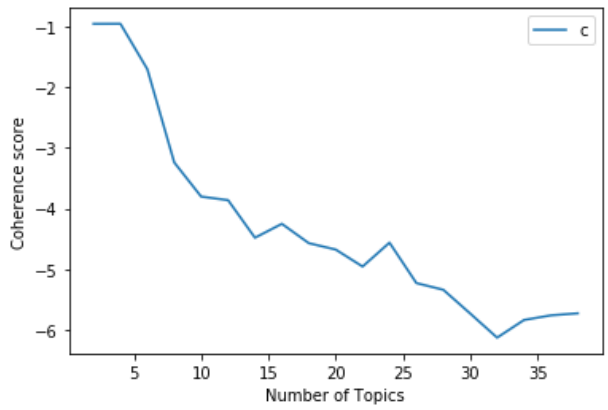
Untuk perhitungan *similarity* digunakan kelas '*Matrix Simmilarity*' yang disediakan oleh *Gensim*. Hasil dari perhitungan akan menunjukan kedekatan kueri terhadap setiap dokumen jurnal per universitas. Universitas dengan nilai rata-rata tertinggi dianggap memiliki kedekatan (*similarity*) yang paling dekat dengan kueri yang sudah disiapkan, dan akan dijadikan rekomendasi.

TABEL III
HASIL LOAD DOKUMEN

	University	Author	Journal Title	File Path
0	BINUS	A Haris Rangkuti,	Deteksi 4 Tanda Vital Pasien Rumah Sakit Berb...	D:\Documents\Journal\Informatika\BINUS\A Haris...
1	BINUS	A Haris Rangkuti,	Klasifikasi Motif Batik Berbasis Kemiripan Cir...	D:\Documents\Journal\Informatika\BINUS\A Haris...
2	BINUS	A. Haris Rangkuti,	Seleksi Citra Berdasarkan Ciri dengan Algoritm...	D:\Documents\Journal\Informatika\BINUS\A. Hari...
3	BINUS	A. Raharto Condrobimo, Albert V. Dian Sano, He...	The Application Of K-Means Algorithm For LQ45 ...	D:\Documents\Journal\Informatika\BINUS\A. Raha...
4	BINUS	Abas Setiawan,	Playing the SOS Game Using Feasible Greedy Str...	D:\Documents\Journal\Informatika\BINUS\Abas Se...
...	...	...	...	...
2473	WIDYA_MANDALA	Yosephat Suryo Susilo, Hartono Pranjoto, Alber...	Sistem Pelacakan dan Pengamanan Kendaraan Berb...	D:\Documents\Journal\Informatika\WIDYA_MANDALA...
2474	WIDYA_MANDALA	Youngky Siswanto, Evy Suryaningsih, Nani Indra...	Pengaruh Suhu Pemasakan dan Laju Penambahan Ai...	D:\Documents\Journal\Informatika\WIDYA_MANDALA...
2475	WIDYA_MANDALA	Yudo Herman Cahyo Adi, Hendro Gunawan,	Video Mixer Yang Dapat Diprogram	D:\Documents\Journal\Informatika\WIDYA_MANDALA...
2476	WIDYA_MANDALA	Yuli Prastiawati, Diana Lestariningsih, Andrew...	Mesin penggiling bumbu pecel otomatis berbasis...	D:\Documents\Journal\Informatika\WIDYA_MANDALA...
2477	WIDYA_MANDALA	Yusup Tanudjaja, Hadi Santosa, Julius Mulyono,	Perancangan Alat Bantu Peletakan Sheet dengan ...	D:\Documents\Journal\Informatika\WIDYA_MANDALA...

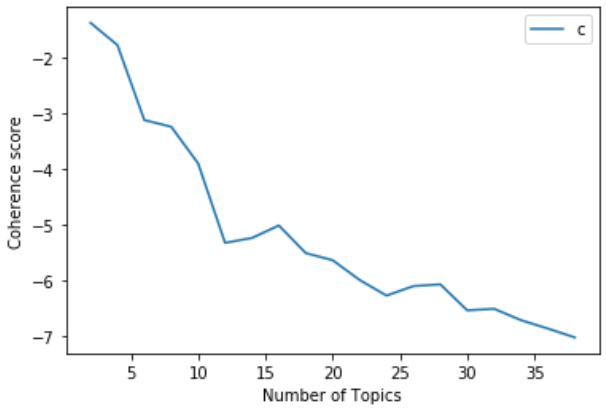
B. Topic Coherence

Setelah *dictionary* dan *corpus* dibentuk dan sebelum model topik dibuat perlu ditentukan berapa banyak topik yang akan dibentuk untuk mencegah topik yang terlalu sempit atau terlalu luas. Untuk penentuan topik maka proses pengukuran *topic coherence* dilakukan. Untuk itu metode UCI, UMass, dan C_V akan digunakan. Proses perhitungan akan dilakukan dengan menggunakan pustaka *Gensim*

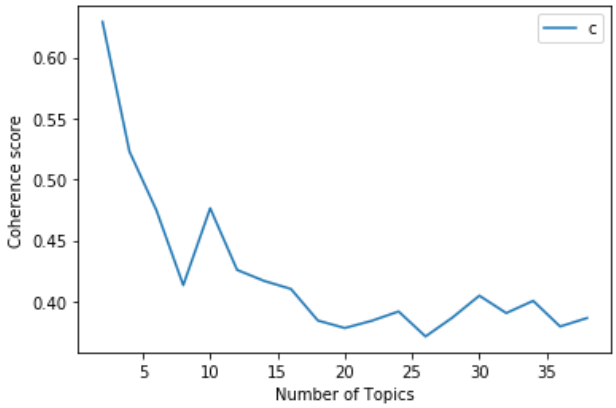


Gambar 7. *Grafik topic coherence* UCI

Pada Gambar 7, 8, dan 9 dapat dilihat bahwa jumlah terbaik ada pada kisaran 2 topik. Oleh karena itu dalam pembentukan topik akan ditentukan bahwa jumlah topik yang akan dibentuk adalah 2.



Gambar 8. Grafik *topic coherence* UMass



Gambar 9. Grafik *topic coherence* C_V

C. Pembentukan topik model

Topik model kemudian dibentuk dengan menggunakan metode LSI, dan menggunakan pustaka *Gensim* untuk Python. Setelah model LSI sudah dibentuk dapat dimunculkan kata-kata pada peringkat 10 teratas yang mewakili setiap kelompok topik.

```
[(0,'0.565*"ajar" + 0.262*"medium" +
0.250*"learning" + 0.221*"siswa" +
0.169*"interaktif" + 0.141*"kembang" +
0.141*"hasil" + 0.123*"student" +
0.115*"multimedia" + 0.112*"teliti"'),
 (1,'-0.343*"ajar" + 0.219*"information" +
0.182*"company" + 0.146*"application" + -
0.137*"siswa" + 0.132*"business" + 0.128*"method"
+ 0.122*"data" + 0.120*"process" +
0.119*"service"')]
```

Gambar 10. Contoh hasil pembentukan model topik

D. Proses Pembentukan Rekomendasi

Untuk dapat membantu menentukan calon mitra jurnal ilmiah pada makalah ini akan dilakukan percobaan pembentukan rekomendasi yang dilakukan berdasarkan kedekatan sebuah topik terhadap jurnal. Percobaan proses pembentukan rekomendasi pada makalah ini dilakukan dengan tiga cara yaitu rekomendasi dengan kata kunci, rekomendasi berdasarkan hubungan topik terhadap jurnal, rekomendasi berdasarkan kumpulan jurnal universitas.

Rekomendasi dengan kata kunci

Rekomendasi kata kunci adalah proses yang melihat kedekatan terhadap kata kunci yang diberikan. Kata kunci yang diberikan dapat berupa kata apapun dan akan disebut kueri kata kunci. Kueri kemudian ditransformasi menjadi korpus, kemudian model LSI diaplikasikan terhadap kueri tersebut. Setelah itu rekomendasi kemudian dihitung antara kueri kata kunci dengan setiap jurnal. Kemudian setelah diperoleh nilai kedekatan untuk setiap jurnal, dihitunglah nilai rata-rata berdasarkan universitas. Dari nilai rata-rata tersebut akan dibentuk rekomendasi berdasarkan nilai tertinggi. Pada makalah ini kata kunci yang ditentukan untuk uji coba adalah: "teknologi informasi", "*information technology*", "data", "jaringan", dan "*artificial intelligence*".

Tabel IV, V, VI, VII, VIII merupakan hasil perhitungan kemiripan antara jurnal dan kata kunci, kemudian diambil rata-rata berdasarkan universitas. Pada Tabel IV dan V terlihat perbedaan hasil kemiripan antara hasil kueri "teknologi informasi" dan "*information technology*" walaupun secara arti sama.

Dari contoh ini dapat terlihat bahwa perbedaan kata walau memiliki makna yang sama dapat mempengaruhi hasil proses LSI. Meskipun demikian, dari hal ini dapat dilihat pula walau terdapat perbedaan bahasa, LSI tetap dapat memunculkan topik walaupun bahasa tercampur. Hal ini dimungkinkan karena setiap *term* (kata) diubah menjadi komponen dalam vektor dan kata akan diproses berdasarkan kemunculan kata tersebut secara bersamaan dengan kata-kata lainnya (*word co-occurrences*). Sehingga meskipun topiknya berbeda, namun kata-kata di dalam topik berpotensi sama. Hal ini menjadi salah satu keunggulan yang diharapkan melalui proses LSI.

TABEL IV
KESAMAAN KATA KUNCI "TEKNOLOGI INFORMASI"

Universitas	Similaritas
UKWMS	0.758805
UNDIKSHA	0.657663
UPI	0.602762
UNIMED	0.455957
UNDIP	0.421972
TELKOM	0.328896
BINUS	0.203004
UKM	0.193805
UGM	0.161158
UI	0.109022
ITB	-0.145228

TABEL V
KESAMAAN KATA KUNCI "*INFORMATION TECHNOLOGY*"

Universitas	Similaritas
BINUS	0.635494
UKM	0.524530
ITB	0.329039
UGM	0.231738
TELKOM	0.185355
UI	0.178089
UNDIP	0.112671
UNDIKSHA	0.047857
UPI	0.044735
UNIMED	0.036253
UKWMS	0.008774

TABEL VI
KESAMAAN KATA KUNCI "DATA"

Universitas	Similaritas
UGM	0.693101
ITB	0.644369
TELKOM	0.614619
UNDIP	0.606476

UKM	0.583101
BINUS	0.554007
UI	0.540212
UPI	0.510424
UKWMS	0.490704
UNDIKSHA	0.460973
UNIMED	0.372912

TABEL VII
KESAMAAN KATA KUNCI JARINGAN

Universitas	**Similaritas**
UKWMS	0.833250
UNDIKSHA	0.696402
UPI	0.691545
UNDIP	0.617695
TELKOM	0.510143
UNIMED	0.465833
UGM	0.425123
UI	0.356596
UKM	0.215865
ITB	0.157982
BINUS	0.116051

TABEL VIII
KESAMAAN KATA KUNCI "*ARTIFICIAL INTELIGENCE*"

Universitas	**Similaritas**
ITB	0.669722
UGM	0.581022
UI	0.482508
TELKOM	0.414425
UNDIP	0.392505
UKM	0.297074
BINUS	0.186082
UPI	0.186010
UKWMS	0.102624
UNIMED	0.095344
UNDIKSHA	0.094186

Rekomendasi berdasarkan hubungan topik terhadap jurnal

Rekomendasi berdasarkan hubungan topik terhadap jurnal adalah rekomendasi yang diambil berdasarkan kedekatan topik terhadap masing-masing jurnal. Rekomendasi dilakukan dengan melihat hubungan topik terhadap jurnal itu sendiri. Misalnya, kelompok topik 1 dan 2 pada Gambar 10 diaplikasikan terhadap masing-masing jurnal. Dari sana akan dapat dilihat seberapa dekat sebuah topik terhadap jurnal. Tabel IX adalah hasil aplikasi topik terhadap masing-masing jurnal. Nilai yang diberikan menandakan seberapa dekat setiap topik terhadap masing-masing makalah dalam jurnal. Kemudian diambil nilai rata-rata dari makalah-makalah tersebut berdasarkan universitas masing-masing dengan cara yang ditampilkan pada Kode Program 1.

```
simGroup = df.groupby('University',
sort=True)['Topic 1'].mean().reset_index()
simGroup.sort_values('Topic 1', ascending=False)
```

Kode Program 1. Kode Python untuk mengambil nilai rata rata kedekatan topik

Dari nilai rata-rata tersebut dapat diamati kedekatan antar universitas yang cocok untuk dijadikan mitra. Pada Tabel X dapat dilihat kedekatan kelompok Topik 1 sebagai topik-topik yang paling sering dibahas di dalam jurnal-jurnal terbitan UNIMED. Dan pada Tabel XI dapat dilihat nilai kedekatan Topik 2 ada pada UGM sehingga bisa diambil kesimpulan Topik 2 paling banyak dibahas di UGM.

Rekomendasi sesuai kumpulan jurnal universitas

Sebagai percobaan berikutnya akan dicari rekomendasi pada bagi salah satu universitas terhadap universitas lainya. Untuk percobaan ini JuTISI dipilih untuk dicarikan rekomendasinya terhadap universitas lain. Proses awal, jurnal dari berbagai universitas yang akan dijadikan rekomendasi diproses hingga menghasilkan model topik dengan proses yang sama sebagaimana yang sudah dilakukan sebelumnya. Kemudian konten jurnal JuTISI ditransformasi menjadi *corpus*, *dictionary*, dan *bow*. Model topik kemudian diaplikasikan kepada JuTISI yang sudah ditransformasi dan kemudian dibandingkan dengan jurnal-jurnal pada universitas lainya. Hasil perbandingan pada TABEL XIII menunjukan nilai kedekatan terhadap jurnal masing-masing universitas. Dari sana bisa diketahui jurnal-jurnal apa saja yang memiliki kedekatan terhadap JuTISI dan siapa penulisnya.

TABEL IX
KEDEKATAN MASING – MASING TOPIK TERHADAP JURNAL

	University	Journal	Topic 1	Topic 2
0	BINUS	Deteksi 4 Tanda Vital Pasien Rumah Sakit Berb...	0.083205	0.149733
1	BINUS	Klasifikasi Motif Batik Berbasis Kemiripan Cir...	0.112207	0.210339
2	BINUS	Seleksi Citra Berdasarkan Ciri dengan Algoritm...	1.789223	4.243989
3	BINUS	The Application Of K-Means Algorithm For LQ45 ...	1.623480	3.616180
4	BINUS	Playing the SOS Game Using Feasible Greedy Str...	0.707048	1.172001
...	...	...	...	...
2465	WIDYA_MANDALA	Sistem Pelacakan dan Pengamanan Kendaraan Berb...	2.360630	1.861508
2466	WIDYA_MANDALA	Pengaruh Suhu Pemasakan dan Laju Penambahan Ai...	2.485027	0.332983
2467	WIDYA_MANDALA	Video Mixer Yang Dapat Diprogram	1.853573	0.191648
2468	WIDYA_MANDALA	Mesin penggiling bumbu pecel otomatis berbasis...	1.815556	0.406435
2469	WIDYA_MANDALA	Perancangan Alat Bantu Peletakan Sheet dengan ...	0.639641	0.472445

TABEL X
RATA – RATA NILAI TOPIK 1

Universitas	Topik 1
UNIMED	10.229574
UGM	4.857787
UNDIKSHA	3.973356
UPI	3.121568
TELKOM	2.060027
WIDYA_MANDALA	1.796863
UKM	1.724983
BINUS	1.578034
UI	1.486171
UNDIP	1.111340
ITB	0.961792

TABEL XI
RATA – RATA NILAI TOPIK 2

Universitas	Topik 2
UGM	6.344867
UNIMED	3.541486
UKM	2.906761
BINUS	2.805072
TELKOM	2.399610
UI	2.065756
ITB	1.831246
UNDIKSHA	1.436769
UPI	1.325012
UNDIP	1.139939
UKWMS	0.641613

```
simGroup =
df.groupby('University')['Similarity']
simGroup.mean()
```

Kode Program 2. Kode Python untuk mengambil nilai rata - rata kedekatan per universitas

Setelah itu, sama dengan proses *similarity* dengan kueri kata kunci, setiap jurnal akan dibandingkan terhadap kueri dan akan diambil nilai rata-rata setiap universitas dengan menggunakan Kode Program 2, dan hasilnya dapat dilihat pada TABEL XII. Pada TABEL XII dapat dilihat bahwa ITB memiliki nilai tertinggi pertama, UGM memiliki nilai kedua, dan BINUS memiliki nilai tertinggi ketiga. Berdasarkan nilai tersebut dapat diberikan rekomendasi bagi JuTISI dalam menjalin kerja sama jurnal ilmiah yaitu dengan ITB, UGM, dan BINUS. Selain itu, pada Tabel XIII disertakan juga nama-nama penulis yang yang memiliki nilai kesamaan antara topik penulis dengan topik makalah dalam JuTISI. Dari nilai tersebut dapat direkomendasikan nama penulis yang dapat dijadikan mitra bestari dalam kerja sama jurnal ilmiah.

TABEL XII
RATA RATA NILAI KEDEKATAN PER UNIVERSITAS

Universitas	Nilai Kedekatan
ITB	0.980855
UGM	0.969292
BINUS	0.948059
UNDIP	0.935152
UI	0.924045
Tel-U	0.907489
UKWMS	0.755674
UPI	0.721170
UNDIKSHA	0.652922
UNIMED	0.486192

VI. KESIMPULAN

Hasil akhir yang diberikan dalam makalah ini merupakan nilai kesamaan yang dimiliki antar jurnal universitas berdasarkan dengan kedekatan topik yang dibangun oleh

p-ISSN : 2443-2210
e-ISSN : 2443-2229

model LSI. Dari nilai kesamaan tersebut rekomendasi dapat dibentuk untuk membantu dalam pemilihan calon rekanan jurnal ilmiah. Sebagai contoh, rekomendasi dari percobaan pertama memberikan jurnal ilmiah universitas mana yang memiliki kedekatan pada kata kunci tertentu berdasarkan makalah yang ada pada portal jurnal ilmiah, sehingga ketika akan menjalin kerja sama jurnal ilmiah universitas memiliki acuan dalam memilih mitra. Pada percobaan ketiga rekomendasi diberikan hingga pada tingkat penulis, dari rekomendasi ini maka bisa didapatkan acuan untuk memilih seorang *reviewer* yang yang cocok bagi universitas ketika akan mempublikasikan jurnal ilmiah.

Melalui proses LSI ini, kemiripan diukur berdasarkan hubungan antar kata yang membentuk sebuah kelompok topik. Sebelum model topik dibentuk, dilakukan pengukuran *topic coherence* untuk menentukan seberapa banyak topik yang akan dibuat. Hal ini dilakukan agar jumlah topik yang akan dibuat tidak terlalu sedikit sehingga kurang mewakili konten jurnal yang ada ataupun tidak terlalu banyak sehingga model topik yang dihasilkan menjadi redundan.

Berpijak pada model topik tersebut kemudian dibentuklah beberapa tahap rekomendasi. Rekomendasi pertama, dengan menggunakan kata kunci yang ditransformasi dengan model LSI untuk melihat kedekatannya terhadap topik kemudian dibandingkan terhadap masing-masing jurnal yang memiliki nilai kedekatan terhadap topik. Pada proses ini ditemukan bahwa bahasa yang berbeda, misalnya Indonesia dan Inggris, meskipun *term* pencarian memiliki arti yang sama, menghasilkan kelompok rekomendasi topik yang berbeda. Hal ini dikarenakan proses LSI bukan sekedar menilai kemiripan kueri, namun juga memperhitungkan kemunculan kata-kata secara bersamaan (*word co-occurrences*). Sehingga meskipun topiknya berbeda, namun kata-kata di dalam topik berpotensi sama. Rekomendasi kedua dihitung berdasarkan nilai kedekatan topik terhadap masing-masing jurnal dan diambil rata-ratanya berdasarkan universitas, sehingga dapat diambil kesimpulan universitas mana saja yang cocok untuk dijadikan mitra.

TABEL XIII
HASIL PERBANDINGAN KEDEKATAN JUTISI TERHADAP JURNAL LAINNYA

	University	Author	Journal Title	Similarity
1368	UI	Nursuci Putri Husain, Nursanti Novi Arisa, Put...	LEAST SQUARES SUPPORT VECTOR MACHINES PARAMET...	1.000000
289	BINUS	Joni Suhartono,	Merencanakan Keamanan Jaringan Komputer	1.000000
1347	UI	Mamluatul Hani`ah, Christian Sri Kusuma Aditya...	CORTICAL BONE SEGMENTATION USING WATERSHED AND...	1.000000
473	BINUS	Shinta Mardallena, Melen Melen, Denen Davinely...	ERP System Evaluation on SOFI XP Based Account...	1.000000
645	ITB	Endang Prasetyaningsih, Suprayogi Suprayogi, T...	Production and Delivery Batch Scheduling with ...	1.000000
...	...	...	...	...
2050	Unimed	Sri Hartini, Julaga Situmorang,	PENGEMBANGAN BAHAN AJAR BERBASIS MULTIMEDIA DE...	-0.062013
1870	Unimed	Atika Mahryani Simamora, . Mukhtar,	PENGEMBANGAN MEDIA PEMBELAJARAN KALIMAT EFEKTI...	-0.065501
2066	Unimed	Tumbur Simangunsong, Mukhtar .,	PENGEMBANGAN MEDIA PEMBELAJARAN BERBASIS MULTI...	-0.069424
1884	Unimed	Darmawaty Tarigan, Sahat Siagian,	PENGEMBANGAN MEDIA PEMBELAJARAN INTERAKTIF PAD...	-0.093518
2101	UPI	Sisilia Sylviani, Fahmi Candra Permana, Rio Gu...	PHET Simulation sebagai Alat Bantu Siswa Sekol...	-0.138492

Rekomendasi ketiga, dilakukan dengan membandingkan koleksi jurnal dari sebuah universitas, dalam hal ini JuTISI Maranatha, dengan universitas lainya untuk mendapatkan rekomendasi calon mitra. Dari hasil eksperimen, terlihat bahwa topik-topik yang diterbitkan oleh JuTISI memiliki kedekatan topik yang sangat erat terhadap jurnal bertemakan teknologi informasi yang diterbitkan oleh ITB, UGM, dan BINUS. Berdasarkan hal tersebut maka ITB, UGM, BINUS dapat direkomendasikan bagi JuTISI apabila akan menjalin hubungan kemitraan jurnal ilmiah.

Untuk penelitian lebih lanjut *dataset* yang digunakan dapat dilakukan proses penerjemahan ke dalam satu bahasa. Hal ini perlu dilakukan untuk memperkecil peluang kemunculan kata-kata yang memiliki arti sama, sehingga mendapatkan model topik yang lebih mengerucut. Untuk data yang diambil dapat diperluas dengan mengambil data jurnal tanpa menentukan tema dari fakultas sehingga mendapatkan jumlah topik yang lebih beragam.

DAFTAR PUSTAKA

[1] T. Karthikeyan; K. Sekaran, D. Ranjith, V. Kumar, & M. Balajee, "Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques," *International Journal of Web Portals*, vol. 11, no. 2, pp. 41 - 52, 2019.

[2] S.V. Broucke, & B. Baesens, "From Web Scraping to Web Crawling," *Practical Web Scraping for Data Science, Apress*, 2018, pp. 156-172.

[3] P, Avar & C, Sandip, "Efficient Focused Web Crawling Approach for Search Engine," *Prosiding International Advanced Computing Conference (IACC), IEEE*, 2013, pp. 908-911.

[4] H. Ma, Hualong, X. Wang, J. Hou, & Y. Lu, "Course Recommendation Based on Semantic Similarity Analysis," *Prosiding IEEE 3rd International Conference on Control Science and Systems Engineering, Beijing*, 2017, pp. 638-641.

[5] B. Sakshi, G. Chetna, & A. Anuja, "User tweets based genre prediction and movie recommendation using LSI and SVD," *Prosiding 9th*

International Conference on Contemporary Computing (IC3), Noida, 2016, pp. 1-6.

[6] F. McCarey, M. O. Cinneide, & N. Kushmerick, "Recommending Library Methods: An Evaluation of the Vector Space Model (VSM) and Latent Semantic Indexing (LSI)," *Lecture Notes in Computer Science*, Springer, 2020, pp. 217-230.

[7] P.W. Foltz, "Using Latent Semantic Indexing for Information Filtering," *ACM SIGOIS Bulletin*, vol. 11, no. 2-3, pp. 40-47, 1990.

[8] T.K. Landauer, P.W. Folts, & D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259-284, 1998.

[9] Das, Kumer Pial, "Semantic Similarity of Documents Using Latent Semantic Analysis," *Prosiding National Conference on Undergraduate*, Kentucky, 2014, pp. 1083-1092.

[10] D.A. Grossman & O. Frieder, *Information Retrieval*, Springer, 2004.

[11] S.T. Dumais, "Latent Semantic Analysis," *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 189-229, 2005.

[12] A. Kontostathis & W.M. Pottenger, "A framework for understanding Latent Semantic," *Information Processing and Management*, vol. 42, no. 1, pp. 56-73, 2006.

[13] K. Stevens, P. Kegelmeyer, D. Andrzejewski & D. Buttler, "Exploring Topic Coherence over many models and many topics," *Prosiding Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 952-961.

[14] S. Syed & M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," *Prosiding IEEE International Conference on Data Science and Advanced Analytics* (DSAA), Tokyo, 2017, pp. 165-174.

[15] M. Roder, A. Both & A. Hinnerburg, "Exploring the Space of Topic Coherence Measures," *Prosiding 8th ACM International Conference on Web Search and Data Mining*, New York, 2015, pp. 399-408.

[16] J. Wang & Y. Guo, "Scrapy-based Crawling and User-behavior Characteristics Analysis on Taobao," *Prosiding IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Sanya, China,2012, pp. 44-52.

[17] R. Rehurek. (2019) Gensim topic modelling for humans. [Online]. Tersedia: https://radimrehurek.com/gensim/index.html.

=== HASIL INDEXING OTOMATIS ===

1. Metode TF-IDF:
- topik (Halaman: 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)
- jurnal (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- data (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 14, 15)
- universitas (Halaman: 1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14)
- proses (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)
- lsi (Halaman: 1, 2, 3, 4, 7, 8, 9, 11, 14, 15)
- dokumen (Halaman: 1, 2, 3, 4, 5, 7, 8, 9, 10)
- rekomendasi (Halaman: 1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14)
- hasil (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)
- web (Halaman: 1, 2, 5, 6, 9, 14, 15)
- model (Halaman: 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15)
- berdasarkan (Halaman: 1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14)
- informasi (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- makalah (Halaman: 1, 2, 4, 5, 6, 7, 9, 11, 12, 13, 14)
- kueri (Halaman: 1, 3, 4, 5, 8, 9, 10, 11, 13, 14)
- issn (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- 2443 (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- memiliki (Halaman: 1, 2, 4, 6, 7, 8, 10, 11, 12, 13, 14)
- kunci (Halaman: 1, 2, 3, 4, 5, 8, 9, 11, 12, 13, 14)
- diambil (Halaman: 1, 2, 3, 5, 6, 8, 11, 12, 13, 14)
- topic (Halaman: 1, 4, 7, 8, 9, 10, 11, 12, 14, 15)
- coherence (Halaman: 1, 4, 5, 7, 8, 9, 10, 11, 14, 15)
- halaman (Halaman: 1, 2, 5, 6, 9)
- dibentuk (Halaman: 1, 4, 5, 7, 8, 9, 10, 11, 14)
- teknik (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- sistem (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- mitra (Halaman: 1, 4, 8, 11, 12, 13, 14)
- 2020 (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- ilmiah (Halaman: 1, 2, 4, 5, 6, 11, 13, 14)
- portal (Halaman: 1, 2, 5, 6, 9, 14)
- nomor (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- desember (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- volume (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- pp (Halaman: 1, 2, 6, 8, 9, 11, 14, 15)
- 2229 (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- 2210 (Halaman: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
- nilai (Halaman: 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14)
- kedekatan (Halaman: 2, 3, 5, 8, 10, 11, 12, 13, 14)
- url (Halaman: 2, 6)
- ekstraksi (Halaman: 2, 5, 6, 9)
- metode (Halaman: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)
- tabel (Halaman: 3, 4, 6, 9, 10, 11, 12, 13, 14)
- matriks (Halaman: 3)
- pembentukan (Halaman: 4, 5, 7, 8, 9, 10, 11)
- perhitungan (Halaman: 4, 5, 7, 8, 10, 11)
- pengukuran (Halaman: 4, 5, 7, 9, 10, 14)
- corpus (Halaman: 5, 7, 8, 9, 10, 12)
- gambar (Halaman: 5, 6, 9, 10, 11, 12)
- ugm (Halaman: 6, 11, 12, 13, 14)
- dictionary (Halaman: 7, 8, 9, 10, 12)

2. Metode RAKE:
- halaman web (Halaman: 2)
- model topik (Halaman: 4, 10)
- jurnal teknik informatika (Halaman: 2, 4, 6, 8, 10, 12, 14)
- jurnal ilmiah (Halaman: 1, 13)
- sim q (Halaman: 4)
- kelompok topik (Halaman: 8, 9)
- jurnal jurnal (Halaman: 12)
- model lsi (Halaman: 4, 8, 11)
- pembentukan topik (Halaman: 4, 9, 10)
- berdasarkan universitas (Halaman: 11, 12)
- data dokumen jurnal (Halaman: 1)
- dokumen jurnal (Halaman: 1, 10)
- proses ekstraksi data (Halaman: 2, 5, 6)
- membentuk model topik (Halaman: 5)
- portal jurnal (Halaman: 6)
- kedekatan topik (Halaman: 12, 13)
- base url (Halaman: 2)
- teknik web scraping (Halaman: 2, 5)
- silver truck (Halaman: 3)
- nama penulis (Halaman: 5)
- allowed domain (Halaman: 6)
- machine learning (Halaman: 8)
- topik topik (Halaman: 8, 12)
- tabel iv (Halaman: 11)
- hubungan topik (Halaman: 12)
- tabel x (Halaman: 12)
- abel xii (Halaman: 13)
- web scraping (Halaman: 14)
- membangun topik topik (Halaman: 1, 4)
- local search algorithm (Halaman: 2)
- web portal jurnal (Halaman: 2)
- proses penarikan data (Halaman: 2)
- nilai singular (Halaman: 3)
- memiliki nilai tertinggi (Halaman: 4, 8)
- pengukuran koheren (Halaman: 4)
- memiliki nilai (Halaman: 4)
- topic coherence (Halaman: 4)
- sepuluh universitas (Halaman: 5)
- data target pembanding (Halaman: 5)
- kumpulan jurnal (Halaman: 5)
- hasil ekstraksi data (Halaman: 6)
- portal jurnal ilmiah (Halaman: 6, 14)
- alternatif tata letak (Halaman: 7)
- merk fikaro (Halaman: 7)
- biaya energi (Halaman: 7)
- metode lsi (Halaman: 7, 11)
- pembentukan model topik (Halaman: 7)
- menghasilkan model topik (Halaman: 7, 12)
- pengukuran topic coherence (Halaman: 7, 14)
- hasil pembentukan topik (Halaman: 8, 9)
- ratanya berdasarkan universitas (Halaman: 8, 14)
- nilai kedekatan topik (Halaman: 8, 12)

- contoh hasil (Halaman: 9)
- ekstraksi data (Halaman: 9)
- corpus dictionary (Halaman: 9, 12)
- proses pembentukan rekomendasi (Halaman: 11)
- kode progra (Halaman: 12)
- tabel xi (Halaman: 12)
- universitas lainya (Halaman: 12)
- jurnal rekomendasi (Halaman: 12)
- diambil nilai (Halaman: 12, 13)
- mengambil nilai (Halaman: 12, 13)
- kode program (Halaman: 13)
- itb ugm (Halaman: 13)
- nilai kedekatan (Halaman: 13)
- tabel xii (Halaman: 13)
- nilai kesamaan (Halaman: 13)
- jurnal ilmiah universitas (Halaman: 14)
- journal topic modeling (Halaman: 1)
- fakultas rangkaian percobaan (Halaman: 1)
- maranatha edu abstract (Halaman: 1)
- laman web science (Halaman: 1)
- koleksi jurnal universitas (Halaman: 1)
- journal topics coherence (Halaman: 1)
- partnering recommendation list (Halaman: 1)
- information technology themes (Halaman: 1)
- cooperation recommendation strategy (Halaman: 1)
- university journal portal (Halaman: 1)
- semantic representation (Halaman: 1)
- using keywords (Halaman: 1)
- categorical similarity (Halaman: 1)
- matched key words (Halaman: 1)
- internet terkadang tersedia (Halaman: 1)
- higher education institution (Halaman: 1)
- dunia pend idikan (Halaman: 1)
- b erdasarkan pengelompokkan (Halaman: 1)
- reviewer mitra bestari (Halaman: 1)
- pendahuluan menemukan mitra (Halaman: 1)
- rekomendasi rekomendasi dibentuk (Halaman: 1)
- menentukan kriteria mitra (Halaman: 1)
- universitas bereputasi (Halaman: 1)
- penulis makalah perbandingan (Halaman: 1)
- lsi kesamaan (Halaman: 1)
- variabel name (Halaman: 2)
- sistem lo kal (Halaman: 2)
- menentukan pembatasan url (Halaman: 2)
- menentukan halaman web (Halaman: 2)
- membatasi proses crawling (Halaman: 2)
- proses web scraping (Halaman: 2)
- web scrapin g (Halaman: 2)
- menghasilkan himpunan topik (Halaman: 2)
- membatasi url (Halaman: 2)
- dihitung kedekata nnya (Halaman: 2)
- menghitung nilai agregasi (Halaman: 2)
- semantic indexing lsi (Halaman: 2)

- dihitung berdasarkan entitas (Halaman: 3)
- truck proses lsi (Halaman: 3)
- proses komputasi svd (Halaman: 3)
- diambil nilai singular (Halaman: 3)
- menyimpan nilai eigenvector (Halaman: 3)
- dokumen lsi bergantung (Halaman: 3)
- memanfaatkan asosiasi berdimensi (Halaman: 3)
- menghasilkan struktur relasi (Halaman: 3)
- proses mencari kesamaan (Halaman: 3)
- perkalian matriks tsdt (Halaman: 3)
- pola kemunculan (Halaman: 3)
- shipment silver truck (Halaman: 3)
- dokumen matriks asli (Halaman: 3)
- gold damaged (Halaman: 3)
- menyusun informasi (Halaman: 3)
- menyaring gangguan (Halaman: 3)
- gold arrived (Halaman: 3)
- k dimensi (Halaman: 3)
- lsi matriks (Halaman: 3)
- kueri q (Halaman: 3)
- berdasarkan tabel (Halaman: 3)
- silver arrived (Halaman: 3)
- dokumen j tabel (Halaman: 3)
- parameter k (Halaman: 3)
- metode temu (Halaman: 3)
- vektor dokumen (Halaman: 3)
- struktur semantik (Halaman: 3)
- dataset struktur (Halaman: 3)
- ruang dimensi (Halaman: 3)
- menggabungkan elemen dasar (Halaman: 4)
- mengaplikasikan cosine similarity (Halaman: 4)
- pengukuran konfirmasi individu (Halaman: 4)
- didapatkan sim q (Halaman: 4)
- urutan kepentingan topik (Halaman: 4)
- prediksi topik terkadang (Halaman: 4)
- terkadang berkorelasi negatif (Halaman: 4)
- diekstrak topic coherence (Halaman: 4)
- menampilkan ukuran koheren (Halaman: 4)
- dibentuk topic modeling (Halaman: 4)
- diukur nilai koherensi (Halaman: 4)
- formula hasil kemiripan (Halaman: 4)
- informasi semantik (Halaman: 4)
- koordinat vektor (Halaman: 4)
- nilai vektor kueri (Halaman: 4)
- pengambilan informasi (Halaman: 4)
- dokumen kueri (Halaman: 4)
- proses pengukuran (Halaman: 4)
- jenis pengukuran (Halaman: 4)
- mengurutkan dokumen (Halaman: 4)
- mengidentifikasi dokumen (Halaman: 4)
- memiliki probabilitas (Halaman: 4)
- konfirmasi ukuran (Halaman: 4)
- koleksi dokumen (Halaman: 4)

- perhitungan kemiripan (Halaman: 4)
- kuncirekomendasi berdasarkan nilai (Halaman: 5)
- gambar data jurnal (Halaman: 5)
- portal jurnal fakultas (Halaman: 5)
- spesifik sumber data (Halaman: 5)
- dicarikan rekomendasi kerja (Halaman: 5)
- jurnal format file (Halaman: 5)
- gambar menggambarkan garis (Halaman: 5)
- nilai pasangan himpunan (Halaman: 5)
- data data jurnal (Halaman: 5)
- target pembanding data (Halaman: 5)
- penentuan koherensi topik (Halaman: 5)
- technology index sinta (Halaman: 5)
- situs jurnal ilmiah (Halaman: 5)
- umass mengkalkulasi koherensi (Halaman: 5)
- ekstraksi data penentuan (Halaman: 5)
- web crawling scrapy (Halaman: 5)
- menghitng frekuensi kemunculan (Halaman: 5)
- corpus iii metodologi (Halaman: 5)
- universitas dianggap (Halaman: 6)
- parameter start url (Halaman: 6)
- memiliki base url (Halaman: 6)
- web portal resmi (Halaman: 6)
- ekstraksi data jurnal (Halaman: 6)
- kriteria allowed domain (Halaman: 6)
- gambar meliputi (Halaman: 6)
- data judul abstrak (Halaman: 6)
- pra pemrosesan data (Halaman: 6)
- mengekstrak struktur data (Halaman: 6)
- contoh preproses data (Halaman: 6)
- situs web (Halaman: 6)
- url pdf (Halaman: 6)
- metode ekstraksi (Halaman: 6)
- pengumpulan data (Halaman: 6)
- pemroses data (Halaman: 6)
- batasan data (Halaman: 6)
- perbaikan tata letak (Halaman: 6)
- mewakili topik topik (Halaman: 6)
- menghilangkan tag html (Halaman: 6)
- bangga mengh apus (Halaman: 6)
- penataan luas areal (Halaman: 7)
- jarak pemindahan bahan (Halaman: 7)
- biaya investasi perpipaan (Halaman: 7)
- mengurangi biaya biaya (Halaman: 7)
- departemen depar temen (Halaman: 7)
- menentukan efisiensi (Halaman: 7)
- algoritma corel ap (Halaman: 7)
- model topik selesai (Halaman: 7)
- proses lsi hasil (Halaman: 7)
- redundan si (Halaman: 7)
- perhitungan topic coherence (Halaman: 7)
- training model (Halaman: 7)
- model bag (Halaman: 7)

- topik machine learning (Halaman: 8)
- kelompok topik divariasikan (Halaman: 8)
- perkembangan siswa (Halaman: 8)
- kueri machine learning (Halaman: 8)
- perusahaan topic coherence (Halaman: 8)
- universitas hasil (Halaman: 8)
- hasil metode (Halaman: 8)
- proses pembentuk lsi (Halaman: 8)
- hasilnya diban dingkan (Halaman: 8)
- aplikasi proses bisnis (Halaman: 8)
- komprehensif proses perhitungan (Halaman: 8)
- proses lsi (Halaman: 8)
- mengubah bentuk kueri (Halaman: 8)
- rekomendasi nilai similarity (Halaman: 8)
- merepresentasikan karakteristik jurnal (Halaman: 8)
- mencari mitra universitas (Halaman: 8)
- memiliki bobot nilai (Halaman: 8)
- vektor lsi (Halaman: 8)
- metode pengukuran (Halaman: 9)
- v diskusi hasil (Halaman: 9)
- dictionary gambar (Halaman: 9)
- tabel data (Halaman: 9)
- otomatis menarik data (Halaman: 9)
- persiapan data (Halaman: 9)
- halaman selesai diekstraksi (Halaman: 9)
- u mass proses (Halaman: 9)
- pustaka gensim proses (Halaman: 9)
- tama membentuk model (Halaman: 9)
- bentuk grafik bergaris (Halaman: 9)
- pembentukan bag (Halaman: 9)
- ditentukanlah kueri (Halaman: 9)
- proses pembentuk (Halaman: 9)
- halaman makalah (Halaman: 9)
- halaman lainya (Halaman: 9)
- metode uci umass (Halaman: 10)
- gensim hasil (Halaman: 10)
- menunjukan kedekatan kueri (Halaman: 10)
- kelas matrix simmilarity (Halaman: 10)
- perhitungan similarity (Halaman: 10)
- proses perhitungan (Halaman: 10)
- universitas universitas (Halaman: 10)
- c v (Halaman: 10)
- penentuan topik (Halaman: 10)
- mencegah topik (Halaman: 10)
- kisaran topik (Halaman: 10)
- corpus dibentuk (Halaman: 10)
- percobaan pembentukan rekomendasi (Halaman: 11)
- topik berpotensi (Halaman: 11)
- memunculkan topik (Halaman: 11)
- hasil perhitungan kemiripan (Halaman: 11)
- perbedaan hasil kemiripan (Halaman: 11)
- jurnal dihitunglah nilai (Halaman: 11)
- k unci rekomendasi (Halaman: 11)

- diproses berdasarkan kemunculan (Halaman: 11)
- word co occurrences (Halaman: 11)
- diperoleh nilai kedekatan (Halaman: 11)
- berdas arkan kedekatan (Halaman: 11)
- perbedaan bahasa lsi (Halaman: 11)
- kunci kueri (Halaman: 11)
- bahasa tercampur (Halaman: 11)
- uji coba (Halaman: 11)
- jurnal tabel ix (Halaman: 12)
- dijadikan rekomendasi diproses (Halaman: 12)
- hasil aplikasi topik (Halaman: 12)
- kedekatan kelompok topik (Halaman: 12)
- k elompok topik (Halaman: 12)
- diambil kesimpulan topik (Halaman: 12)
- bow model topik (Halaman: 12)
- jur nal nilai (Halaman: 12)
- dicari rekomendasi (Halaman: 12)
- memiliki kedekatan (Halaman: 12)
- diamati kedekatan (Halaman: 12)
- dijadikan mitra (Halaman: 12)
- univer sitas (Halaman: 12)
- pr oses (Halaman: 12)
- jutisi dipilih (Halaman: 12)
- dicarikan rekomendasinya (Halaman: 12)
- ugm memiliki nilai (Halaman: 13)
- jurnal universitas berdasarkan (Halaman: 13)
- topik penulis deng (Halaman: 13)
- proses similarity (Halaman: 13)
- memiliki nilai kesamaan (Halaman: 13)
- topik makalah (Halaman: 13)
- jurnal tabel x (Halaman: 13)
- nama nama penulis (Halaman: 13)
- direkomendasikan nama penulis (Halaman: 13)
- dijadikan mitra bestari (Halaman: 13)
- menjalin kerja (Halaman: 13)
- web portals vol (Halaman: 14)
- memiliki nilai kedekatan (Halaman: 14)
- mengambil data jurnal (Halaman: 14)
- svd prosiding th (Halaman: 14)
- rekomendasi calon mitra (Halaman: 14)
- mode l lsi (Halaman: 14)
- mempublikasikan jurnal ilmiah (Halaman: 14)
- memiliki kedekata n (Halaman: 14)
- ranjith v kumar (Halaman: 14)
- control science (Halaman: 14)
- mewakili konten jurnal (Halaman: 14)
- membandingkan koleksi jurnal (Halaman: 14)
- menentuka n tema (Halaman: 14)
- karthikeyan k sekaran (Halaman: 14)
- memperkecil peluang kemunculan (Halaman: 14)
- model topik dibentuk (Halaman: 14)
- tahap rekomendasi rekomendasi (Halaman: 14)
- infor mation science (Halaman: 15)

- data science (Halaman: 15)
- spruit full text (Halaman: 15)
- kontostathis w (Halaman: 15)
- technology vol (Halaman: 15)
- management vol (Halaman: 15)
- empirical methods (Halaman: 15)
- web search (Halaman: 15)
- hinnerburg exploring (Halaman: 15)

3. Metode Word2Vec:
- Kata kunci: manfaat
  Tidak ada kata mirip yang ditemukan di dokumen.