LATENT DIRICHLET ALLOCATION UNTUK PEMODELAN TOPIK ABSTRAK DOKUMEN SKRIPSI

(Studi Kasus: Abstrak Dokumen Skripsi Mahasiswa Statistika UII Tahun Angkatan 2011-2015)

TUGAS AKHIR

Ella Anggraini

16611033

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2020

LATENT DIRICHLET ALLOCATION UNTUK PEMODELAN TOPIK ABSTRAK DOKUMEN SKRIPSI

(Studi Kasus: Abstrak Dokumen Skripsi Mahasiswa Statistika UII Tahun Angkatan 2011-2015)

TUGAS AKHIR

Ella Anggraini

16611033

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2020

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	*Latent Dirichlet Allocation* Untuk Pemodelan Topik Abstrak Dokumen Skripsi
Nama Mahasiswa	:	Ella Anggraini
NIM	:	16611033

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 18 Maret 2020

Pembimbing

Ayundyah Kesumawati, S.Si., M.Si.

HALAMAN PENGESAHAN

TUGAS AKHIR

LATENT DIRICHLET ALLOCATION UNTUK PEMODELAN TOPIK ABSTRAK DOKUMEN SKRIPSI

(Studi Kasus: Abstrak Dokumen Skripsi Mahasiswa Statistika UII Tahun Angkatan 2011-2015)

Nama Mahasiswa : Ella Anggraini

NIM : 16611033

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL : 3 April 2020

Nama Penguji:	**Tanda Tangan**
1. Dr. RB. Fajriya Hakim, M.Si.	……..
2. Rahmadi Yotenka, M.Sc.	……..
3. Ayundyah Kesumawati, S.Si., M.Si.	……..

Mengetahui,
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Prof. Riyanto, S.Pd., M.Si., Ph.D

KATA PENGANTAR

Assalamu'alaikum Wr.Wb

Alhamdulillah, dengan mengucap segala puji dan syukur bagi Allah SWT yang telah memberikan nikmat dan hidayah-Nya berupa kesabaran, kekuatan, serta kelancaran dalam penulisan ini dari awal hingga terselesaikan penulisan ini. Sholawat serta salah tak lupa dihaturkan kepada Nabi Muhammad SAW, beserta keluarga, sahabat dan umatnya.

Penulisan ini yaitu dengan judul "*LATENT DIRICHLET ALLOCATION* UNTUK PEMODELAN TOPIK ABSTRAK DOKUMEN SKRIPSI". Pada penulisan ini tak luput dari bantuan berbagai pihak, berupa bimbingan, saran, kritik dan bantuan yang lainnya yang diberikan kepada penulis. Dengan begitu izinkan penulis mengucapkan terima kasih kepada:

1. Ibuku Siti Rahmah dan Ayahku Muklis yang tercinta yang sudah memberikan seluruh tenaga dan fikirannya dalam mendukung semua proses perkuliahan dari awal hingga selesai, selalu memberikan semangat dan motivasi untuk selalu kuat dalam keadaan apapun. Tak lupa Adik Ferdi dan Adik Tasya yang memberikan semangat tersendiri kepada penulis untuk selalu memberikan contoh yang baik. Penulis sangat menyayangi mereka.
2. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam.
3. Bapak Dr. Edy Widodo, S.Si., M.Si, selaku Ketua Program Studi Statistika, Universitas Islam Indonesia.
4. Ibu Ayundyah Kesumawati, S.Si., M.Si. selaku dosen pembimbing yang telah memberikan ilmunya, waktu luang dan dengan sabar membimbing hingga terselesaikan penyusunan ini, selalu memberikan semangat serta motivasi kepada penulis.
5. Ibu Arum Handini Primandari, S.Pd.Si.,M.Sc. yang telah memberikan banyak arahan dan bimbingan kepada penulis.

6. Bapak serta Ibu Dosen Program Studi Statistika, Universitas Islam Indonesia yang telah memberikan banyak ilmunya selama masa perkuliahaan dan membagikan banyak cerita pengalamannya kepada penulis.
7. Sahabat terbaik “Bukan Grup” yaitu Hanum, Diana, Anggit, Revika, Yesi, Billa, dan Nadia yang sudah menemani dari awal masuk kuliah hingga sekarang, yang selalu menemani, memberikan motivasi, memberikan banyak pelajaran dan berbagai pengalaman.
8. Partner terbaik yaitu Vegi Kresnadi yang selalu memberikan semangat dan selalu menjadi partner terbaik untuk penulis dan tetap menjadi partner sampai seterusnya.
9. Keluarga besar Marching Band UII yang sangat-sangat besar memberikan banyak pengalaman diluar dari perkuliahan yaitu penulis dapat berorganisasi dalam sela-sela perkuliahan. Teruntuk anak Snareku tersayang yaitu Intan, Novy, Bunga dan Sultan sudah memberikan banyak canda, tawa, tangis yang tidak akan pernah terlupakan.
10. Teman-teman seperjuangan dalam bimbingan yaitu Kides, Anis, Cinmey, Laras, Dea, dan Mita yang berjuang bersama dalam penyusunan TA dan saling bertukar pikiran jika terdapat suatu masalah.
11. Seluruh teman-teman Statistika angkatan 2016 yang telah memberikan banyak informasi dalam perkuliahan dan berjuang dalam perkuliahan selama ini.
12. Teman-teman KKN yaitu Ratih, Iffa, Poppy, Wendy, Bibi, Rafi dan bang Affan yang memberikan banyak pengalaman dan pelajaran selama 1 bulan ditempat KKN Majan, Purworejo.
13. Serta seluruh pihak-pihak lainnya yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan bantuan dalam penyusuan tugas akhir ini dengan penuh perjuangan.

Meskipun penyusunannya telah dibuat dengan sebaik-baiknya, namun tugas akhir ini masih jauh dari kata sempurna. Dengan begitu segala kritik dan saran sangat diharapkan dalam penulisan ini. Semoga tugas akhir ini dapat memberikan

manfaat bagi penulis dan semua yang membutuhkan. Semoga Allah senantiasa memberikan berkah dan hidayah-Nya untuk kita semua, Aamiin Ya Allah.

Wassalamualaikum Wr.Wb

Yogyakarta, 3 April 2020

Ella Anggraini

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 3 April 2020

(dengan materai 6000)

Ella Anggraini

ABSTRAK

LATENT DIRICHLET ALLOCATION UNTUK PEMODELAN TOPIK ABSTRAK DOKUMEN SKRIPSI

(Studi Kasus: Abstrak Dokumen Skripsi Mahasiswa Statistika UII Tahun Angkatan 2011-2015)

Ella Anggraini
Program Studi Statistika, Fakultas MIPA
Universitas Islam Indonesia

Skripsi atau tugas akhir merupakan salah satu syarat sebuah kelulusan dalam setiap universitas dalam bentuk penelitian sesuai dengan jurusan masing-masing, tentunya dalam penyusunan skripsi harus menggunakan topik dan metode yang belum pernah digunakan sebelumnya. Sehingga semakin lama, topik dan metode yang diambil semakin bervariasi dan mengikuti perkembangan. Di Universitas Islam Indonesia terdapat web yang berisi dokumen skripsi yaitu https://dspace.uii.ac.id/. Dengan begitu semakin bertambahnya tahun maka semakin banyak pula dokumen skripsi dengan metode dan topik yang beragam, salah satunya yang berada di Prodi Statistika UII. Dengan semakin banyak dokumen skripsi yang ada sehingga penulis melakukan analisis Topic Modeling menggunakan metode Latent Dirichlet Allocation (LDA) dari abstak dokumen skripsi mahasiswa Statistika UII. Dengan analisis Topic Modeling yang digunakan untuk mengetahui topik apa saja yang sering muncul. Diperoleh jumlah topiknya sebanyak 3 yaitu topik ke-1 mengenai Artificial Intelligence (AI), topik ke-2 mengenai Statistika pada Bidang Kesehatan, dan topik ke-3 mengenai Statistika Perekonomian.

Kata Kunci: *Abstrak Skripsi, Topic Modeling, Latent Dirichlet Allocation*

ABSTRACT

LATENT DIRICHLET ALLOCATION FOR TOPIC MODELING ABSTRACT THESIS DOCUMENTS

Ella Anggraini
Program Studi Statistika, Fakultas MIPA
Universitas Islam Indonesia

Thesis or final project is one of the requirements for graduation in every university in the form of research in accordance with their respective majors, of course in the preparation of the thesis must use topics and methods that have never been used before. So that the longer, the topics and methods taken are increasingly varied and keep up with developments. At the Islamic University of Indonesia there is a web containing thesis documents, namely https://dspace.uii.ac.id/. With so increasing number of years, there are also more thesis documents with diverse methods and topics, one of which is in the UII Statistics Study Program. With more and more thesis documents available, the authors conducted a Topic Modeling analysis using the Latent Dirichlet Allocation (LDA) method of abstraction of UII Statistics student thesis documents. With Topic Modeling analysis used to find out what topics often arise. Obtained the number of topics as much as 3 topics, the first topic is Artificial Intelligence (AI), the second topic is Statistics in Health, and the third topic is Economic Statistics.

Keywords: *Thesis Abstract, Topic Modeling, Latent Dirichlet Allocation*

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Pendidikan merupakan sebuah upaya yang terencana dalam proses pembimbingan dan pembelajaran bagi individu untuk dapat berkembang dan tumbuh menjadi seseorang yang mandiri, bertanggungjawab, kreatif, berilmu, sehat, dan berakhlak mulia baik dilihat dari aspek jasmani maupun rohani. Pendidikan juga merupakan salah satu hal terpenting dalam membentuk kepribadian seseorang. Pendidikan terdiri dari pendidikan informal dan non formal. Pendidikan formal yaitu jalur pendidikan yang terstruktur dan berjenjang yang terdiri atas pendidikan dasar, menengah, atas dan pendidikan tinggi (Ilma, 2015). Pada pendidikan tinggi seperti pada jenjang perkuliahan Sarjana, Magister, Doktor.

Pada jenjang perkuliahan sarjana, mahasiswa biasanya menempuh perkuliahan kurang lebih 4 tahun. Dalam menempuh gelar sarjana, mahasiswa harus membuat skripsi atau tugas akhir yang menjadi salah satu syarat sebuah kelulusan dalam setiap universitas. Dalam penyusunan skripsi, tentunya mahasiswa menyusun dengan topik dan metode yang belum pernah digunakan sebelumnya. Mahasiswa harus mencari metode dan topik yang akan digunakan yang biasanya diambil sesuai dengan jurusan masing-masing. Dalam mengambil topik skripsi, biasanya mahasiswa memilih topiknya sendiri atau diarahkan oleh dosen pembimbing masing-masing. Semakin lama, topik dan metode yang diambil semakin bervariasi dan mengikuti perkembangan pada masa sekarang. Pada penyusunan skripsi salah satunya yaitu terdapat bagian intisari atau *abstract*.

Abstrak merupakan bagian paling penting dalam makalah ilmiah, artikel atau laporan penelitian, dimana memungkinkan pembaca dapat dengan mudah mengidentifikasi secara cepat dan akurat pada isi dasar dari dokumen. Dalam sebuah dokumen, kemudian untuk pembaca memutuskan harus membaca atau tidak keseluruhan dokumen tersebut, menentukan relevansi dari dokumen tersebut dengan kepentingan mereka, memungkinkan dengan menggunakan abstrak dari

dokumen tersebut. Maka abstrak wajib disertai untuk masing-masing makalah ilmiah. Untuk menulis ringkasan singkat dari keseluruhan kandungan dokumen dalam kalimat yang mampu mewakili keseluruhan isi dokumen dengan jelas tentu saja tidak begitu mudah. Abstrak harus berisi informasi spesifik yang cukup dimana dapat memenuhi kebutuhan pembaca yang sedang mencari sumber informasi terkait (Nasution, 2017).

Di Universitas Islam Indonesia telah tersedia web yang dapat diakses oleh mahasiswa untuk mencari dokumen skripsi yaitu https://dspace.uii.ac.id/. Dimana semakin bertambahnya tahun maka semakin banyak pula dokumen skripsi dengan metode dan topik yang beragam, salah satunya yang berada di Prodi Statistika UII. Mahasiswa dalam mengambil topik skripsi juga berkonsultasi terlebih dahulu dengan dosen pembimbing masing-masing, sehingga dosen juga berperan dalam pengambilan topik yang diambil oleh mahasiswa bimbingannya. Banyak pula mahasiswa yang mengajukan topik skripsi terkadang mengambil metode yang telah digunakan peneliti terdahulu namun mengganti topik yang digunakan saja. Dengan semakin banyaknya dokumen skripsi yang masuk dan tidak terdapat metode apakah yang banyak diambil oleh mahasiswa Statistika sehingga penulis akan melakukan analisis *Topic Modeling* dari abstak dokumen skripsi mahasiswa Statistika UII. *Topic modeling* merupakan sebuah topik yang terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen memiliki probabilitas masing-masing dari beberapa topik yang dihasilkan (Putra & Kusumawardani, 2017). Dengan analisis *Topic Modeling* yang digunakan untuk mengetahui topik apa saja yang sering muncul dalam dokumen tersebut. Menurut Blei (2003) untuk dapat menganalisis *Topic modeling* penulis menggunakan metode *Latent Dirichlet Allocation* (LDA) (Putra & Kusumawardani, 2017).

Intuisi dari LDA itu sendiri yaitu sebuah dokumen memiliki beberapa topik yang terdapat di dalamnya (Herwanto, 2018). Dengan menggunakan LDA untuk penelitian ini yang dimana LDA dapat membuat dokumen menghasilkan beberapa topik dari abstrak dokumen statistika UII, sehingga hasil yang didapatkan dapat menjadi pertimbangan dimana topik apa saja yang sudah banyak diambil oleh mahasiswa statistika yang telah menyelesaikan skripsinya. Dengan begitu

mahasiswa yang akan menyusun skripsi dan dosen pembimbingnya dapat mengetahui atau mempunyai acuan dalam menentukan topik dan metode dalam penyusunan skripsi.

Penelitian yang dilakukan oleh Putra dan Kusumawardani (2017) mengenai Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation* (LDA). Data yang digunakan oleh penulis yaitu laporkan peristiwa atau kejadian yang diperoleh dari partisipasi masyarakat. Jumlah laporan perhari yang tinggi dan berbagai macam topik dari laporan tersebut menyebabkan kesusahan dalam mengidentifikasi suatu topiknya sehingga akan menghabiskan banyak waktu jika dilakukan secara manual oleh manusia. Dengan demikian sangat dibutuhkan pemodelan topik yang dapat secara otomatis mengklasifikasi pesan media sosial tersebut ke dalam topik-topik yang muncul dari hasil pemodelannya. Pemodelan topik dilakukan dengan metode *Latent Dirichlet Allocation* (LDA). Dengan metode LDA didapatkan hasil jumlah topik yang terdapat dalam pesan media sosial yaitu 4 topik dengan nilai *perplexity* terbaik yaitu sebesar 213.41 dan diuji kemudahannya untuk diinterpretasi oleh manusia dengan uji koherensi topik yang terdiri dari *word intrusion task* dan *topic intrusion task*. Kesimpulannya dengan menggunakan metode LDA, model yang didapatkan dapat diinterpretasi manusia dengan baik.

Hasil yang diperoleh dari analisis Topic Modeling dengan *Latent Dirichlet Allocation* ini diharapkan akan menjadi pertimbangan oleh mahasiswa serta dosen dalam menentukan topik atau metode yang akan di ambil oleh mahasiswa Statistika UII dalam penyusunan Skripsi.

1.2 Rumusan Masalah

Berdasarkan latar belakakang masalah, berikut ini merupakan rumusan yang dapat diidentifikasi:

1. Bagaimana gambaran umum dari abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015?
2. Bagaimana implementasi pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) pada data abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015?

1.3 Batasan Masalah

Berikut ini merupakan batasan masalah yang digunakan dalam penelitian ini:

1. Data yang digunakan merupakan data sekunder yang diperoleh dari *website* https://dspace.uii.ac.id/ dan didapatkan dari perpustakaan Universitas Islam Indonesia.
2. Data yang digunakan yaitu data abstrak skripsi.
3. Penelitian ini hanya difokuskan pada mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015.
4. Data yang diambil merupakan data sampel.
5. Data diolah dengan menggunakan *software* Anaconda 3.5.3.1 dengan metode analisis yang digunakan yaitu *Topic Modelling* dengan *Latent Dirichlet Allocation* (LDA). *Microsoft Excel* 2007 sebagai penyimpanan data dan pengolahan diagram.

1.4 Tujuan Penelitian

Berikut ini merupakan tujuan dari penelitian yang akan dilakukan:

1. Mengetahui gambaran umum dari abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015.
2. Mengetahui implementasi pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) pada data abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015.

1.5 Manfaat Penelitian

Berikut ini merupakan manfaat yang bisa diambil dari penelitian ini:

1. Memberikan gambaran umum dari abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015 yang digunakan dalam penelitian ini berupa proporsi jumlah dokumen yang digunakan dan lama pengerjaan TA tiap tahun angkatannya.
2. Hasil dari pemodelan topik berupa topik dan metode apa yang sering diambil oleh mahasiswa dalam penyusunan skripsi dapat menjadi bahan pertimbangan oleh Prodi Statistika UII yang meliputi mahasiswa serta dosen Statistika UII

dalam menentukan topik dan metode yang akan di ambil oleh mahasiswa Statistika UII dalam penyusunan Skripsi.

BAB II
TINJAUAN PUSTAKA

Pada penelitian yang akan dilakukan, peneliti membutuhkan beberapa penelitian terdahulu yang dijadikan sebagai bahan acuan yang bertujuan untuk dapat mengetahui hubungan antara penulisan terdahulu dengan penulisan yang dilakukan penulis saat ini agar terhindar dari plagiasi.

Penelitian yang dilakukan oleh Rahmawati, Sihwi, & Suryani (2014) mengenai Analisa *Clustering* Menggunakan Metode *K-Means* dan *Hierarchical Clustering* (Studi Kasus: Dokumen Skripsi Jurusan Kimia, Fmipa, Universitas Sebelas Maret). pengelompokan teks dapat dilakukan dengan *text mining* menggunakan metode *clustering*. Data pada penelitian menggunakan bagian abstrak dari skipsi. Hasil dari *clustering* didapatkan 16 *cluster* yang kemudian dianalisa keterkaitannya antar dokumennya dan diperkirakan tema dari tiap *cluster*nya, keterkaitannya dengan dosen yang mengajar. Dari hasil *cluster* memperlihatkan bahwa keahlian dosen mempengaruhi variasi tema penelitian yang dilakukan mahasiswa. Didapatkan juga bahwa banyaknya penelitian di suatu tema berkaitan dengan minat sisa dan proyek dosen di jurusan Kimia.

Penelitian yang dilakukan oleh Prilianti & Wijaya (2014) mengenai Aplikasi *Text Mining* untuk Automasi Penentuan Tren Topik Skripsi dengan Metode *K-Means Clustering*. Suatu pengetahuan mengenai tren topik skripsi mahasiswa di Universitas atau pada program studi dapat membawa manfaat yang positif untuk pengembangan kurikulum atau *roadmap* penelitian skala institusi. Akan tetapi, tekhnologi yang dapat cepat mendapatkan pengetahuan dengan menyeluruh dari penelitian-penelitian mahasiswa melalui skripsi sangat terbatas dibandingkan dengan penyimpanan dokumen yang tersedia. Maka dilakukan penelitian ini dikembangkan aplikasi yang dapat memperoleh pegetahuan dari topik-topik skripsi mahasiswa yang terdapat dalam *repository* digital perpustakaan Universitas. Penelitian ini menggunakan *text mining* dan algoritma *k-means clustering* terhadap

dokumen abstrak skripsi. Dari hasil uji coba didapatkan hasil yaitu 89% responden menyatakan tren topik skripsi yang dihasilkan dari aplikasi sesuai dengan kondisi sesungguhnya.

Penelitian yang dilakukan oleh Putra & Kusumawardani (2017) mengenai Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation* (LDA). Pada Radio Suara Surabaya yang mengembangkan tentang siaran interaktif dengan basis jurnalistik masyarakat, yang melibatkan partisipasi dari masyarakat dalam melaporkan peristiwa atau kejadian kepada penyiar radio yang bertugas. Jumlah laporan perhari yang tinggi dan berbagai macam topik dari laporan tersebut menyebabkan kesusahan dalam mengidentifikasi suatu topiknya sehingga akan menghabiskan banyak waktu jika dilakukan secara manual oleh manusia. Dengan demikian sangat dibutuhkan pemodelan topik yang dapat secara otomatis mengklasifikasi pesan media sosial tersebut ke dalam topik-topik yang muncul dari hasil pemodelannya. Pemodelan topik dilakukan dengan metode *Latent Dirichlet Allocation* (LDA). Dengan metode LDA didapatkan hasil jumlah topik yang terdapat dalam pesan media sosial yaitu 4 topik dengan nilai *perplexity* terbaik yaitu sebesar 213.41 dan diuji kemudahannya untuk diinterpretasi oleh manusia dengan uji koherensi topik yang terdiri dari *word intrusion task* dan *topic intrusion task*. Kesimpulannya dengan menggunakan metode LDA, model yang didapatkan dapat diinterpretasi manusia dengan baik.

Penelitian yang dilakukan oleh Februariyanti & Santoso (2017) mengenai *Hierarchical Agglomerative Clustering* Untuk Pengelompokan Skripsi Mahasiswa. *Text mining* pada umumnya dapat digunakan untuk mengelompokkan suatu data penelitian yang berbentuk teks. Data dari penelitian ini merupakan dokumen skripsi Program Studi Sistem Informasi Fakultas Teknologi Informasi Universitas Stikubank Semarang dimana akan mengelompokkan judul skripsi dari mahasiswa. Untuk dapat mengetahui dari kemiripan antar judul dari dokumen skripsi maka dilakukan proses menghitung jarak antar objek atau kemiripan dari masing-masing judul dari dokumen skripsi dengan menggunakan algoritma *dice coefficient*. Objek judul dengan jarak terdekat merupakan objek judul yang paling mirip. Didapatkan hasil dokumen paling mirip yang diperoleh dari proses perhitungan jarak kemiripan

antar dokumen yaitu dengan jarak sebesar 0.2. Kemudian hasil dari clustering yaitu dilakukan pemotongan 2 titik yaitu menghasilkan 1 buah *cluster* dari pemotongan dendrogram pada titik 0.25 dan menghasilkan 5 buah *cluster* dari pemotongan dendrogram pada titik 0.5.

Penelitian yang dilakukan oleh Agustina (2017) mengenai Analisis dan visualisasi suara pelanggan pada pusat layanan pelanggan dengan pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA) studi kasus: PT. Petromkimia Gresik. Pada perusahaan tersebut salah satunya yaitu memproduksi pupuk dimana pupuk merupakan kebutuhan penting dalam bidang pertanian. Kepuasan pelanggan menjadi salah satu aspek yang diperhatikan oleh perusahaan salah satunya yaitu dengan adanya aplikasi Pusat Layanan Pelanggan yang berguna untuk mengoptimalisasi suara pelanggannya. Pada pencatatan suara pelanggan yang telah masuk dengan berbagai media tetapi belum terdapat analisis terhadap topik apakah yang sering pelanggan suarakan maka dilakukan analisis topik dengan menggunakan pemodelan LDA. Didapatkan hasil yaitu 35 topik yang berhasil diidentifikasi kemudian dikelompokkan ke dalam 7 kategori. Nilai *perplexity* yang diperoleh sebesar 34.92 dengan standar deviasi 0.49 pada 20 iterasi dan nilai akurasinya sebesar 83.7%. Kemudian hasil identifikasi divisualisasikan dalam *dashboard* berbasis web.

Penelitian yang dilakukan oleh Wicaksana, Adikara, & Adinugroho (2018) mengenai *Clustering* Dokumen Skripsi Dengan Menggunakan *Hierarchical Agglomerative Clustering*. Pada Ruang Baca yang berada di Fakultas Ilmu Komputer dan Perpusatakaan Pusat Universitas Brawijaya terdapat suatu masalah yaitu tidak adanya pengkategorian seluruh dokumen skripsi yang disimpan. *Hierarchical Agglomerative Clustering* (HAC) mengelompokkan dokumen secara iterative mulai dari *cluster* terkecil hingga 1 *cluster* terbesar. Data yang digunakan yaitu judul dari dokumen skripsi Teknik Informatika Universitas Brawijaya. Nilai *cosine distance* digunakan untuk memperoleh nilai kemiripan antar dokumen. Pilihan jarak yang digunakan sebagai parameter pada proses *clustering* adalah *single linkage, complete linkage* dan *average linkage*. Diperoleh hasil pengujian dari 100 dokumen yang digunakan yaitu nilai *Silhouette Coefficient* dari *single*

linkage yaitu 0,10125, *complete linkage* yaitu 0,155733 dan *average linkage* yaitu 0,160428. Didapatkan hasil yang lebih baik yaitu *Average linkage*.

Penelitian yang dilakukan oleh Herwanto (2018) mengenai dokumen *clustering* dengan *latent dirichlet allocation* dan *ward hierarichal clustering*. Pada saat ini terdapat sangat banyak konten informasi dalam bentuk berita dari berbagai sumber setiap harinya. Dengan banyaknya konten tersebut menuntut organisasi konten yang baik untuk pencarian informasi yang diinginkan dapat dilakukan dengan mudah maka dilakukan clustering dokumen. Penelitian ini menggunakan kombinasi pemodelan topik latent dirichlet allocation dengan *ward hierarchical clustering*. Pada LDA digunakan untuk representasi vektor dokumen berupa distribusi topik dengan tujuan mengurai dimensi vektor yang biasanya terlalu panjang jika menggunakan *tf-idf*. Didapatkan hasil *silhouette coefficient* yang baik yaitu 0.7. Performa *silhouette coefficient* pada representasi pemodelan topik lebih baik dibandingkan dengan representasi dengan *tf-idf*.

Penelitian yang dilakukan oleh Wirasakti, Permadi, Hartanto, & Hartatik (2020) mengenai Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik. Pada penulisan ini yaitu untuk mengetahui kata kunci yang cocok untuk digunakan dalam sebuah publikasi artikel pada sebuah blog dengan menggunakan model LDA yaitu model probabilitas yang dapat menghasilkan beberapa topik. Data yang diambil yaitu berasal dari blog/artikel. lalu melakukan pemotongan artikel per bagian kemudian dilakukan preprocessing data, dan melakukan pengubahan vektor data menjadi corpus untuk dimodelkan dengan LDA, yang kemudian dilakukan *clustering* dengan menggunakan *K-Means*. Didapatkan hasil model LDA yaitu 4 topik dengan 8 kata. Yang memiliki nilai probabilitas tertinggi diantaranya mesin (0.09375857), maksimal (0.064600445), mazda (0.10009629), varian (0.07572112), cx-8 (0.10170187), mazda (0.101048954), mobil (0.09820121), dan mazda (0.05679208).

Pada referensi di atas kemudian dijadikan dalam bentuk **Tabel 2.1** dibawah ini yang merupakan tabel rangkuman perbandingan penelitian terdahulu dengan penulisan ini tentang *Topic Modeling* menggunakan *Latent Dirichlet Allocation*

(LDA) pada abstrak dokumen skripsi mahasiswa Statistika Universitas Islam Indonesia.

Tabel 2.1 Perbandingan Dengan Penulis Terdahulu

No	Penulis	Judul	Metode	Persamaan	Perbedaan
1	Rahmawati, Sihwi, & Suryani (2014)	Analisa *Clustering* Menggunakan Metode *K-Means* dan *Hierarchical Clustering* (Studi Kasus: Dokumen Skripsi Jurusan Kimia, Fmipa, Universitas Sebelas Maret)	*K-Means* dan *Hierarchical Clustering*	Sama-sama menggunakan dokumen skripsi	Penulis sekarang menggunakan metode *Latent Dirichlet Allocation*
2	Prilianti & Wijaya (2014)	Aplikasi *Text Mining* untuk Automasi Penentuan Tren Topik Skripsi dengan Metode *K-Means Clustering*	*K-Means Clustering*	Sama-sama menggunakan dokumen skripsi	Penulis sekarang menggunakan metode *Latent Dirichlet Allocation*

No	Penulis	Judul	Metode	Persamaan	Perbedaan
3	Putra & Kusumawardani (2017)	Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation* (LDA)	*Latent Dirichlet Allocation* (LDA)	Sama-sama menggunakan *Latent Dirichlet Allocation*	Studi kasus penulis sekarang yaitu abstrak dokumen skripsi
4	Februariyanti & Santoso (2017)	*Hierarchical Agglomerative Clustering* Untuk Pengelompokan Skripsi Mahasiswa	*Hierarchical Agglomerative Clustering*	Sama-sama menggunakan dokumen skripsi	Penulis sekarang menggunakan metode *Latent Dirichlet Allocation*
5	Agustina (2017)	Analisis dan visualisasi suara pelanggan pada pusat layanan pelanggan dengan pemodelan topik	*Latent Dirichlet Allocation* (LDA)	Sama-sama menggunakan *Latent Dirichlet Allocation*	Studi kasus penulis sekarang yaitu abstrak dokumen skripsi

No	Penulis	Judul	Metode	Persamaan	Perbedaan
		menggunakan *Latent Dirichlet Allocation* (LDA) studi kasus: PT. Petromkimia Gresik.			
6	Wicaksana, Adikara, & Adinugroho (2018)	*Clustering* Dokumen Skripsi Dengan Menggunakan *Hierarchical Agglomerative Clustering*	*Hierarchical Agglomerative Clustering*	Sama-sama menggunakan dokumen skripsi	Penulis sekarang menggunakan metode *Latent Dirichlet Allocation*
7	Herwanto (2018)	Dokumen *Clustering* Dengan *Latent Dirichlet Allocation* dan *Ward Hierarichal Clustering*	LDA dan *Ward Hierarichal Clustering*	Sama-sama menggunakan *Latent Dirichlet Allocation*	Studi kasus penulis sekarang yaitu abstrak dokumen skripsi
8	Wirasakti, Permadi, Hartanto, &	Pembuatan Kata Kunci Otomatis Dalam Artikel	LDA dan *K-Means Clustering*	Sama-sama menggunakan *Latent*	Studi kasus penulis sekarang yaitu

No	Penulis	Judul	Metode	Persamaan	Perbedaan
	Hartatik (2020)	Dengan Pemodelan Topik		*Dirichlet Allocation*	abstrak dokumen skripsi

Dengan beberapa penelitian yang telah dilakukan oleh peneliti terdahulu, maka penulis sekarang melakukan penelitian mengenai dokumen abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia Tahun 2011-2015 menggunakan pemodelan topik dengan *Latent Dirichlet Allocation* (LDA).

BAB III
LANDASAN TEORI

3.1 Skripsi

Skripsi yang bisa dikatakan sebagai karya tulis ilmiah merupakan sebuah paparan tulisan dari hasil penelitian calon sarjana S1 yang membahas suatu permasalahan di bidang ilmu tertentu dengan menggunakan kaidah-kaidah yang berlaku. Skripsi juga merupakan salah satu syarat yang digunakan oleh perguruan tinggi agar mahasiswa dapat memperoleh gelar sarjana strata 1 (Sari, Windarto, Hartama, & Solikhun, 2018). Pada penyusunan skripsi, terdapat beberapa bab yang harus disusun. Namun tidak hanya bab-bab yang harus disusun namun ada Bagian penting yang masuk dalam penyusunan skripsi tersebut yaitu abstrak.

Abstrak merupakan bagian paling penting dalam makalah ilmiah, artikel atau laporan penelitian, dimana memungkinkan pembaca dapat dengan mudah mengidentifikasi secara cepat dan akurat pada isi dasar dari dokumen. Dalam sebuah dokumen, kemudian untuk pembaca memutuskan harus membaca atau tidak keseluruhan dokumen tersebut, menentukan relevansi dari dokumen tersebut dengan kepentingan mereka, memungkinkan dengan menggunakan abstrak dari dokumen tersebut. Maka abstrak wajib disertai untuk masing-masing makalah ilmiah. Untuk menulis ringkasan singkat dari keseluruhan kandungan dokumen dalam kalimat yang mampu mewakili keseluruhan isi dokumen dengan jelas tentu saja tidak begitu mudah. Abstrak harus berisi informasi spesifik yang cukup dimana dapat memenuhi kebutuhan pembaca yang sedang mencari sumber informasi terkait (Nasution, 2017).

3.2 Analisis Deskriptif

Analisis deskriptif merupakan suatu metode analisis yang pada umumnya menampilkan statistik melalui sebuah perhitungan matematis yang berfungsi untuk mengetahui dari gambaran penyebaran data sampel atau populasi. Sehingga, data akan lebih mudah dipahami dan lebih informatif. Pada analisis deskriptif yang

biasanya ditampilkan diantaranya seperti rata-rata, mean, modus, dan statistik lainnya. Umumnya, analisis deskriptif sering mengkaji tentang berbagai hal terkait aktivitas yang menonjol, kewajaran suatu kejadian maupun hubungan antar variabel data (Wahyudin, Tosida, & Andria, 2019).

3.3 *Text Mining*

Text mining merupakan suatu teknik yang digunakan untuk menangani dari pengelompokan, klasifikasi, ekstraksi informasi dan pengambilan informasi (Hudaya, Fakhrurroja, & Alamsyah, 2019). *Text Mining* merupakan suatu analisis data yang terdapat bahasa alami dengan menggunakan teknik dan alat untuk merancang, menemukan dan mengekstrak pengetahuan pada data yang tidak terstruktur. Pada data yang tidak terstruktur sehingga dapat menjadi data dengan topik yang lebih terstruktur dan dapat lebih mudah dianalisis dengan penambangan teks berfungsi dengan mengubah kata atau kalimat (Kabiru & Sari, 2019). Pada proses *text mining* terdapat empat tahap yaitu *text preprocessing* (pemrosesan awal terhadap teks), *text transformation* (transformasi teks), *feature selection* (pemilihan fitur), dan *pattern discovery* (penemuan pola) (Sanjaya & Absar, 2015).

3.4 *Preprocessing*

Preprocessing merupakan suatu proses awal pada klasifikasi dokumen dengan tujuan untuk menyiapkan data agar menjadi terstruktur. Hasil yang diperoleh dari *preprocessing* akan berupa nilai numerik sehingga dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut (Gusti. A, Akbar, & Akbar, 2016).

Pada tahap *preprocessing* dimana tahap yang sangat penting dilakukan untuk menghilangkan kata-kata dalam dokumen yang tidak dibutuhkan. Tahap ini berpengaruh yang dapat memberikan hasil setelah dilakukan pengolahan data teks tersebut. Jika tahap ini dilakukan dengan baik maka hasil dari tahap *preprocessing* akan mendapatkan hasil yang baik pula. Pada proses *preprocessing* terdiri dari beberapa proses yaitu *case folding*, *Remove punctuation, stopword,* dan *tokenizing*.

3.3.1. *Case Folding*

Tahap awal yang dilakukan dalam *preprocessing* yaitu *case folding* dimana pada proses ini akan mengubah karakter dari huruf besar menjadi huruf kecil. Kemudian karakter yang akan diterima hanya "a" hingga "z" (Gusti. A, Akbar, & Akbar, 2016). Dalam hal ini dengan tujuan untuk dapat menghindari adanya dua kata yang sama namun dianggap berbeda oleh program yang dikarenakan perbedaan huruf kapital dan huruf kecil (Bagus C.W, 2017).

3.3.2. *Remove Punctuation*

Remove punctuation merupakan proses yang dilakukan untuk membuang karakter yang tidak digunakan berupa tanda baca, angka, markup/html/tag, spesial karakter ($, %, &, stc) (Cendana & Permana, 2019) atau yang biasa dikenal dengan istilah *noise*. *Noise* yaitu suatu bentuk data yang dimana nantinya akan mengganggu pada proses pengolahan tersebut (Bagus C.W, 2017).

3.3.3. *Stopwords*

Stopwords merupakan kata-kata yang terdapat dalam dokumen yang tidak bermakna sehingga dapat dihilangkan (Gusti. A, Akbar, & Akbar, 2016).

3.3.4 *Tokenizing*

Tokenizing merupakan membagi suatu teks atau pemotongan berdasarkan dari tiap kata yang menyusunnya (Gusti. A, Akbar, & Akbar, 2016). Dengan tujuan yaitu agar dalam proses yang selanjutnya menjadi lebih mudah yang mengenai penghitungan kata, pembobotan kata sampai dengan transformasi kedalam bentuk vektor yang berdimensi tinggi (Bagus C.W, 2017).

3.5 Pembobotan *Term Frequency-Invers Document Frequency* (TF-IDF)

Metode TF-IDF adalah suatu cara yang digunakan untuk memberikan bobot hubungan pada suatu *term* (kata) terhadap suatu dokumen, yang merupakan sebuah gabungan dari dua konsep perhitungan bobot yaitu gabungan dari metode *term*

frequency (tf)/frekuensi kemunculan dari sebuah kata pada sebuah dokumen tertentu dan metode *inverse document frequency* (idf)/*inverse* dari frekuensi dokumen yang mengandung kata tersebut (Karmayasa, 2012).

Pembobotan dasar yang dilakukan dalam melakukan *term weighting* (pembobotan kata) yaitu menghitung frekuensi kemunculan kata dalam dokumen. Frekuensi kemuculan (TF) yaitu menunjukkan sejauh mana kata tersebut mewakili dari isi dokumen. Kemudian semakin besar dari kemunculan suatu kata (*term*) pada dokumen maka akan memberikan nilai kesesuaian yang semakin besar. Pada pembobotan juga akan memperhitungkan faktor dari kebalikan frekuensi dokumen yang mengandung suatu kata (IDF) (Karmayasa, 2012).

Pada *Term Frequency* terbagi menjadi beberapa jenis algoritma yaitu (Yoren, 2018):

- *Raw Term Frequency* (TF Murni), yaitu nilai TF yang didapatkan dengan menghitung frekuensi kemunculan suatu term pada dokumen. Jika muncul kata sebanyak 4 kali maka kata tersebut akan bernilai 4.
- *Binary Term Frequency* (TF Binary), yaitu penyeragaman bobot pada dokumen dimana jika muncul minimal satu kali dalam dokumen maka memiliki nilai 1 dan jika kata tidak muncul sama sekali maka memiliki nilai 0.
- *Logarithmic Term Frequency* (TF Logaritmik), yaitu mengurangi tingkat kepentingan kemunculan kata dalam menghitung bobot dokumen terhadap suatu kata dengan melakukan log pada TF yang dapat diperoleh dari persamaan berikut:

  TF = 1+Log(TF) (0.1)
- *Augmented Term Frequency* (TF Normalisasi) , yaitu nilai TF dengan menggunakan perbandingan antara frekuensi sebuah kata dengan jumlah keseluruhan kata pada dokumen.

  $\mathrm{TF} = 0.5 + 0.5\, x \left(\frac{TF}{maxTF}\right)$ (0.2)

  Kemudian untuk mengitung nilai IDF yaitu sebagai berikut:

  $\mathrm{IDF_j} = \ln \frac{D}{DF_j}$ (0.3)

Keterangan :

D : jumlah semua dokumen yang ada dalam koleksi

DF_j : jumlah dokumen yang mengandung term (j)

Rumus pada metode TF dan IDF merupakan gabungan dari formula perhitungan *row* TF dan Formula IDF. Pada penelitian ini jenis formula tf yang digunakan yaitu *Raw Term Frequency* (TF Murni), Sehingga jika digunakan dalam setiap kata pada setiap dokumen akan dihitung bobotnya dengan menggunakan rumus sebagai berikut:

$$W_{ij} = TF_{ij} * IDF_j \quad (0.4)$$

$$W_{ij} = TF_{ij} * \ln\left(\frac{D}{DF_j}\right) \quad (0.5)$$

Keterangan :

i = dokumen ke-d

j = kata ke-t dari kata kunci

W_{ij} = bobot *term* (j) terhadap dokumen (i)

TF_{ij} = jumlah kemunculan *term* (j) dalam dokumen (i)

Dapat dilihat pada persamaan , didapatkan hasil bahwa berapapun besarnya nilai TF_{ij} jika nilai D = DFj maka akan didapatkan hasil 0 pada perhitungan IDF. Dengan begitu akan ditambahkan 1 pada sisi IDF, Sehingga perhitungan bobotnya dirumuskan menjadi sebagai berikut:

$$W_{ij} = TF_{ij} * \left(\ln\left(\frac{D}{DF_j}\right) + 1\right) \quad (0.6)$$

Fungsi dari metode TF-IDF ini yaitu untuk mencari representasi nilai dari setiap dokumen dalam koleksi yang dimiliki. Kemudian akan dibentuk suatu vektor antara dokumen dan *term* yang ditentukan berdasarkan nilai bobot *term* dalam dokumen. Semakin besar dari nilai perhitungan bobot yang didapatkan maka akan semakin tinggi pula tingkat similaritas dokumen terhadap *term* (Fattah, 2016).

3.6 *Topic Modeling*

Konsep *topic modeling* menurut Blei (2003) terdiri dari entitas-entitas yaitu "kata", "dokumen", dan "*corpora*". "Kata" dianggap sebagai unit dasar dari data diskrit dalam dokumen, didefinisikan sebagai item dari kosa kata yang diberi indeks untuk setiap kata unik pada dokumen. "Dokumen" adalah susunan N kata-kata.

Sebuah *corpus* adalah kumpulan M dokumen dan *corpora* merupakan bentuk jamak dari *corpus*. Sementara "*topic*" adalah distribusi dari beberapa kosakata yang bersifat tetap. Secara sederhana, setiap dokumen dalam *corpus* mengandung proporsi tersendiri dari topik-topik yang dibahas sesuai kata-kata yang terkandung di dalamnya. *Topic modeling* merupakan sebuah topik yang terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen memiliki probabilitas masing-masing dari beberapa topik yang dihasilkan. Secara sederhana, *Topic Modeling* adalah algortima yang memiliki tujuan untuk menemukan suatu topik yang tersembunyi dari rangkaian kata pada dokumen yang tidak terstruktur. Untuk menemukan topik yang berada antara teks tersebut dengan algoritma *Topic Modeling* menganalisis dari teks asli, bagaimana topik-topik dapat saling terhubung satu dengan yang lain, bagaimana tema-tema bisa berubah dari waktu ke waktu, sehingga bisa dikembangkan untuk pencarian, atau dengan meringkas teks yang terdapat dalam dokumen. (Putra & Kusumawardani, 2017).

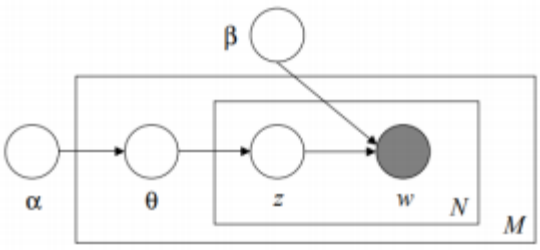
Topic Modeling merupakan dokumen teks yang terdiri dari kata-kata, topik yang dapat dituliskan dalam banyak dokumen dapat dinyatakan dengan kombinasi kata-kata yang saling terkait dan teknik yang dapat digunakan untuk menyimpulkan suatu topik yang tersembunyi dalam sebuah dokumen teks. Karena *topic modeling* ini mewakili dari setiap dokumen sebagai kombinasi kompleks dari beberapa topik dan setiap topiknya sebagai kombinasi kompleks dari beberapa kata, kemudian juga digunakan sebagai alat *text mining* untuk mengklasifikasikan sebuah dokumen berdasarkan hasil kesimpulan topik (Nugroho & Alamsyah, 2018). Salah satu pemodelan topik yang paling populer yaitu *Latent Dirichlet Allocation* (LDA) (Putra & Kusumawardani, 2017).

3.7 *Latent Dirichlet Allocation (*LDA)

Latent Dirichlet Allocation merupakan salah satu metode yang dapat dipilih dalam melakukan analisis untuk dokumen yang memiliki ukuran sangat besar. LDA itu sendiri bisa digunakan untuk meringkas, melakukan klasterisasi, menghubungkan atau memproses data yang sangat besar dikarenakan LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen.

Distribusi yang digunakan yaitu distribusi Dirichlet, yang digunakan untuk memperoleh distribusi topik per-dokumen, dalam proses generatif, hasil yang didapatkan dari Dirichlet digunakan untuk mengalokasikan kata-kata dalam dokumen untuk topik yang berbeda. Pada LDA, dokumen-dokumen adalah objek yang bisa diamati, namun topik, distribusi topik per-dokumen, penggolongan setiap kata untuk topik per-dokumen adalah struktur tersembunyi. Menurut Blei (2003), LDA merupakan model probabilistik generatif dari kumpulan tulisan yang dapat disebut corpus. Ide dasar dari metode LDA yaitu setiap dokumen direpresentasikan sebagai campuran acak atas topik yang tersembunyi, dimana setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang terdapat didalamnya (Putra & Kusumawardani, 2017).

Blei merepresentasikan metode LDA sebagai model *probabilistic* secara visual seperti pada **Gambar 3.1** berikut.



Gambar 3.1 Model Representasi LDA (Blei, et al., 2003)

Dapat dilihat dari **Gambar 3.1** diatas yaitu representasi metode LDA menurut (Blei, et al., 2003) dimana terdapat tingkatan pada pemodelan dengan LDA. Parameter α dan β yaitu parameter distribusi topik yang berada pada tingkatan *corpus*, adalah kumpulan dari M dokumen. Untuk parameter α yang digunakan dalam menentukan distribusi topik dokumen, jika nilai alpha semakin besar dalam suatu dokumen, menandakan bahwa campuran topik yang dibahas dalam dokumen semakin banyak. Untuk parameter β yang digunakan untuk menentukan distribusi kata dalam topik. Jika nilai beta semakin tinggi, maka semakin banyak kata-kata yang terdapat di dalam topik, namum jika nilai beta semakin kecil, maka semakin sedikit kata-kata yang terdapat di dalam topik sehingga topik tersebut mengandung kata-kata yang lebih spesifik. pada variabel θ_m yaitu variabel yang berada di tingkat dokumen (M). Variabel θ merepresentasikan distribusi topik untuk dokumen

tersebut. Jika nilai θ semakin tinggi, maka semakin banyak topik yang terdapat di dalam dokumen, jika nilai θ semakin kecil, maka semakin spresifik pada topik tertentu. Pada variabel Z_n dan W_n yaitu variabel tingkat kata (N). Variabel Z merepresentasikan topik dari kata tertentu pada sebuah dokumen, pada variabel W merepresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen. Berdasarkan penjelasan notasi sebelumnya, proses generatif pada LDA akan berkorespondensi pada *joint distribution* dari variabel yang tersembunyi dan variabel yang terobsesi. Berikut merupakan perhitungan probabilitas dari sebuah *corpus* berdasarkan notasi yang telah dijelaskan (Putra & Kusumawardani, 2017).

$$p(D|\,\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\,\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\,\theta_d) p(w_{dn}|\,Z_{dn},\beta) \right) d\theta_d \quad (0.7)$$

Dapat dilihat bahwa pada notasi β mendeskripsikan topik, dimana pada setiap β merupakan distribusi dari sejumlah kata. Pada Variabel θ_d adalah variabel level dokumen dengan satu kali sampel per dokumen yang merepresentasikan proporsi topik untuk dokumen ke *d*. Pada notasi z_{dn} dan w_{dn} merupakan representasi variabel di level kata dengan satu kali sampel untuk masing-masing kata pada setiap dokumen.

3.8 *Topic Coherence*

Topic modeling membahas mengenai kumpulan dari sebuah kata-kata dari sebuah dokumen ataupun *corpus*. Berdasarkan dari kata-kata yang terdapat dalam dokumen yang digunakan, penggalian dari relasi topik dilakukan dengan asumsi bahwa pada satu dokumen meliputi suatu set kecil dari topik yang ringkas, dimana topik-topik ini perlu dikorelasikan dengan interpretasi manusia. Pada penelitian ini akan menggunakan validasi topik dengan menggunakan *coherence topic* (Putra, 2017).

Topic coherence yaitu dimana satu set dari kata-kata yang dihasilkan pada topik model dengan dinilai berdasarkan tingkat koherensi atau dalam diinterpretasi oleh manusia dengan tingkat kemudahannya. *Topic Coherence* mengukur nilai dari suatu topik dengan mengukur tingkat kesamaan semantik antara kata-kata yang ada dalam topik. Pengkukuran ini dapat membantu dalam membedakan antara topik

yang dapat diinterpretasi secara semantik dengan topik yang memiliki keterkaitan secara statistik (Putra, 2017). *Topic Coherence* merupakan suatu ukuran yang akan digunakan untuk mengevaluasi *Topic Modeling*, dimana jika *coherence skor* topik yang tinggi maka model yang dihasilkan tersebut yang baik. Menurut Wisdom (2017) *Topic Coherence* dapat dianggap memberikan kemampuan interpretasi lebih baik terhadap hasil dari *Topic Modeling* dibandingkan dengan *Perplexity*. Namun hasil dari matriks *perplexity* terkadang tidak memiliki korelasi yang baik pada interpretasi model oleh manusia (Listari, 2019).

3.9 *Multidimensional Scalling*

Multidimensional Scalling (MDS) merupakan salah satu kategori yang merupakan metode interdependensi dalam analisis multivariat. MDS merupakan suatu teknik yang digunakan untuk mencari hubungan dari antar data secara spasial. Pada hubungan dari analisis MDS akan membentuk sebuah grafik (*map*) untuk menggambarkan dari posisi sebuah objek dengan objek lain berdasarkan dari *similarity* (kemiripan) dari objek-objek tersebut (Pradita, Satyahadewi, & Perdana, 2019). Kegunaan dari MDS yaitu untuk menyajikan objek-objek dengan visual berdasarkan dengan kemiripan yang dimiliki. Kegunaan yang lain yaitu mengelompokkan objek-objek yang memiliki kemiripan yang dilihat berdasarkan beberapa peubah atau atribut yang dianggap dapat mengelompokkan objek-objek tersebut (Irmawati, 2017).

3.10 *Principal Component Analysis*

Principal component analysis (PCA) merupakan suatu teknik statistik yang digunakan untuk mengurangi dimensi dari dataset yang jumlahnya banyak menjadi variabel-variabel yang saling terhubung. PCA secara khusus melakukan ekstraksi faktor-faktor yang memiliki korelasi tinggi dengan mempertahankan dari karakter aslinya. Maka pada akhirnya akan didapatkan variabel yang tidak memiliki korelasi dengan satu sama lain. Pada dasarnya prosedur dari PCA memiliki tujuan untuk menyederhanakan variabel yang diamati dengan cara mereduksi atau menyusutkan dimensinya, yang dilakukan dengan cara menghilangkan korelasi diantara variabel bebas dengan melakukan transormasi variabel bebas asal ke variabel baru yang

tidak berkorelasi sama sekali atau yang disebut *principal component*. Setelah beberapa komponen hasil PCA yang bebas dari multikolinearitas didapatkan, maka komponen-komponen tersebut akan menjadi variabel bebas baru yang diregresikan atau dianalisis pengaruhnya terhadap sebuah variabel tak bebas dengan melakukan analisis regresi (Koesriputranto, 2015).

BAB IV
METODOLOGI PENELITIAN

4.1 Populasi dan Sampel

Populasi pada penelitian ini yaitu semua dokumen abstrak pada skripsi mahasiswa yang sudah selesai di jurusan Statistika Universitas Islam Indonesia. Sampel yang digunakan pada penelitian ini yaitu Dokumen abstrak pada skripsi Mahasiswa jurusan Statistika Universitas Islam Indonesia tahun angkatan 2011-2015.

4.2 Jenis dan Sumber Data

Jenis data yang digunakan dalam penelitain ini adalah data sekunder. Data yang digunakan hanya yang diperoleh dari *website* yang dimiliki oleh Universitas Islam Indonesia yaitu https://dspace.uii.ac.id/ dan data diperoleh dari perpustakaan UII pada tanggal 26 dan 30 Desember 2019.

4.3 Variabel Penelitian

Pada penelitian ini digunakan variabel, diantaranya yaitu Tahun, Abstrak, Sk Dosen Pembimbing, Tanggal Pendadaran, dan Lama TA. Berikut ini (**Tabel 4.1**) merupakan penjelasan dan definisi operasional variabelnya:

Tabel 4.1 Definisi Operasional Variabel

Variabel	Definisi Operasional Variabel
Tahun	Tahun pada saat mahasiswa masuk kuliah
Abstrak	Abstrak merupakan ringkasan singkat dari keseluruhan kandungan dokumen dalam kalimat yang mampu mewakili keseluruhan isi dokumen dengan jelas.
Sk Pembimbing Skripsi	Waktu/tanggal dikeluarkannya surat keputusan mengenai pengangkatan dosen pembimbing skripsi
Tanggal Pendadaran	Waktu pada saat mahasiswa melakukan sidang skripsi

Variabel	Definisi Operasional Variabel
Lama Penyusunan TA	Waktu mahasiswa dalam mengerjakan skripsi yang diperoleh dari : (Tanggal Pendadaran – SK Pembimbing Skripsi) / 30)

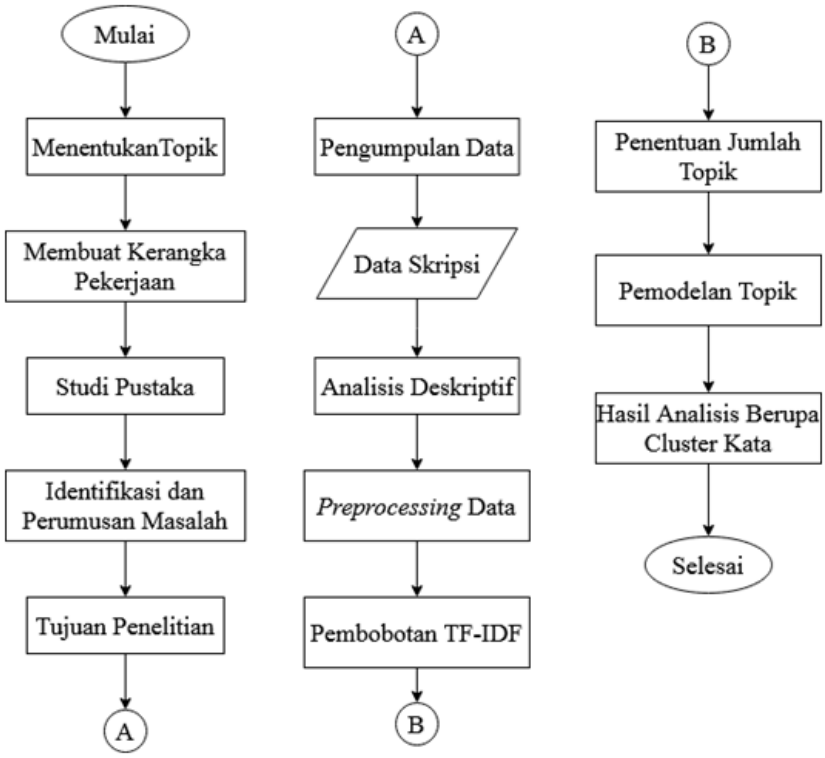
4.4 Metode Analisis Data

Software yang digunakan dalam analisis ini adalah *Microsoft Excel* 2007 yang digunakan untuk sebagai penyimpanan data, pengolahan diagram dan Anaconda 3.5.3.1. Adapun metode analisis data yang digunakan sebagai berikut:

1. Analisis Deskriptif untuk mengetahui gambaran umum dari abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia angkatan 2011-2015.
2. Pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) pada data abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia angkatan 2011-2015.

4.5 Tahapan Analisis / Diagram Alir *Topic Modeling* dengan *Latent Dirichlet Allocation* (LDA)

Berikut ini merupakan langkah-langkah yang dilakukan dalam penelitian ini mulai dari pengambilan data hingga tahap analisis.



Gambar 4.1 Tahapan Analisis Penelitian

Tahapan yang dilakukan oleh peneliti setelah melakukan analisis deskriptif yaitu melakukan analisis *Topic Modeling* dengan *Latent Dirichlet Allocation* (LDA). Tahapan yang dilakukan yang dapat dilihat pada **Gambar 4.1** diatas.

Pada analisis *Topic Modeling* dengan *Latent Dirichlet Allocation*, peneliti dalam melakukan analisis, *packages* yang digunakan dari python diantaranya yaitu *packages Gensim* dan *pyLDAvis.* Penjelasan mengenai *packages* yaitu sebagai berikut:

1. *Gensim*

Gensim merupakan *library* dari *python* untuk implementasi LDA. *Gensim* ini dirancang yang secara otomatis mengekstraksi topik semantik dari dokumen dengan seefisien mungkin menurut Rehurek (2019) (Listari, 2019). *Gensim* memiliki kemampuan dapat melakukan pemrosesan pada data teks mentah. *Gensim* juga dapat menyediakan beberapa algoritma di dalamnya diantaranya yaitu *Latent Dirichlet Allocation* (LDA). Menurut Rehurek (2017) terdapat beberapa fitur yang tersedia dari *Gensim* yaitu sebagai berikut (Listari, 2019):

- *Memory Independence*, seluruh data *corpus* tidak perlu di alokasikan pada RAM atau pada memori karena dalam satu waktu pada saat proses *training* terhadap data *corpus* tersebut dilakukan.
- Mengimplementasikan secara efisien yang terkait dengan algoritma ruang vektor yang populer salah satunya yaitu *Term Frequency-Invers Document Frequency* (TF-IDF).
- Memiliki kemampuan untuk melakukan *similarity query* terhadap dokumen dalam representasi semantiknya.

Pada penelitian ini secara khusus menggunakan beberapa fitur dari *Gensim* yaitu *gensim.corpora* dan *gensim.model.* Pada modul *gensim.corpora* yang digunakan untuk membangun *dictionary* dari data teks sebelum melakukan proses pembuatan model LDA dengan memanggil modul *Dictionary.* Kemudian modul *gensim.model* yang digunakan untuk membangun model LDA dengan memanggil dari modul *LdaModel.* Kemudian akan dipanggil pada saat melalukan proses perhitungan *coherence* model dengan meng-*import* modul *CoherenceModel*.

2. *PyLDAvis*

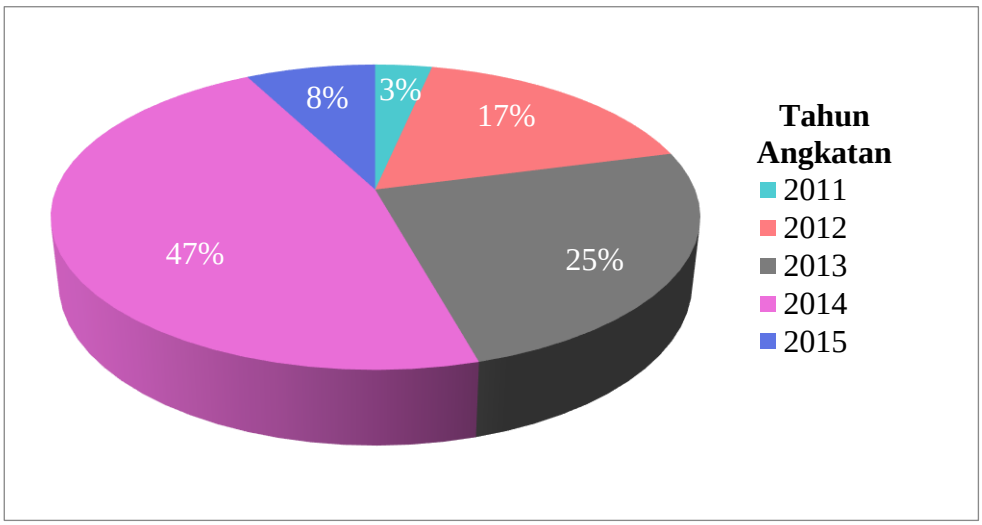
PyLDAvis merupakan *library* yang terdapat pada *Python* untuk visualisasi model topik interaktif. Pada *PyLDAvis* ini dirancang untuk dapat membantu pengguna menafsirkan topik dalam model topik yang sesuai dengan kumpulan data teks. *Packages* ini juga mengekstraksi informasi dari model topik LDA yang dipasang untuk dapat menginformasikan visualisasi berbasis *web* interaktif. Visualisasi yang dimaksudkan yaitu untuk digunakan dalam *notebook IPython* namun dapat juga disimpan dalam *file* HTML yang berdiri sendiri untuk memudahkan berbagi (pypi.org, 2018).

BAB V
HASIL DAN PEMBAHASAN

5.1 Analisis Deskriptif

Analisis deskriptif pada penelitian ini yaitu untuk mengetahui gambaran umum dari abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia tahun 2011-2015 yang meliputi jumlah Skripsi Mahasiswa berdasarkan tahun dan rata-rata lama pengerjaan TA setiap tahunnya.

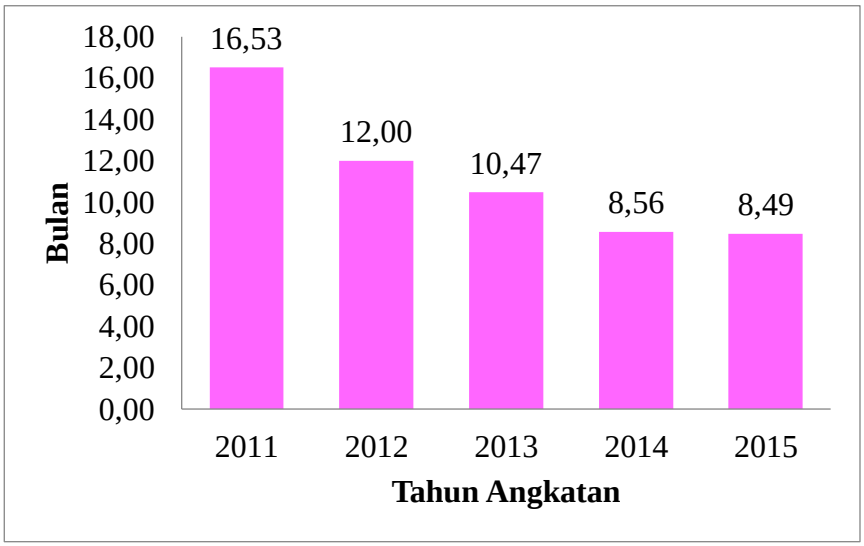
Berikut ini merupakan jumlah Skripsi Mahasiswa berdasarkan tahunnya.



Gambar 5.1 Jumlah Abstrak Dokumen Skripsi Mahasiswa

Berdasarkan **Gambar 5.1** diatas yaitu jumlah skripsi Mahasiswa berdasarkan tahunnya yang digunakan dalam penelitian ini. Didapatkan informasi bahwa proporsi jumlah skripsi mahasiswa berdasarkan tahunnya yang digunakan dalam penelitian ini dengan proporsi terbesar pada tahun angkatan 2014 sebesar 47% dokumen dan proporsi terkecil pada tahun angkatan 2011 sebesar 3% dokumen. Hal ini dikarenakan jumlah mahasiswa terdaftar pada angkatan 2014 adalah sebanyak 279 mahasiswa.

Berikut ini merupakan histogram rata-rata lama pengerjaan TA setiap tahunnya.



Gambar 5.2 Rata-rata Lama Pengerjaan TA

Berdasarkan **Gambar 5.2** diatas yaitu rata-rata lama pengerjaan TA berdasarkan tahunnya yang digunakan dalam penelitian ini. Didapatkan informasi bahwa rata-rata lama pengerjaan TA pada tiap tahunnya semakin kecil atau pada grafik **Gambar 5.2** semakin menurun, yang dapat diartikan bahwa lama pengerjaan TA tiap tahun angkatannya semakin baik.

Tahap berikutnya adalah melakukan *topic modeling* LDA dengan menggunakan abstrak dari dokumen skripsi.

5.2 *Preprocessing*

Pada tahap *preprocessing* dimana tahap yang sangat penting dilakukan untuk menghilangkan kata-kata dalam dokumen yang tidak dibutuhkan. Tahap ini merupakan tahap pertama yang dilakukan dalam mulai menganalisis. Berikut merupakan beberapa data yang digunakan.

Tabel 5.1 Data Awal Penulisan

Abstrak
Perkembangan bandara di Indonesia dalam infrastruktur transportasi telah sangat cepat, ini dapat dilihat oleh banyak perusahaan atau yang melayani layanan penerbangan domestik dan internasional. Serta jumlah uang yang dihabiskan di Bandara Internasional Soekarno-Hatta yang menghabiskan waktu setiap hari untuk acara khusus atau beberapa acara dan acara seperti pada Idul Fitri, liburan Natal dan Tahun Baru. Oleh karena itu, analisis peramalan diperlukan untuk

Abstrak
menentukan pengembangan pesawat domestik di Bandara Internasional Soekarno-Hatta pada bulan Desember 2018. Data yang digunakan terdiri dari data sekunder dari Badan Pusat Statistik (BPS), yaitu data jumlah penumpang pesawat di Bandara Soekarno-Hatta. -Hatta International Airport pada Januari 2015 hingga November 2018. Analisis yang digunakan dengan membandingkan dua metode untuk mendapatkan hasil peramalan yang akurat adalah dengan metode Triple Exponential Smoothing dan Fuzzy Time Series Ruey Chyn Tsaur. Hasil penelitian menggunakan metode Ruey Chyn Tsaur Fuzzy Time Series lebih cocok daripada metode Triple Exponential Smoothing, karena memiliki nilai kesalahan kecil MAPE (Mean Absolute Persentase Error) sebesar 4,54% dan MSE (Mean Square Error) dari 9,773. 836.405. Jumlah penumpang pesawat domestik pada bulan Desember 2018 di Bandara Internasional Soekarno-Hatta adalah 1.767.893 penumpang, yang berarti mengurangi jumlah bulan-bulan sebelumnya.
Dalam beberapa tahun terakhir ini, Bitcoin telah menarik banyak perhatian karena sifatrnya yang mendukung teknologi enkripsi dan unit moneter. Bitcoin adalah mata uang elektronik yang memungkinkan pembayarn secara online tanpa melalui lembaga keuangan. Bitcoin menjadi investasi yang menjanjikan bagi para pedagang finansial karena harganya yang fluktuatif berpotensi menghasilkan laba tinggi (semakin tinggi risikonya, semakin tinggi pula pengembaliannya). Tidak seperti stok konvensiaonal, Bitcoin diperdagangkan selama 24 jam sehari tanpa periode tutup, sehingga meningkatkan risiko. Prediksi nilai Bitcoin diharapkan dapat meminimalkan risiko dengan mempertimbangkan beberapa informasi seperti informasi blockchain, faktor ekonomi makro, dan rasio mata uang global. Namun, multikolinearitas diantara variabel-variabel independen menyebabkan metode regresi tidak dapat digunakan. Penelitian ini menggunakan Bayesian Regularization Neural Network (BRNN) yang merupakan asumsi bebas. Metode ini adalah Single Hidden Layer Feed Forward Neural Network (SLNN) yang memanfaatkan konsep Bayesian untuk mengoptimalkan bobot, bias, dan kekuatan koneksi. Data yang digunakan adalah data time series dari 23 Januari 2017, hingga 23 Januari 2019. Regresi dengan subset digunakan untuk mengurangi variabel independen, dari total 25 variabel menjadi 14 variabel. Hasilnya menunjukkan bahwa model BRNN yang dibentuk dapat memprediksi nilai Bitcoin dengan baik, diperoleh nilai prediksi tidak jauh berbeda dari data aktual, dengan akurasi sebesar 91,1% berdasarkan nilai MAPE.

Pada tahap *preprocessing* ini terdiri dari beberapa proses seperti *case folding*, *remove punctuation*, *stopword*, dan *tokenizing*.

5.2.1 *Case Folding*

Case folding dimana pada proses ini akan mengubah karakter dari huruf besar menjadi huruf kecil dengan tujuan jika terdapat kata yang sama namun

penulisannya berbeda huruf kapital atau tidak maka tidak dianggap kata yang berbeda. Hasil dari *case folding* yaitu sebagai berikut.

Tabel 5.2 Hasil *Case Folding*

Abstrak
perkembangan bandara di indonesia dalam infrastruktur transportasi telah sangat cepat, ini dapat dilihat oleh banyak perusahaan atau yang melayani layanan penerbangan domestik dan internasional. serta jumlah uang yang dihabiskan di bandara internasional soekarno-hatta yang menghabiskan waktu setiap hari untuk acara khusus atau beberapa acara dan acara seperti pada idul fitri, liburan natal dan tahun baru. oleh karena itu, analisis peramalan diperlukan untuk menentukan pengembangan pesawat domestik di bandara internasional soekarno-hatta pada bulan desember 2018. data yang digunakan terdiri dari data sekunder dari badan pusat statistik (bps), yaitu data jumlah penumpang pesawat di bandara soekarno-hatta. -hatta international airport pada januari 2015 hingga november 2018. analisis yang digunakan dengan membandingkan dua metode untuk mendapatkan hasil peramalan yang akurat adalah dengan metode triple exponential smoothing dan fuzzy time series ruey chyn tsaur. hasil penelitian menggunakan metode ruey chyn tsaur fuzzy time series lebih cocok daripada metode triple exponential smoothing, karena memiliki nilai kesalahan kecil mape (mean absolute persentase error) sebesar 4,54% dan mse (mean square error) dari 9,773. 836.405. jumlah penumpang pesawat domestik pada bulan desember 2018 di bandara internasional soekarno-hatta adalah 1.767.893 penumpang, yang berarti mengurangi jumlah bulan-bulan sebelumnya.
dalam beberapa tahun terakhir ini, bitcoin telah menarik banyak perhatian karena sifatrnya yang mendukung teknologi enkripsi dan unit moneter. bitcoin adalah mata uang elektronik yang memungkinkan pembayarn secara online tanpa melalui lembaga keuangan. bitcoin menjadi investasi yang menjanjikan bagi para pedagang finansial karena harganya yang fluktuatif berpotensi menghasilkan laba tinggi (semakin tinggi risikonya, semakin tinggi pula pengembaliannya). tidak seperti stok konvensiaonal, bitcoin diperdagangkan selama 24 jam sehari tanpa periode tutup, sehingga meningkatkan risiko. prediksi nilai bitcoin diharapkan dapat meminimalkan risiko dengan mempertimbangkan beberapa informasi seperti informasi blockchain, faktor ekonomi makro, dan rasio mata uang global. namun, multikolinearitas diantara variabel-variabel independen menyebabkan metode regresi tidak dapat digunakan. penelitian ini menggunakan bayesian regularization neural network (brnn) yang merupakan asumsi bebas. metode ini adalah single hidden layer feed forward neural network (slnn) yang memanfaatkan konsep bayesian untuk mengoptimalkan bobot, bias, dan kekuatan koneksi. data yang digunakan adalah data time series dari 23 Januari 2017, hingga 23 Januari 2019. regresi dengan subset digunakan untuk mengurangi variabel independen, dari total 25 variabel menjadi 14 variabel. hasilnya menunjukkan bahwa model brnn yang dibentuk dapat memprediksi nilai bitcoin dengan

Abstrak
baik, diperoleh nilai prediksi tidak jauh berbeda dari data aktual, dengan akurasi sebesar 91,1% berdasarkan nilai mape.

Dapat dilihat pada **Tabel 5.2** diatas merupakan hasil dari proses *case folding* dimana yang diberi garis bawah yaitu pada sebelumnya merupakan karakter huruf besar yang kemudian menjadi huruf kecil semua.

5.2.2 *Remove Punctuation*

Remove punctuation merupakan proses yang dilakukan untuk membuang karakter seperti tanda baca yang tidak digunakan seperti yang terdapat dalam **Tabel 5.3** berikut.

Tabel 5.3 Hasil dari *Remove Punctuation*

Sebelum	Sesudah
perkembangan bandara di indonesia dalam infrastruktur transportasi telah sangat cepat, ini dapat dilihat oleh banyak perusahaan atau yang melayani layanan penerbangan domestik dan internasional. serta jumlah uang yang dihabiskan di bandara internasional soekarno-hatta yang menghabiskan waktu setiap hari untuk acara khusus atau beberapa acara dan acara seperti pada idul fitri, liburan natal dan tahun baru. oleh karena itu, analisis peramalan diperlukan untuk menentukan pengembangan pesawat domestik di bandara internasional soekarno-hatta pada bulan desember 2018. data yang digunakan terdiri dari data sekunder dari badan pusat statistik (bps). yaitu data jumlah penumpang pesawat di bandara soekarno-hatta. -hatta international airport pada januari 2015 hingga november 2018. analisis yang digunakan dengan membandingkan dua metode untuk mendapatkan hasil peramalan	perkembangan bandara di indonesia dalam infrastruktur transportasi telah sangat cepat ini dapat dilihat oleh banyak perusahaan atau yang melayani layanan penerbangan domestik dan internasional serta jumlah uang yang dihabiskan di bandara internasional soekarnohatta yang menghabiskan waktu setiap hari untuk acara khusus atau beberapa acara dan acara seperti pada idul fitri liburan natal dan tahun baru oleh karena itu analisis peramalan diperlukan untuk menentukan pengembangan pesawat domestik di bandara internasional soekarnohatta pada bulan desember data yang digunakan terdiri dari data sekunder dari badan pusat statistik bps yaitu data jumlah penumpang pesawat di bandara soekarnohatta hatta international airport pada januari hingga november analisis yang digunakan dengan membandingkan dua metode untuk mendapatkan hasil peramalan yang akurat adalah dengan metode triple exponential smoothing dan fuzzy time series ruey chyn tsaur hasil penelitian menggunakan metode ruey chyn

Sebelum	Sesudah
yang akurat adalah dengan metode triple exponential smoothing dan fuzzy time series ruey chyn tsaur. hasil penelitian menggunakan metode ruey chyn tsaur fuzzy time series lebih cocok daripada metode triple exponential smoothing. karena memiliki nilai kesalahan kecil mape (mean absolute persentase error) sebesar 4,54% dan mse (mean square error) dari 9,773. 836.405. jumlah penumpang pesawat domestik pada bulan desember 2018 di bandara internasional soekarno-hatta adalah 1.767.893 penumpang. yang berarti mengurangi jumlah bulan-bulan sebelumnya.	tsaur fuzzy time series lebih cocok daripada metode triple exponential smoothing karena memiliki nilai kesalahan kecil mape mean absolute persentase error sebesar dan mse mean square error dari jumlah penumpang pesawat domestik pada bulan desember di bandara internasional soekarnohatta adalah penumpang yang berarti mengurangi jumlah bulanbulan sebelumnya
dalam beberapa tahun terakhir ini, bitcoin telah menarik banyak perhatian karena sifatrnya yang mendukung teknologi enkripsi dan unit moneter. bitcoin adalah mata uang elektronik yang memungkinkan pembayarn secara online tanpa melalui lembaga keuangan. bitcoin menjadi investasi yang menjanjikan bagi para pedagang finansial karena harganya yang fluktuatif berpotensi menghasilkan laba tinggi (semakin tinggi risikonya, semakin tinggi pula pengembaliannya). tidak seperti stok konvensiaonal, bitcoin diperdagangkan selama 24 jam sehari tanpa periode tutup, sehingga meningkatkan risiko. prediksi nilai bitcoin diharapkan dapat meminimalkan risiko dengan mempertimbangkan beberapa informasi seperti informasi blockchain, faktor ekonomi makro, dan rasio mata uang global. namun, multikolinearitas diantara variabel-variabel independen menyebabkan	dalam beberapa tahun terakhir ini bitcoin telah menarik banyak perhatian karena sifatrnya yang mendukung teknologi enkripsi dan unit moneter bitcoin adalah mata uang elektronik yang memungkinkan pembayarn secara online tanpa melalui lembaga keuangan bitcoin menjadi investasi yang menjanjikan bagi para pedagang finansial karena harganya yang fluktuatif berpotensi menghasilkan laba tinggi semakin tinggi risikonya semakin tinggi pula pengembaliannya tidak seperti stok konvensiaonal bitcoin diperdagangkan selama jam sehari tanpa periode tutup sehingga meningkatkan risiko prediksi nilai bitcoin diharapkan dapat meminimalkan risiko dengan mempertimbangkan beberapa informasi seperti informasi blockchain faktor ekonomi makro dan rasio mata uang global namun multikolinearitas diantara variabelvariabel independen menyebabkan metode regresi tidak dapat digunakan penelitian ini menggunakan bayesian

Sebelum	Sesudah
metode regresi tidak dapat digunakan. penelitian ini menggunakan bayesian regularization neural network (brnn) yang merupakan asumsi bebas. metode ini adalah single hidden layer feed forward neural network (slnn) yang memanfaatkan konsep bayesian untuk mengoptimalkan bobot, bias, dan kekuatan koneksi. data yang digunakan adalah data time series dari 23 Januari 2017, hingga 23 Januari 2019. regresi dengan subset digunakan untuk mengurangi variabel independen, dari total 25 variabel menjadi 14 variabel. hasilnya menunjukkan bahwa model brnn yang dibentuk dapat memprediksi nilai bitcoin dengan baik, diperoleh nilai prediksi tidak jauh berbeda dari data aktual, dengan akurasi sebesar 91,1% berdasarkan nilai mape.	regularization neural network brnn yang merupakan asumsi bebas metode ini adalah single hidden layer feed forward neural network slnn yang memanfaatkan konsep bayesian untuk mengoptimalkan bobot bias dan kekuatan koneksi data yang digunakan adalah data time series dari januari hingga januari regresi dengan subset digunakan untuk mengurangi variabel independen dari total variabel menjadi variabel hasilnya menunjukkan bahwa model brnn yang dibentuk dapat memprediksi nilai bitcoin dengan baik diperoleh nilai prediksi tidak jauh berbeda dari data aktual dengan akurasi sebesar berdasarkan nilai mape

Dapat dilihat pada **Tabel 5.3** diatas pada kolom sebelum yang diberi tanda *underline* dan berwarna yang kemudian dihilangkan sehingga pada kolom sesudah sudah tidak terdapat karakter-karakter tersebut.

5.2.3 *Stopwords*

Stopwords merupakan kata-kata yang sering muncul dalam dokumen teks dan kata dalam dokumen tersebut tidak berguna.

Tabel 5.4 Hasil dari *Stopwords*

Sebelum	Sesudah
perkembangan bandara di indonesia dalam infrastruktur transportasi telah sangat cepat ini dapat dilihat oleh banyak perusahaan atau yang melayani layanan penerbangan domestik dan internasional serta jumlah uang yang dihabiskan di bandara internasional soekarnohatta yang menghabiskan waktu setiap hari untuk acara	perkembangan bandara indonesia infrastruktur transportasi cepat perusahaan melayani layanan penerbangan domestik internasional uang dihabiskan bandara internasional soekarnohatta menghabiskan acara khusus acara acara idul fitri liburan natal analisis peramalan menentukan pengembangan pesawat domestik bandara internasional soekarnohatta desember data data

Sebelum	Sesudah
khusus atau beberapa acara dan acara seperti pada idul fitri liburan natal dan tahun baru oleh karena itu analisis peramalan diperlukan untuk menentukan pengembangan pesawat domestik di bandara internasional soekarnohatta pada bulan desember data yang digunakan terdiri dari data sekunder dari badan pusat statistik bps yaitu data jumlah penumpang pesawat di bandara soekarnohatta hatta international airport pada januari hingga november analisis yang digunakan dengan membandingkan dua metode untuk mendapatkan hasil peramalan yang akurat adalah dengan metode triple exponential smoothing dan fuzzy time series ruey chyn tsaur hasil penelitian menggunakan metode ruey chyn tsaur fuzzy time series lebih cocok daripada metode triple exponential smoothing karena memiliki nilai kesalahan kecil mape mean absolute persentase error sebesar dan mse mean square error dari jumlah penumpang pesawat domestik pada bulan desember di bandara internasional soekarnohatta adalah penumpang yang berarti mengurangi jumlah bulanbulan sebelumnya	sekunder badan pusat statistik bps data penumpang pesawat bandara soekarnohatta hatta international airport januari november analisis membandingkan metode hasil peramalan akurat metode triple exponential smoothing fuzzy time series ruey chyn tsaur hasil metode ruey chyn tsaur fuzzy time series cocok metode triple exponential smoothing memiliki nilai kesalahan mape mean absolute persentase error mse mean square error penumpang pesawat domestik desember bandara internasional soekarnohatta penumpang mengurangi bulanbulan
dalam beberapa tahun terakhir ini bitcoin telah menarik banyak perhatian karena sifatrnya yang mendukung teknologi enkripsi dan unit moneter bitcoin adalah mata uang elektronik yang memungkinkan pembayarn secara online tanpa melalui lembaga keuangan bitcoin menjadi investasi yang menjanjikan bagi para pedagang finansial karena harganya yang fluktuatif	bitcoin menarik perhatian sifatrnya mendukung teknologi enkripsi unit moneter bitcoin mata uang elektronik pembayarn online lembaga keuangan bitcoin investasi menjanjikan pedagang finansial harganya fluktuatif berpotensi menghasilkan laba risikonya pengembaliannya stok konvensiaonal bitcoin diperdagangkan jam sehari periode tutup meningkatkan risiko prediksi nilai bitcoin diharapkan meminimalkan risiko

Sebelum	Sesudah
berpotensi menghasilkan laba tinggi semakin tinggi risikonya semakin tinggi pula pengembaliannya tidak seperti stok konvensiaonal bitcoin diperdagangkan selama jam sehari tanpa periode tutup sehingga meningkatkan risiko prediksi nilai bitcoin diharapkan dapat meminimalkan risiko dengan mempertimbangkan beberapa informasi seperti informasi blockchain faktor ekonomi makro dan rasio mata uang global namun multikolinearitas diantara variabelvariabel independen menyebabkan metode regresi tidak dapat digunakan penelitian ini menggunakan bayesian regularization neural network brnn yang merupakan asumsi bebas metode ini adalah single hidden layer feed forward neural network slnn yang memanfaatkan konsep bayesian untuk mengoptimalkan bobot bias dan kekuatan koneksi data yang digunakan adalah data time series dari januari hingga januari regresi dengan subset digunakan untuk mengurangi variabel independen dari total variabel menjadi variabel hasilnya menunjukkan bahwa model brnn yang dibentuk dapat memprediksi nilai bitcoin dengan baik diperoleh nilai prediksi tidak jauh berbeda dari data aktual dengan akurasi sebesar berdasarkan nilai mape	mempertimbangkan informasi informasi blockchain faktor ekonomi makro rasio mata uang global multikolinearitas variabelvariabel independen menyebabkan metode regresi penelitian bayesian regularization neural network brnn asumsi bebas metode single hidden layer feed forward neural network slnn memanfaatkan konsep bayesian mengoptimalkan bobot bias kekuatan koneksi data data time series januari januari regresi subset mengurangi variabel independen total variabel variabel hasilnya model brnn dibentuk memprediksi nilai bitcoin diperoleh nilai prediksi berbeda data aktual akurasi berdasarkan nilai mape

Dapat dilihat pada **Tabel 5.4** diatas pada kolom sebelum yang diberi tanda *underline* dan berwarna yang kemudian dihilangkan sehingga pada kolom sesudah tidak terdapat kata-kata tersebut.

5.2.4 *Tokenizing*

Tokenizing merupakan membagi suatu teks yang dapat menjadi elemen-elemen kecil agar memudahkan pada tahap selanjutnya seperti **Tabel 5.5** berikut.

Tabel 5.5 Hasil dari *Tokenizing*

Sebelum	Sesudah
perkembangan bandara indonesia infrastruktur transportasi cepat perusahaan melayani layanan penerbangan domestik internasional uang dihabiskan bandara internasional soekarnohatta menghabiskan acara khusus acara acara idul fitri liburan natal analisis peramalan menentukan pengembangan pesawat domestik bandara internasional soekarnohatta desember data data sekunder badan pusat statistik bps data penumpang pesawat bandara soekarnohatta hatta international airport januari november analisis membandingkan metode hasil peramalan akurat metode triple exponential smoothing fuzzy time series ruey chyn tsaur hasil metode ruey chyn tsaur fuzzy time series cocok metode triple exponential smoothing memiliki nilai kesalahan mape mean absolute persentase error mse mean square error penumpang pesawat domestik desember bandara internasional soekarnohatta penumpang mengurangi bulanbulan	['perkembangan', 'bandara', 'indonesia', 'infrastruktur', 'transportasi', 'cepat', 'perusahaan', 'melayani', 'layanan', 'penerbangan', 'domestik', 'internasional', 'uang', 'dihabiskan', 'bandara', 'internasional', 'soekarnohatta', 'menghabiskan', 'acara', 'khusus', 'acara', 'acara', 'idul', 'fitri', 'liburan', 'natal', 'analisis', 'peramalan', 'menentukan', 'pengembangan', 'pesawat', 'domestik', 'bandara', 'internasional', 'soekarnohatta', 'desember', 'data', 'data', 'sekunder', 'badan', 'pusat', 'statistik', 'bps', 'data', 'penumpang', 'pesawat', 'bandara', 'soekarnohatta', 'hatta', 'international', 'airport', 'januari', 'november', 'analisis', 'membandingkan', 'metode', 'hasil', 'peramalan', 'akurat', 'metode', 'triple', 'exponential', 'smoothing', 'fuzzy', 'time', 'series', 'ruey', 'chyn', 'tsaur', 'hasil', 'metode', 'ruey', 'chyn', 'tsaur', 'fuzzy', 'time', 'series', 'cocok', 'metode', 'triple', 'exponential', 'smoothing', 'memiliki', 'nilai', 'kesalahan', 'mape', 'mean', 'absolute', 'persentase', 'error', 'mse', 'mean', 'square', 'error', 'penumpang', 'pesawat', 'domestik', 'desember', 'bandara', 'internasional', 'soekarnohatta', 'penumpang', 'mengurangi', 'bulanbulan']
bitcoin menarik perhatian sifatrnya mendukung teknologi enkripsi unit moneter bitcoin mata uang elektronik pembayarn online lembaga keuangan bitcoin investasi menjanjikan pedagang finansial harganya fluktuatif berpotensi menghasilkan laba risikonya pengembaliannya stok konvensiaonal bitcoin diperdagangkan jam sehari periode tutup meningkatkan risiko prediksi nilai bitcoin diharapkan	['bitcoin', 'menarik', 'perhatian', 'sifatrnya', 'mendukung', 'teknologi', 'enkripsi', 'unit', 'moneter', 'bitcoin', 'mata', 'uang', 'elektronik', 'pembayarn', 'online', 'lembaga', 'keuangan', 'bitcoin', 'investasi', 'menjanjikan', 'pedagang', 'finansial', 'harganya', 'fluktuatif', 'berpotensi', 'menghasilkan', 'laba', 'risikonya', 'pengembaliannya', 'stok', 'konvensiaonal', 'bitcoin', 'diperdagangkan', 'jam', 'sehari', 'periode', 'tutup', 'meningkatkan', 'risiko',

Sebelum	Sesudah
meminimalkan risiko mempertimbangkan informasi informasi blockchain faktor ekonomi makro rasio mata uang global multikolinearitas variabelvariabel independen menyebabkan metode regresi penelitian bayesian regularization neural network brnn asumsi bebas metode single hidden layer feed forward neural network slnn memanfaatkan konsep bayesian mengoptimalkan bobot bias kekuatan koneksi data data time series januari januari regresi subset mengurangi variabel independen total variabel variabel hasilnya model brnn dibentuk memprediksi nilai bitcoin diperoleh nilai prediksi berbeda data aktual akurasi berdasarkan nilai mape	'prediksi', 'nilai', 'bitcoin', 'diharapkan', 'meminimalkan', 'risiko', 'mempertimbangkan', 'informasi', 'informasi', 'blockchain', 'faktor', 'ekonomi', 'makro', 'rasio', 'mata', 'uang', 'global', 'multikolinearitas', 'variabelvariabel', 'independen', 'menyebabkan', 'metode', 'regresi', 'penelitian', 'bayesian', 'regularization', 'neural', 'network', 'brnn', 'asumsi', 'bebas', 'metode', 'single', 'hidden', 'layer', 'feed', 'forward', 'neural', 'network', 'slnn', 'memanfaatkan', 'konsep', 'bayesian', 'mengoptimalkan', 'bobot', 'bias', 'kekuatan', 'koneksi', 'data', 'data', 'time', 'series', 'januari', 'januari', 'regresi', 'subset', 'mengurangi', 'variabel', 'independen', 'total', 'variabel', 'variabel', 'hasilnya', 'model', 'brnn', 'dibentuk', 'memprediksi', 'nilai', 'bitcoin', 'diperoleh', 'nilai', 'prediksi', 'berbeda', 'data', 'aktual', 'akurasi', 'berdasarkan', 'nilai', 'mape']

Dapat dilihat pada **Tabel 5.5** diatas pada kolom sebelum dalam bentuk kalimat yang kemudian pada kolom sesudah setiap katanya menjadi terpisah.

5.3 Pembobotan *Term Frequency-Invers Document Frequency* (TF-IDF)

Pada tahap ini merupakan tahap pembobotan yang dimana setelah dilakukan *preprocessing* pada sebelumnya akan dilakukan pengubahan data yang berbentuk kata menjadi dalam bentuk numerik dengan menggunakan pembobotan TF-IDF. Pembobotan TF-IDF adalah gabungan dari metode *Term Frequency* (TF) dengan metode *Inverse Document Frequency* (IDF). Peneliti menggunakan beberapa sampel kata yang akan digunakan untuk menghitung pembobotan tersebut. Berikut merupakan *output* TF yang dihasilkan.

Tabel 5.6 Sampel Hasil dari TF

Dokumen	abad	abdullah	able	…	perakitan	peramalan	…
1	0	0	0	…	0	0	…
2	0	0	0	…	0	0	…
3	0	0	0	…	0	0	…
4	0	0	0	…	0	0	…
5	0	0	0	…	0	2	…
:	:	:	:	:	:	:	…
:	:	:	:	:	:	:	…
413	0	0	0	…	0	0	…
414	0	0	0	…	0	0	…
415	0	0	0	…	0	0	…
416	0	0	0	…	0	0	…

Dapat dilihat pada **Tabel 5.6** diatas merupakan hasil yang diperoleh *output* TF yang merupakan frekuensi kemunculan *term* (t) pada dokumen (D). Jika nilainya 0 maka kemunculan *term* (t) pada dokumen (D) tidak ada, namun jika bernilai 1 maka kemunculan *term* (t) pada dokumen (D) tersebut terdapat 1. Setelah didapatkan hasil TF kemudian dapat mencari hasil dari TF-IDF. Berikut ini merupakan *output* dari perhitungan TF-IDF.

Tabel 5.7 Sampel Hasil dari Perhitungan TF-IDF

Dokumen	abad	abdullah	able	…	perakitan	peramalan	…
1	0	0	0	…	0	0	…
2	0	0	0	…	0	0	…
3	0	0	0	…	0	0	…
4	0	0	0	…	0	0	…
5	0	0	0	…	0	6.58603	…
:	:	:	:	:	:	:	…
:	:	:	:	:	:	:	…
413	0	0	0	…	0	0	…
414	0	0	0	…	0	0	…
415	0	0	0	…	0	0	…
416	0	0	0	…	0	0	…

Dapat dilihat pada **Tabel 5.7** diatas merupakan hasil yang diperoleh *output* perhitungan TF-IDF yang didapatkan dari perkalian nilai TF dengan IDF. Sebelum diperoleh nilai TF-IDF seperti yang terdapat dalam **Tabel 5.7** diatas, maka peneliti harus mendapatkan terlebih dahulu nilai TF dan IDF.

Peneliti menggunakan sampel untuk perhitungan TF-IDF secara manual yaitu kata “peramalan” pada dokumen ke-5. Nilai TF sudah didapatkan pada **Tabel 5.6** diatas, kemudian untuk menghitung nilai IDF diperlukan nilai DF yang merupakan jumlah dokumen yang mengandung *term* (t) atau terdapat berapa banyak dokumen yang mengandung kata "peramalan" yaitu sebanyak 42. Selanjutnya untuk menghitung nilai IDF, dimana IDF merupakan *inverse* dari DF. Dengan menggunakan persamaan untuk perhitungan manual IDF. Jumlah dokumen (D) pada penelitian ini sebanyak 416 dokumen.

$$\text{IDF}_{\text{peramalan}} = \ln\frac{416}{42} = 2.293015642$$

Setelah didapatkan nilai TF dan IDF, kemudian untuk menghitung nilai TF-IDF yaitu dengan melakukan perkalian TF dengan IDF dengan menggunakan **Persamaan 3.6.**

$$W = 2 * \left(\ln\left(\frac{416}{42}\right) + 1\right)$$

$$W = 2 * (2.293015642 + 1)$$

$$W = 2 * (3.293015642)$$

$$W = 6.586031284$$

$$W \approx 6.58603$$

Hasil perhitungan manual dari TF-IDF kata “peramalan” pada dokumen ke-5 (D5) yang telah didapatkan kemudian dirangkum sebagai berikut.

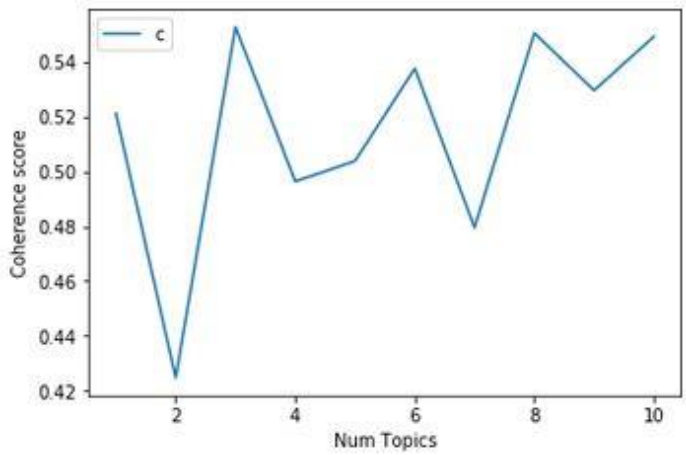
Tabel 5.8 Sampel Hasil dari Perhitungan TF-IDF “peramalan”

Term (t)	TF	DF	IDF	TF-IDF
	D5			D5
peramalan	2	42	2.293015642	6.586031284

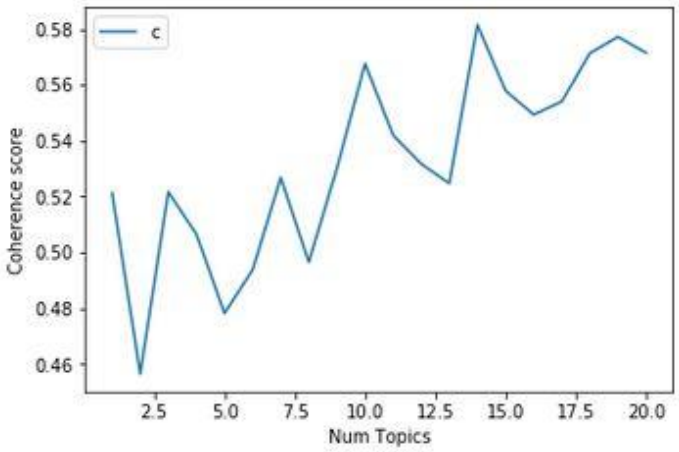
5.4 Hasil *Topic Modeling* dengan *Latent Dirichlet Allocation* (LDA)

Dalam menentukan hasil pemodelan dapat dilakukan dengan melihat pada visualisasi dari grafik *coherence score*. *Coherence score* merupakan suatu ukuran yang akan digunakan untuk mengevaluasi *Topic Modeling*, dimana jika *coherence skor* topik yang tinggi maka model yang dihasilkan tersebut yang baik. Grafik dari *coherence score* yang dihasilkan terdapat naik dan turun, peneliti akan

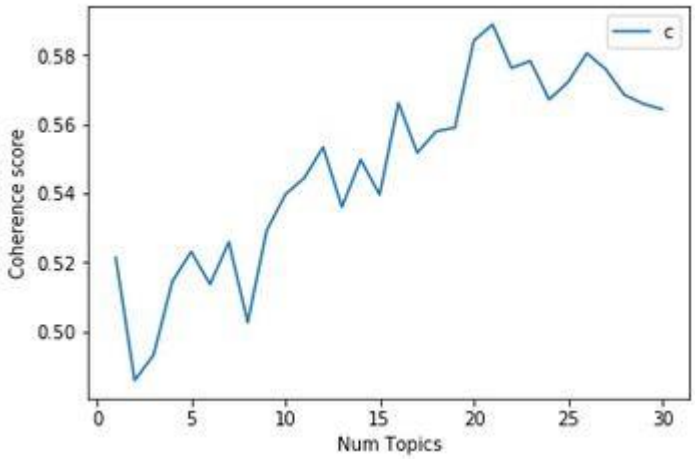
menunjukkan beberapa perbandingan dalam grafiknya yaitu dimana topik dimulai dari 0-11,0-21,0-31 yang dapat dilihat pada **Gambar 5.3** sebagai berikut:



Gambar 5.3 Grafik *Coherence Score* Dengan Limit Topik 0-11



Gambar 5.4 Grafik *Coherence Score* Dengan Limit Topik 0-21



Gambar 5.5 Grafik *Coherence Score* Dengan Limit Topik 0-31

Berdasarkan **Gambar 5.3, Gambar 5.4, Gambar 5.5** diatas didapatkan informasi bahwa pada grafik *coherence score* diatas memiliki pola yang berulang dan semakin banyak limit topiknya maka semakin tinggi nilai *coherence* yang dihasilkan. Kemudian dari perbandingan grafik tersebut maka peneliti memutuskan untuk menggunakan limit topik 0-11 dengan jumlah topik sebanyak 3 yang memiliki nilai *coherence* terbesar yaitu sebesar 0.5528, dengan begitu jumlah topik tersebut yang akan menjadi acuan untuk membuat model selanjutnya. pada tabel berikut ini merupakan *coherence score* yang dihasilkan:

Tabel 5.9 *Coherence Score*

Num Topic	*Coherence Score*	*Num Topic*	*Coherence Score*
1	0.5212	6	0.5377
2	0.4247	7	0.4796
3	0.5528	8	0.5506
4	0.4964	9	0.5296
5	0.5039	10	0.5493

Setelah didapatkan jumlah topiknya berdasarkan grafik *coherence score*, kemudian akan didapatkan model LDA berdasarkan banyak topiknya yaitu 3 dengan jumlah kata yang ditampilkan dalam model 10 kata yang memiliki bobot masing-masing dari tiap kata tersebut.

5.4.1 Model LDA Topik ke-1

Pada topik ke-1 didapatkan model LDA yaitu sebagai berikut:

Tabel 5.10 Model LDA Topik 1

```
'0.004*"convolutional_neural" + 0.004*"tingkat_akurasi" + '
  '0.004*"deep_learning" + 0.004*"kabupaten_sleman" +
0.004*"neural_network" + '
  '0.003*"citra" + 0.003*"analisis_regresi" +
0.003*"kendaraan_bermotor" + '
  '0.003*"tanaman" + 0.003*"cluster"'
```

Kemudian setelah diperoleh model LDA maka model tersebut dapat dilihat dengan visualisasi PyLDAvis dan keterkaitan antar kata yang dihasilkan. Pada panel sisi kiri merupakan pemetaan jarak dari antar topik (*intertopic distance map*) via *multidimensional scaling* yang terdapat juga *cluster* topik yang berbentuk lingkaran dengan nomor tertentu pada setiap *cluster* topik. Sedangkan pada panel sisi kanan terdapat 30 buah terminologi yang paling relevan untuk topik tertentu.

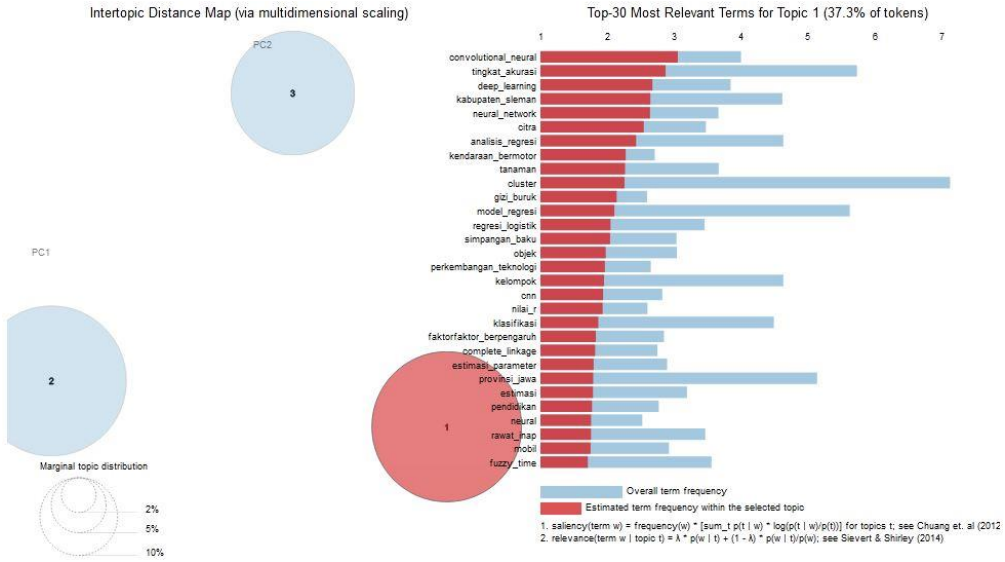
Pada letak untuk tiap topiknya yaitu tergantung pada titik koordinat dari masing-masing topik dengan melihat berdasarkan *principal component* (PC). Nilai PC untuk tiap topiknya yaitu pada tabel berikut.

Tabel 5.11 Nilai *Principal Component*

Topik	PC1	PC2
1	0.022057	-0.015207
2	-0.025566	-0.009700
3	0.003510	0.024908

Dapat dilihat pada **Tabel 5.11** diatas yang merupakan nilai *principal component* (PC) untuk masing-masing topiknya yang dipakai dalam visualisasi PyLDAvis. Untuk nilai PC1 yang digunakan sebagai titik koordinat pada sumbu X, sedangkan nilai PC2 yang digunakan sebagai titik koordinat pada sumbu Y.

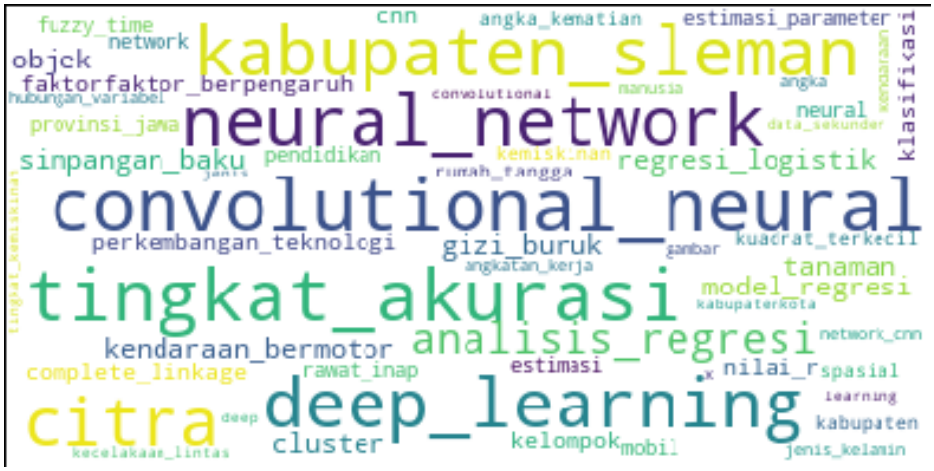
Selanjutnya dapat dilihat pada **Gambar 5.6** dibawah dipilih topik 1 maka lingkarannya akan berubah warna menjadi merah yang kemudian pada *bar chart* panel sisi kanan akan berubah berwarna merah yang memperlihatkan estimasi *term frequncy* pada topik yang pilih. Kemudian untuk letak topik 1 berada pada kuadran ke IV yang dapat dilihat pada **Tabel 5.11** dimana nilai PC1 bernilai positif dan nilai PC2 bernilai negatif.



Gambar 5.6 Visualisasi Topik 1 Dengan PyLDAvis

Berdasarkan model LDA yang diperoleh untuk topik ke-1 pada **Tabel 5.10** dan visualisasi topik ke-1 pada **Gambar 5.6** diatas atau untuk lebih jelasnya dapat

dilihat pada **Lampiran 3**, maka dapat disimpulkan bahwa pada model LDA topik ke-1 yang banyak muncul dan berkaitan satu sama lain yaitu mengenai metode *Deep Learning* dengan *Convolutional Neural Network* yang berarti bahwa topik 1 mengenai *Artificial Intelligence* (AI). Kemudian dari model tersebut juga dapat ditampilkan dengan bentuk *wordcloud* sebagai berikut.



Gambar 5.7 *Wordcloud* Topik Ke-1

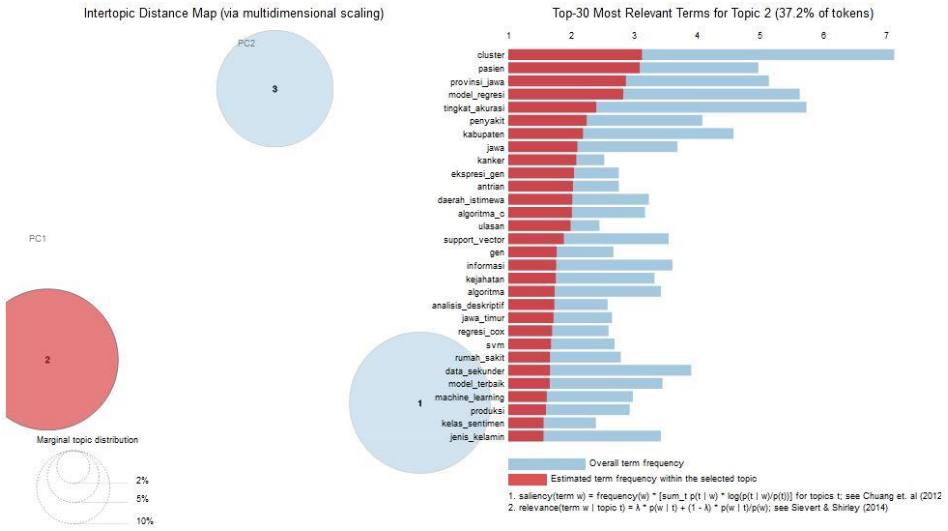
5.4.2 Model LDA Topik ke-2

Pada topik ke-2 didapatkan model LDA yaitu sebagai berikut:

Tabel 5.12 Model LDA Topik 2

```
'0.004*"cluster" + 0.004*"pasien" + 0.004*"provinsi_jawa" + '
  '0.004*"model_regresi" + 0.003*"tingkat_akurasi" +
0.003*"penyakit" + '
  '0.003*"kabupaten" + 0.003*"jawa" + 0.003*"kanker" +
0.003*"ekspresi_gen"'
```

Kemudian setelah diperoleh model LDA maka model tersebut dapat dilihat dengan visualisasi PyLDAvis dan keterkaitan antar kata yang dihasilkan. Kemudian untuk letak topik 2 berada pada kuadran ke III yang dapat dilihat pada **Tabel 5.11** dimana nilai PC1 bernilai negatif dan nilai PC2 bernilai negatif. Dapat dilihat visualisasi PyLDAvis topik 2 pada **Gambar 5.8** dibawah ini.



Gambar 5.8 Visualisasi Topik 2 Dengan PyLDAvis

Berdasarkan model LDA yang diperoleh untuk topik ke-2 pada **Tabel 5.12** dan visualisasi topik ke-2 pada **Gambar 5.8** diatas atau untuk lebih jelasnya dapat dilihat pada **Lampiran 3**, maka dapat disimpulkan bahwa pada model LDA topik ke-2 yang banyak muncul dan berkaitan satu sama lain yaitu mengenai metode Model Regresi Survival, karena dalam model diperoleh kata model regresi, pasien, penyakit, dan kanker, yang berarti bahwa topik 2 mengenai Statistika pada Bidang Kesehatan. Kemudian dari model tersebut juga dapat ditampilkan dengan bentuk *wordcloud* sebagai berikut.

Gambar 5.9 *Wordcloud* Topik Ke-2

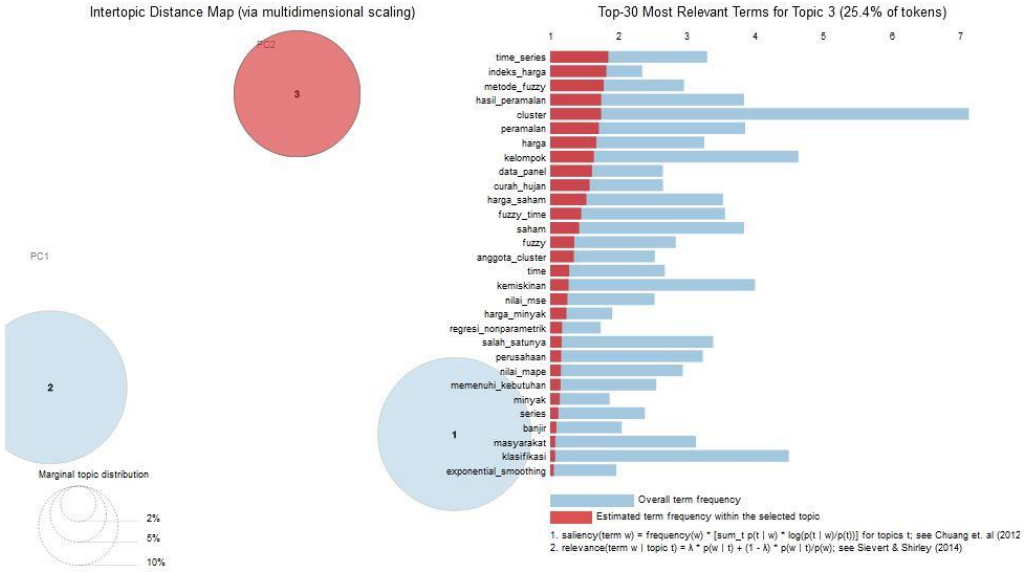
5.4.3 Model LDA Topik ke-3

Pada topik ke-3 didapatkan model LDA yaitu sebagai berikut:

Tabel 5.13 Model LDA Topik 3

```
'0.004*"time_series" + 0.004*"indeks_harga" + 0.003*"metode_fuzzy"
+ ''0.003*"hasil_peramalan" + 0.003*"cluster" + 0.003*"peramalan"
+ ''0.003*"harga" + 0.003*"kelompok" + 0.003*"data_panel" + '
'0.003*"curah_hujan"'
```

Kemudian setelah diperoleh model LDA maka model tersebut dapat dilihat dengan visualisasi PyLDAvis dan keterkaitan antar kata yang dihasilkan. Kemudian untuk letak topik 3 berada pada kuadran ke I yang dapat dilihat pada **Tabel 5.11** dimana nilai PC1 bernilai positi dan nilai PC2 bernilai positif. Dapat dilihat visualisasi PyLDAvis topik 2 pada **Gambar 5.10** dibawah ini.



Gambar 5.10 Visualisasi Topik 3 Dengan PyLDAvis

Berdasarkan model LDA yang diperoleh untuk topik ke-3 pada **Tabel 5.13** dan visualisasi topik ke-3 pada **Gambar 5.10** diatas atau untuk lebih jelasnya dapat dilihat pada **Lampiran 3**, maka dapat disimpulkan bahwa pada model LDA topik ke-3 yang banyak muncul dan berkaitan satu sama lain yaitu mengenai metode *Time Series* atau Peramalan, yang berarti bahwa topik 3 mengenai Statistika Perekonomian. Kemudian dari model tersebut juga dapat ditampilkan dengan bentuk *wordcloud* sebagai berikut.

Gambar 5.11 *Wordcloud* Topik Ke-3

BAB VI
PENUTUP

6.1 Kesimpulan

Tuliskan kesimpulan yang menjawab rumusan masalah:

1. Gambaran umum pada abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia yaitu jumlah skripsi tahun angkatan 2011-2015 yang digunakan dalam penelitian ini sebanyak 416 dimana jumlah terbanyak pada tahun angkatan 2014 dan jumlah paling sedikit yaitu tahun angkatan 2011. Kemudian didapatkan informasi bahwa rata-rata lama pengerjaan TA menurut data yang digunakan dalam penelitian ini pada tiap tahunnya semakin kecil atau pada grafik semakin menurun, yang dapat diartikan bahwa lama pengerjaan TA tiap tahun angkatannya semakin baik.
2. Hasil dari analisis pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) pada data abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia angkatan 2011-2015 diperoleh jumlah topik sebanyak 3 dengan *coherence score* sebesar 0.5528 dengan kata yang saling berkaitan. Topik yang dihasilkan yaitu sebagai berikut:
   - Model LDA topik ke-1 yaitu mengenai *Artificial Intelligence* (AI).
   - Model LDA topik ke-2 yaitu mengenai Statistika pada Bidang Kesehatan.
   - Model LDA topik ke-3 yaitu mengenai Statistika Perekonomian.

6.2 Saran

Saran yang diberikan berdasarkan hasil penelitian ini yaitu:

1. Jumlah data yang digunakan pada penelitian berikutnya bisa lebih banyak lagi.
2. Dapat melalukan penyempunaan proses *preprocessing* diantaranya yaitu pada bagian *stopwords* dengan menyesuaikan kata apa saja yang dapat dihilangkan yang tidak menghilangkan makna kata yang terkandung. Kemudian dapat ditambahkan dengan proses *stemming*.
3. Dapat mengembangkan pemodelan topik dengan metode yang lainnya.

DAFTAR PUSTAKA

Agustina, A. (2017). *ANALISIS DAN VISUALISASI SUARA PELANGGAN PADA PUSAT LAYANAN PELANGGAN DENGAN PEMODELAN TOPIK MENGGUNAKAN LATENTDIRICHLET ALLOCATION (LDA) STUDI KASUS: PT.* Surabaya: DEPARTEMEN SISTEM INFORMASI Fakultas Teknologi Informasi Institut Teknologi Sepuluh Nopember.

Bagus C.W, D. (2017). *Text Mining pada Media Sosial Twitter (Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2).* Jember: Universitas Jember.

Cendana, M., & Permana, S. D. (2019). PRA-PEMROSESAN TEKS PADA GRUP WHATSAPP UNTUK PEMODELAN TOPIK. *Jurnal Mantik Penusa , Volume 3, No. 3*, 112.

Fattah, R. (2016). *Twitter Text Mining Untuk Informasi Gempa Bumi Menggunakan T-IDF Di Indonesia.* Malang: Universitas Islam Negeri Maulana Malik Ibrahim.

Februariyanti, H., & Santoso, D. B. (2017). HIERARCHICAL AGGLOMERATIVE CLUSTERING UNTUK PENGELOMPOKAN SKRIPSI MAHASISWA. *Prosiding SINTAK* , 33-40.

Gusti. A, I., Akbar, A. L., & Akbar, M. S. (2016). Optimasi Naive Bayes Dengan Pemilihan Fitur Dan Pembobotan Gain Ratio. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi* , 23-24.

Herwanto, G. B. (2018). DOCUMENT CLUSTERING DENGAN LATENT DIRICHLET ALLOCATION DAN WARD HIERARICHAL CLUSTERING. *Jurnal Pseudocode , Volume V Nomor 2*, 29-37.

Hudaya, C. S., Fakhrurroja, H., & Alamsyah, A. (2019). ANALISIS PERSEPSI KONSUMEN TERHADAP BRAND GO-JEK PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE SENTIMENT ANALYSIS DAN TOPIC MODELLING. *Jurnal Mitra Manajemen (JMM Online) , Vol. 3 No.6*, 667.

Ilma, N. (2015). Peran Pendidikan Sebagai Modal Utama Membangun Karakter Bangsa. *TADBIRJurnal Manajemen Pendidikan Islam* , 82-83.

Irmawati. (2017). *PENERAPAN ANALISIS MULTIDIMENSIONAL SCALING PADA PEMETAAN KARAKTERISTIK KEMISKINAN DI PROVINSI SULAWESI SELATAN*. Makassar: Universitas Islam Negeri Alauddin.

Kabiru, I. N., & Sari, P. K. (2019). ANALISA KONTEN MEDIA SOSIAL E-COMMERCE PADA INSTAGRAM MENGGUNAKAN METODE SENTIMEN ANALYSIS DAN LDA-BASED TOPIC MODELING (STUDI KASUS: SHOPEE INDONESIA). *e-Proceeding of Management , Vol.6, No.1*, 14.

Karmayasa, O. (2012). IMPLEMENTASI VECTOR SPACE MODEL DAN BEBERAPA NOTASI METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) PADA SISTEM TEMU KEMBALI INFORMASI. *JELIKU - Jurnal Elektronik Ilmu Komputer Udayana , Volume 1 No 1*, 2.

Koesriputranto, A. (2015). *Prediksi Harga Saham Di Indonesia Dengan Menggunakan Metode Hybrid Principal Component Analysis dan Support Vector Machine (PCA-SVM).* Surabaya: Institut Teknologi Sepuluh Nopember.

Listari. (2019). *INISIASI NATURAL LANGUAGE PROCESSING (NLP) DAN KLASIFIKASI JENIS WISATA KULINER UNTUK PROGRAM CHATBOT, (Studi Kasus : Informasi Wisata Kuliner Daerah Istimewa Yogyakarta).* Yogyakarta: Statistika, UII.

Nasution, M. K. (2017). Abstrak - Suatu Karya Ilmiah. *Teknik Penulisan Karya Ilmiah, Bagian 3* , 2.

Nugroho, D. D., & Alamsyah, A. (2018). ANALISIS KONTEN PELANGGAN AIRBNB PADA NETWORK SOSIAL MEDIA TWITTER CONTENT ANALYSIS OF AIRBNB CUSTOMER BASED ON TWITTER SOCIAL MEDIA. *e-Proceeding of Management* , 1623 & 1626.

Pradita, D., Satyahadewi, N., & Perdana, H. (2019). ANALISIS PERBANDINGAN METODE MULTIDIMENSIONAL SCALING (MDS) DAN WEIGHTED MULTIDIMENSIONAL SCALING (WMDS). *Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster) , Volume 08, No. 1*, 149.

Prilianti, K. R., & Wijaya, H. (2014). Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering. *Jurnal Cybermatika , Vol. 2 No. 1*, 1-6.

Putra, I. M. (2017). *Analisis Topik Informasi Publik Media Sosial Di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA).* Surabaya: Institut Teknologi Sepuluh Nopember.

Putra, I. M., & Kusumawardani, R. P. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). *JURNAL TEKNIK ITS Vol. 6, No. 2* , A312.

pypi.org. (2018, Juni 6). Retrieved Februari 23, 2020, from pyLDAvis 2.1.2 : https://pypi.org/project/pyLDAvis/

Rahmawati, L., Sihwi, S. W., & Suryani, E. (2014). ANALISA CLUSTERING MENGGUNAKAN METODE K-MEANS DAN HIERARCHICAL CLUSTERING (STUDI KASUS : DOKUMEN SKRIPSI JURUSAN KIMIA, FMIPA, UNIVERSITAS SEBELAS MARET). *ITSMART: Jurnal Teknologi dan Informasi , Vol 3, No 2*.

Sanjaya, S., & Absar, E. A. (2015). Pengelompokan Dokumen Menggunakan Winnowing Fingerprint dengan Metode K-Nearest Neighbour. *Jurnal CoreIT , Vol.1, No.2,* 52.

Sari, D. R., Windarto, A. P., Hartama, D., & Solikhun. (2018). Sistem Pendukung Keputusan untuk Rekomendasi Kelulusan Sidang Skripsi Menggunakan Metode AHP-TOPSIS. *Jurnal Teknologi dan Sistem Komputer* , 1.

Wahyudin, I., Tosida, E. T., & Andria, F. (2019). *Teori dan Panduan Praktis Data Science dan Big Data.* Bogor: Lembaga Penelitian dan Pengabdian pada Masyarakat Universitas Pakuan.

Wicaksana, D. A., Adikara, P. P., & Adinugroho, S. (2018). Clustering Dokumen Skripsi Dengan Menggunakan Hierarchical Agglomerative Clustering. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer , Vol. 2, No. 12*, 6227-6234.

Wirasakti, L. A., Permadi, R., Hartanto, A. D., & Hartatik. (2020). Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik. *JURNAL MEDIA INFORMATIKA BUDIDARMA , Volume 4, Nomor 1*, 27-31.

Yoren. (2018). *PERBANDINGAN RAW TF DAN BINARY TF PADA SISTEM PENCARIAN DI SITUS MUSEUM WAYANG KEKAYON YOGYAKARTA.* Yogyakarta: PROGRAM STUDI TEKNIK INFORMATIKA JURUSAN TEKNIK INFORMATIKA FAKULTAS SAINS DAN TEKNOLOGI UNIVERSITAS SANATA DHARMA.

LAMPIRAN

Lampiran 1 Data Skripsi Tahun 2011-2015

No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
1	15611122	2015	Aditya Hidayatullah	IMPELMENTASI LOGARITMA SPECTRAL BICLUSTERING PADA DATA ESKPRESIGEN (Studi kasus : Data Pasien Pada Studi "Gene Expression Change in Blood During Etanol Exposure"/ GSE 20486)	Etanol (alcohol) adalah nama suatu golongan senyawa organik yang mengandung unsur C,H dan O. Etanol dalam ilmu kimia disebut juga sebagai etil alkohol dengan rumus senyawa C2H5OH. Badan Narkotika Nasional (BNN) memperkirakan ada 3,2 juta orang (1,5% dari total populasi) di Indonesia mempunyai riwayat menggunakan NAPZA diantaranya 46% adalah perilaku minum alcohol. Dampak dari perilaku minum alcohol salah satunya adalah mabuk, pengendara mabuk merupakan faktor yang beresiko menyebabkan kecelakaan lalu lintas, yang menyebabkan kejadian meninggal dunia. Dalam kasus yang lain tahun 1994-2000 sekitar 11,5% kecelakaan penerbangan umum berhubungan dengan alkohol terkait dengan pilot yang memiliki riwayat DWI	09 September 2018	13/07/19	10.23

No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
					(mengemudi sambil mabuk). Penelitian ini adalah untuk mengetahui pengelompokan karateristik pengaruh dalam tingkat ekspresi gen yang terpapar etanol dengan Spectral Biclustering, menggunakan RNA yang diekstraksi dari seluruh darah, ketika etanol memasuki sistem darah. Dengan analisis microarray. Hasil dari analisis didapatkan 5 bicluster dari 54 sampel dan 201 gen dengan nilai rata-rata ekspresi gen terbesar adalah X213350_at yang berada pada bicluster 4.			
2	15611026	2015	Rina Sriwiji	STUDI EMPIRIS PADA PEMODELAN DAN PREDIKSI HARGA BITCOIN BERDASARKAN	Dalam beberapa tahun terakhir ini, Bitcoin telah menarik banyak perhatian karena sifatrnya yang mendukung teknologi enkripsi dan unit moneter. Bitcoin adalah mata uang elektronik yang memungkinkan pembayarn secara online tanpa melalui lembaga keuangan. Bitcoin menjadi investasi yang menjanjikan bagi para	09 September 2018	24 Mei 2019	8.57

No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
				INFORMASI BLOCKCHAIN MENGGUNAKAN BAYESIAN REGULARIZATION NEURAL NETWORK	pedagang finansial karena harganya yang fluktuatif berpotensi menghasilkan laba tinggi (semakin tinggi risikonya, semakin tinggi pula pengembaliannya). Tidak seperti stok konvensiaonal, Bitcoin diperdagangkan selama 24 jam sehari tanpa periode tutup, sehingga meningkatkan risiko. Prediksi nilai Bitcoin diharapkan dapat meminimalkan risiko dengan mempertimbangkan beberapa informasi seperti informasi blockchain, faktor ekonomi makro, dan rasio mata uang global. Namun, multikolinearitas diantara variabel-variabel independen menyebabkan metode regresi tidak dapat digunakan. Penelitian ini menggunakan Bayesian Regularization Neural Network (BRNN) yang merupakan asumsi bebas. Metode ini adalah Single Hidden Layer Feed Forward Neural Network (SLNN) yang			

No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
					memanfaatkan konsep Bayesian untuk mengoptimalkan bobot, bias, dan kekuatan koneksi. Data yang digunakan adalah data time series dari 23 Januari 2017, hingga 23 Januari 2019. Regresi dengan subset digunakan untuk mengurangi variabel independen, dari total 25 variabel menjadi 14 variabel. Hasilnya menunjukkan bahwa model BRNN yang dibentuk dapat memprediksi nilai Bitcoin dengan baik, diperoleh nilai prediksi tidak jauh berbeda dari data aktual, dengan akurasi sebesar 91,1% berdasarkan nilai MAPE.			
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
416	11611071	2011	Syf. Cindy Ayu W.A	ANALISIS FAKTOR-FAKTOR YANG BERPENGARUH	Provinsi Kalimantan Barat memiliki angka kecelakaan lalu lintas yang cukup tinggi, khususnya di Kota Pontianak. Penelitian ini bertujuan untuk melihat profil kecelakaan lalu lintas di Kota Pontianak pada tahun	1-Apr-15	11-Feb-16	10.53

No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
				TERHADAP TINGKAT KEPARAHAN KORBAN KECELAKAAN LALU LINTAS DI KOTA PONTIANAK MENGGUNAKAN REGRESI LOGISTIK ORDINAL	2014, mengetahui faktor-faktor yang mempengaruhi tingkat keparahan korban dan mengetahui besarnya peluang tingkat keparahan korban kecelakaan lalu lintas berdasarkan faktor-faktor yang mempengaruhi terjadinya kecelakaan lalu lintas. Data yang digunakan dalam penelitian bersumber dari Polda Kalimantan Barat. Analisis data menggunakan analisis deskriptif dan regresi logistik ordinal. Berdasarkan hasil penelitian, didapatkan bahwa angka kecelakaan tertinggi pada tahun 2014 terjadi pada bulan April yaitu sebanyak 50 kejadian. Korban kecelakaan lalu lintas umumnya mengalami luka ringan. Faktor-faktor yang mempengaruhi tingkat keparahan korban adalah jenis kecelakaan tunggal, jenis kecelakaan tabrak lari dan usia korban.			

Link Data Skripsi lengkap:

http://bit.ly/DataSkripsiku

Lampiran 2 *Script* dan *Output Topic Modeling*

```
#membuka file
import pandas as pd
data = pd.read_excel('Skripsiku.xlsx')
data.head()
```

	No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
0	1	15611122	2015	Aditya Hidayatullah	IMPELMENTASI LOGARITMA SPECTRAL BICLUSTERING P...	Etanol (alcohol) adalah nama suatu golongan se...	2018-09-09	2019-07-13	10.233333
1	2	15611026	2015	Rina Sriwiji	STUDI EMPIRIS PADA PEMODELAN DAN PREDIKSI HARG...	Dalam beberapa tahun terakhir ini, Bitcoin tel...	2018-09-09	2019-05-24	8.566667
2	3	15611061	2015	Nita Tri Anggraini	INISIASI NATURAL LANGUAGE PROCESSING (NLP) DAN...	Perkembangan teknologi saat ini mempermudah ke...	2018-09-09	2019-02-15	5.300000
3	4	15611035	2015	Rohmat Apriyanto	ANALISIS FAKTOR-FAKTOR KONDISI DARAH IBU HAMIL...	Ibu hamil yang mengalami keterlambatan kelahir...	2018-09-09	2019-02-14	5.266667
4	5	15611005	2015	Sefiana Anggraini	PERBANDINGAN METODE TRIPLE EXPONENTIAL SMOOTHI...	Perkembangan bandara di Indonesia dalam infras...	2018-09-09	2019-02-11	5.166667

```
#lower Casing
data['Abstrak']  = data['Abstrak'].apply(lambda x: " ".join(x.lower() for x in x.split()))
data['Abstrak'].head()
```

```
0    etanol (alcohol) adalah nama suatu golongan se...
1    dalam beberapa tahun terakhir ini, bitcoin tel...
2    perkembangan teknologi saat ini mempermudah ke...
3    ibu hamil yang mengalami keterlambatan kelahir...
4    perkembangan bandara di indonesia dalam infras...
Name: Abstrak, dtype: object
```

```
# Removing Punctuation
data['Abstrak'] = data['Abstrak'].str.replace('[^\w\s]','')
data['Abstrak'].head()
```

```
0    etanol alcohol adalah nama suatu golongan seny...
1    dalam beberapa tahun terakhir ini bitcoin tela...
2    perkembangan teknologi saat ini mempermudah ke...
3    ibu hamil yang mengalami keterlambatan kelahir...
4    perkembangan bandara di indonesia dalam infras...
Name: Abstrak, dtype: object
```

```
#remove angka
data['Abstrak']=data['Abstrak'].str.replace(r'[\d+]', '')
data['Abstrak']
```

```
0      etanol alcohol adalah nama suatu golongan seny...
1      dalam beberapa tahun terakhir ini bitcoin tela...
2      perkembangan teknologi saat ini mempermudah ke...
3      ibu hamil yang mengalami keterlambatan kelahir...
4      perkembangan bandara di indonesia dalam infras...
5      perkembangan sosial media membawa banyak penga...
6      mata merupakan organ sensorik yang sangat komp...
7      ekuitas merek diciptakan tidak hanya oleh bebe...
8      pisang merupakan buah yang banyak dikonsumsi d...
9      saat ini teknologi informasi berkembang pesat ...
10     saham merupakan surat berharga yang dapat dibe...
11     gizi buruk adalah keadaan kurang gizi tingkat ...
12     bioinformatika merupakan kajian yang tak lepas...
13     bioinformatika adalah konsep biologi dan mener...
14     gandum adalah salah satu bijibijian tanaman se...
15     banyaknya institusi perguruan tinggi yang ada ...
16     lingkungan menjadi salah satu faktor yang berp...
17     tuberkulosis tb merupakan masalah kesehatan du...
18     kereta api adalah sarana perkeretaapian dengan...
```

```
from xlsxwriter.utility import xl_rowcol_to_cell
saveresult = pd.ExcelWriter('removepunct.xlsx', engine='xlsxwriter')
data['Abstrak'].to_excel(saveresult, index=False, sheet_name='report')
saveresult.save()
```

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

factory = StopWordRemoverFactory()
stopwords = factory.get_stop_words()
print(stopwords)
```

```
['yang', 'untuk', 'pada', 'ke', 'para', 'namun', 'menurut', 'antara', 'dia', 'dua', 'ia', 'seperti', 'jika', 'ji
ka', 'sehingga', 'kembali', 'dan', 'tidak', 'ini', 'karena', 'kepada', 'oleh', 'saat', 'harus', 'sementara', 'se
telah', 'belum', 'kami', 'sekitar', 'bagi', 'serta', 'di', 'dari', 'telah', 'sebagai', 'masih', 'hal', 'ketika',
'adalah', 'itu', 'dalam', 'bisa', 'bahwa', 'atau', 'hanya', 'kita', 'dengan', 'akan', 'juga', 'ada', 'mereka', '
sudah', 'saya', 'terhadap', 'secara', 'agar', 'lain', 'anda', 'begitu', 'mengapa', 'kenapa', 'yaitu', 'yakni', '
daripada', 'itulah', 'lagi', 'maka', 'tentang', 'demi', 'dimana', 'kemana', 'pula', 'sambil', 'sebelum', 'sesuda
h', 'supaya', 'guna', 'kah', 'pun', 'sampai', 'sedangkan', 'selagi', 'sementara', 'tetapi', 'apakah', 'kecuali',
'sebab', 'selain', 'seolah', 'seraya', 'seterusnya', 'tanpa', 'agak', 'boleh', 'dapat', 'dsb', 'dst', 'dll', 'da
hulu', 'dulunya', 'anu', 'demikian', 'tapi', 'ingin', 'juga', 'nggak', 'mari', 'nanti', 'melainkan', 'oh', 'ok',
'seharusnya', 'sebetulnya', 'setiap', 'setidaknya', 'sesuatu', 'pasti', 'saja', 'toh', 'ya', 'walau', 'tolong',
'tentu', 'amat', 'apalagi', 'bagaimanapun']
```

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

factory = StopWordRemoverFactory()
stopword =  factory.get_stop_words()

# Kalimat
data['Abstrak'] = data['Abstrak'].apply(lambda x: " ".join(x for x in x.split() if x not in stopword))
data['Abstrak'].head()
```

```
0    etanol alcohol nama suatu golongan senyawa org...
1    beberapa tahun terakhir bitcoin menarik banyak...
2    perkembangan teknologi mempermudah kegiatan be...
3    ibu hamil mengalami keterlambatan kelahiran ba...
4    perkembangan bandara indonesia infrastruktur t...
Name: Abstrak, dtype: object
```

```
stopword
```

```
['yang',
 'untuk',
 'pada',
 'ke',
 'para',
 'namun',
 'menurut',
 'antara',
 'dia',
 'dua',
 'ia',
 'seperti',
 'jika',
 'jika',
 'sehingga',
 'kembali',
 'dan',
 'tidak',
 'ini',
```

```
# Import Stopword Factory class
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

#Create factory
factory = StopWordRemoverFactory()
more_stopword = ['_jijnwijuj','_jijnwijyj','a__','aapl','universitas','islam','tujuan','penelitian',
                 'studi','mahasiswa','mempengaruhi']

# Tambahkan Stopword Baru
stopwordplus = factory.get_stop_words()+stopwords()+more_stopword

data['Abstrak'] = data['Abstrak'].apply(lambda x: " ".join(x for x in x.split() if x not in stopwordplus))
data['Abstrak'].head()
```

```
0    etanol alcohol nama golongan senyawa organik m...
1    bitcoin menarik perhatian sifatrnya mendukung ...
2    perkembangan teknologi mempermudah kegiatan be...
3    hamil mengalami keterlambatan kelahiran bayi h...
4    perkembangan bandara indonesia infrastruktur t...
Name: Abstrak, dtype: object
```

```
from xlsxwriter.utility import xl_rowcol_to_cell
saveresult = pd.ExcelWriter('printstopword.xlsx', engine='xlsxwriter')
data['Abstrak'].to_excel(saveresult, index=False, sheet_name='report')
saveresult.save()
```

```
text = data['Abstrak']
text_list =  [i.split() for i in text]
```

```
print(len(text_list))
```

```
416
```

```
print(text_list)
```

```
[['etanol', 'alcohol', 'nama', 'golongan', 'senyawa', 'organik', 'mengandung', 'unsur', 'ch', 'o', 'etanol', '
ilmu', 'kimia', 'etil', 'alkohol', 'rumus', 'senyawa', 'choh', 'badan', 'narkotika', 'nasional', 'bnn', 'juta
', 'orang', 'total', 'populasi', 'indonesia', 'riwayat', 'napza', 'perilaku', 'minum', 'alcohol', 'dampak', 'p
erilaku', 'minum', 'alcohol', 'salah', 'satunya', 'mabuk', 'pengendara', 'mabuk', 'faktor', 'beresiko', 'menye
babkan', 'kecelakaan', 'lintas', 'menyebabkan', 'kejadian', 'meninggal', 'dunia', 'kecelakaan', 'penerbangan',
'berhubungan', 'alkohol', 'terkait', 'pilot', 'memiliki', 'riwayat', 'dwi', 'mengemudi', 'mabuk', 'pengelompok
an', 'karateristik', 'pengaruh', 'tingkat', 'ekspresi', 'gen', 'terpapar', 'etanol', 'spectral', 'biclustering
', 'rna', 'diekstraksi', 'darah', 'etanol', 'memasuki', 'sistem', 'darah', 'analisis', 'microarray', 'hasil',
'analisis', 'didapatkan', 'bicluster', 'sampel', 'gen', 'nilai', 'ratarata', 'ekspresi', 'gen', 'terbesar', 'x
_at', 'bicluster'], ['bitcoin', 'menarik', 'perhatian', 'sifatrnya', 'mendukung', 'teknologi', 'enkripsi', 'un
it', 'moneter', 'bitcoin', 'mata', 'uang', 'elektronik', 'pembayarn', 'online', 'lembaga', 'keuangan', 'bitcoi
n', 'investasi', 'menjanjikan', 'pedagang', 'finansial', 'harganya', 'fluktuatif', 'berpotensi', 'menghasilkan
', 'laba', 'risikonya', 'pengembaliannya', 'stok', 'konvensiaonal', 'bitcoin', 'diperdagangkan', 'jam', 'sehar
i', 'periode', 'tutup', 'meningkatkan', 'risiko', 'prediksi', 'nilai', 'bitcoin', 'diharapkan', 'meminimalkan
', 'risiko', 'mempertimbangkan', 'informasi', 'informasi', 'blockchain', 'faktor', 'ekonomi', 'makro', 'rasio
', 'mata', 'uang', 'global', 'multikolinearitas', 'variabelvariabel', 'independen', 'menyebabkan', 'metode', '
regresi', 'bayesian', 'regularization', 'neural', 'network', 'brnn', 'asumsi', 'bebas', 'metode', 'single', 'h
idden', 'layer', 'feed', 'forward', 'neural', 'network', 'slnn', 'memanfaatkan', 'konsep', 'bayesian', 'mengop
timalkan', 'bobot', 'bias', 'kekuatan', 'koneksi', 'data', 'data', 'time', 'series', 'januari', 'januari', 're
```

```
import re

# Function to Tokenize words
def tokenize(text):
    tokens = re.split('\W+', text) #W+ means that either a word character (A-Za-z0-9_) or a dash (-) can go there.
    return tokens

data['Abstrak'] = data['Abstrak'].apply(lambda x: tokenize(x.lower()))
#We convert to lower as Python is case-sensitive.
data.head()
```

	No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
0	1	15611122	2015	Aditya Hidayatullah	IMPELMENTASI LOGARITMA SPECTRAL BICLUSTERING P...	[etanol, alcohol, nama, golongan, senyawa, org...	2018-09-09	2019-07-13	10.233333
1	2	15611026	2015	Rina Sriwiji	STUDI EMPIRIS PADA PEMODELAN DAN PREDIKSI HARG...	[bitcoin, menarik, perhatian, sifatnya, mendu...	2018-09-09	2019-05-24	8.566667
2	3	15611061	2015	Nita Tri Anggraini	INISIASI NATURAL LANGUAGE PROCESSING (NLP) DAN...	[perkembangan, teknologi, mempermudah, kegiata...	2018-09-09	2019-02-15	5.300000
3	4	15611035	2015	Rohmat Apriyanto	ANALISIS FAKTOR-FAKTOR KONDISI DARAH IBU HAMIL...	[hamil, mengalami, keterlambatan, kelahiran, b...	2018-09-09	2019-02-14	5.266667
4	5	15611005	2015	Sefiana Anggraini	PERBANDINGAN METODE TRIPLE EXPONENTIAL SMOOTHI...	[perkembangan, bandara, indonesia, infrastrukt...	2018-09-09	2019-02-11	5.166667

```
data.to_csv("ella.csv", sep=',')
import pandas as pd
data = pd.read_csv('ella.csv')
data.head()
```

	Unnamed: 0	No	NIM	Tahun	Nama	Judul	Abstrak	SK Pembimbing Skripsi	Tanggal Pendadaran	Lama TA
0	0	1	15611122	2015	Aditya Hidayatullah	IMPELMENTASI LOGARITMA SPECTRAL BICLUSTERING P...	['etanol', 'alcohol', 'nama', 'golongan', 'sen...	2018-09-09	2019-07-13	10.233333
1	1	2	15611026	2015	Rina Sriwiji	STUDI EMPIRIS PADA PEMODELAN DAN PREDIKSI HARG...	['bitcoin', 'menarik', 'perhatian', 'sifatnya...	2018-09-09	2019-05-24	8.566667
2	2	3	15611061	2015	Nita Tri Anggraini	INISIASI NATURAL LANGUAGE PROCESSING (NLP) DAN...	['perkembangan', 'teknologi', 'mempermudah', '...	2018-09-09	2019-02-15	5.300000
3	3	4	15611035	2015	Rohmat Apriyanto	ANALISIS FAKTOR-FAKTOR KONDISI DARAH IBU HAMIL...	['hamil', 'mengalami', 'keterlambatan', 'kelah...	2018-09-09	2019-02-14	5.266667
4	4	5	15611005	2015	Sefiana Anggraini	PERBANDINGAN METODE TRIPLE EXPONENTIAL SMOOTHI...	['perkembangan', 'bandara', 'indonesia', 'infr...	2018-09-09	2019-02-11	5.166667

```
from sklearn.feature_extraction.text import CountVectorizer
count_vectorizer = CountVectorizer(encoding='latin-1', ngram_range=(1, 1), tokenizer=None, analyzer = 'word', stop_
countvec = count_vectorizer.fit_transform(data.Abstrak).toarray()
countvec
```

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

```
countvec2 = pd.DataFrame(countvec)
countvec2
```

	0	1	2	3	4	5	6	7	8	9	...	6097	6098	6099	6100	6101	6102	6103	6104	6105	6106
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	3	...	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

```
kata_kata = count_vectorizer.get_feature_names()
countvec3 = pd.DataFrame(countvec, columns=kata_kata)
countvec3
```

	abad	abdullah	able	abnormal	abnormalitas	absolut	absolute	abstrak	acak	acara	...	Xjk	Yj	diarenyeri	expusia	muntah	nj	pusing	tebuP
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	1	0	0	3	...	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	

```
from xlsxwriter.utility import xl_rowcol_to_cell
saveresult = pd.ExcelWriter('ellaaatf.xlsx', engine='xlsxwriter')
countvec3.to_excel(saveresult, index=False, sheet_name='report')
saveresult.save()
```

```
from sklearn.feature_extraction.text import TfidfTransformer
transformer = TfidfTransformer(norm=None, use_idf=True, smooth_idf=False, sublinear_tf=False)
tfidf = transformer.fit_transform(countvec)
tfidf
```

```
<416x6107 sparse matrix of type '<class 'numpy.float64'>'
        with 26284 stored elements in Compressed Sparse Row format>
```

```
tfidf1 = tfidf.toarray()
tfidf1
```

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

```
tfidf2 = pd.DataFrame(tfidf1)
tfidf2
```

	0	1	2	3	4	5	6	7	8	9	...	6097	6098	6099	6100	6101	6102	6103	6104	6105	6106
0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.000000	4.63279	0.0	0.000000	21.092056	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

```
kata_kata2 = count_vectorizer.get_feature_names()
df1 = pd.DataFrame(tfidf1, columns=kata_kata2)
df1
```

	abad	abdullah	able	abnormal	abnormalitas	absolut	absolute	abstrak	acak	acara	...	Xjk	Yj	diarenyeri	expusia	muntah	nj	p
0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.000000	4.63279	0.0	0.000000	21.092056	...	0.0	0.0	0.0	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
7	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
8	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	
9	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	

```
from xlsxwriter.utility import xl_rowcol_to_cell
saveresult = pd.ExcelWriter('ellatfidf.xlsx', engine='xlsxwriter')
df1.to_excel(saveresult, index=False, sheet_name='report')
saveresult.save()
```

```
#Create Bigram & Trigram Models
from gensim.models import Phrases
# Add bigrams and trigrams to docs,minimum count 10 means only that appear 10 times or more.
bigram = Phrases(text_list, min_count=10)
trigram = Phrases(bigram[text_list])

for idx in range(len(text_list)):
    for token in bigram[text_list[idx]]:
        if '_' in token:
            # Token is a bigram, add to document.
            text_list[idx].append(token)
    for token in trigram[text_list[idx]]:
        if '_' in token:
            # Token is a bigram, add to document.
            text_list[idx].append(token)
```

```
from gensim import corpora, models
# Create a dictionary representation of the documents.
dictionary = corpora.Dictionary(text_list)

dictionary.filter_extremes(no_below=5, no_above=0.2)
#no_below (int, optional) - Keep tokens which are contained in at least no_below documents.
#no_above (float, optional) - Keep tokens which are contained in no more than no_above documents (fraction of total

print(dictionary)
```

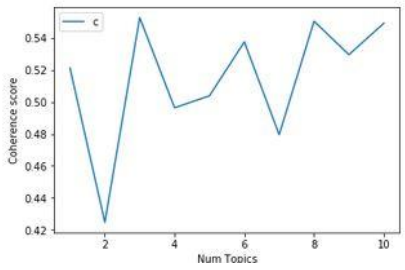
```
Dictionary(1359 unique tokens: ['badan', 'berhubungan', 'dampak', 'darah', 'dunia']...)
```

```
#https://radimrehurek.com/gensim/tut1.html
#build corpus

doc_term_matrix = [dictionary.doc2bow(doc) for doc in text_list]
#The function doc2bow converts document (a list of words) into the bag-of-words format

print(len(doc_term_matrix))
print(doc_term_matrix[100])

tfidf = models.TfidfModel(doc_term_matrix) #build TF-IDF model
corpus_tfidf = tfidf[doc_term_matrix]
```

```
416
[(5, 2), (6, 6), (7, 7), (8, 1), (67, 1), (82, 1), (231, 3), (256, 5), (280, 1), (294, 1), (410, 5), (415, 1),
(456, 1), (585, 1), (686, 1), (769, 1), (953, 1), (989, 1), (1050, 1), (1208, 1), (1209, 2), (1214, 1)]
```

```
from gensim.models.coherencemodel import CoherenceModel
from gensim.models.ldamodel import LdaModel
from gensim.corpora.dictionary import Dictionary
from numpy import array
#function to compute coherence values
def compute_coherence_values(dictionary, corpus, texts, limit, start, step):
    coherence_values = []
    model_list = []
    for num_topics in range(start, limit, step):
        model = LdaModel(corpus=corpus, id2word=dictionary, num_topics=num_topics, iterations=100)
        model_list.append(model)
        coherencemodel = CoherenceModel(model=model, texts=texts, dictionary=dictionary, coherence='c_v')
        coherence_values.append(coherencemodel.get_coherence())

    return model_list, coherence_values
```

```
start=1
limit=11
step=1
model_list, coherence_values = compute_coherence_values(dictionary, corpus=corpus_tfidf,
                                                         texts=text_list, start=start, limit=limit, step=step)
#show graphs
import matplotlib.pyplot as plt
x = range(start, limit, step)
plt.plot(x, coherence_values)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```



```
# Print the coherence scores
for m, cv in zip(x, coherence_values):
    print("Num Topics =", m, " has Coherence Value of", round(cv, 4))
```

```
Num Topics = 1  has Coherence Value of 0.5212
Num Topics = 2  has Coherence Value of 0.4247
Num Topics = 3  has Coherence Value of 0.5528
Num Topics = 4  has Coherence Value of 0.4964
Num Topics = 5  has Coherence Value of 0.5039
Num Topics = 6  has Coherence Value of 0.5377
Num Topics = 7  has Coherence Value of 0.4796
Num Topics = 8  has Coherence Value of 0.5506
Num Topics = 9  has Coherence Value of 0.5296
Num Topics = 10  has Coherence Value of 0.5493
```

```
?LdaModel
```

```
from pprint import pprint

model = LdaModel(corpus=corpus_tfidf, id2word=dictionary,random_state=4, num_topics=3)
pprint(model.print_topics())
```

```
[(0,
  '0.004*"convolutional_neural" + 0.004*"tingkat_akurasi" + '
  '0.004*"deep_learning" + 0.004*"kabupaten_sleman" + 0.004*"neural_network" + '
  '0.003*"citra" + 0.003*"analisis_regresi" + 0.003*"kendaraan_bermotor" + '
  '0.003*"tanaman" + 0.003*"cluster"'),
 (1,
  '0.004*"time_series" + 0.004*"indeks_harga" + 0.003*"metode_fuzzy" + '
  '0.003*"hasil_peramalan" + 0.003*"cluster" + 0.003*"peramalan" + '
  '0.003*"harga" + 0.003*"kelompok" + 0.003*"data_panel" + '
  '0.003*"curah_hujan"'),
 (2,
  '0.004*"cluster" + 0.004*"pasien" + 0.004*"provinsi_jawa" + '
  '0.004*"model_regresi" + 0.003*"tingkat_akurasi" + 0.003*"penyakit" + '
  '0.003*"kabupaten" + 0.003*"jawa" + 0.003*"kanker" + 0.003*"ekspresi_gen"')]
```

```
import pandas as pd
top_words_per_topic = []
for t in range(model.num_topics):
    top_words_per_topic.extend([(t, ) + x for x in model.show_topic(t, topn = 10)])

#pd.DataFrame(top_words_per_topic, columns=['Topic', 'Word', 'P']).to_csv("top_words.csv")
df = pd.DataFrame(top_words_per_topic, columns=['Topic', 'Word','P']).to_csv("top_words11.csv")
print(df)
```

```
None
```

```
import gensim
import pyLDAvis.gensim;pyLDAvis.enable_notebook()

data = pyLDAvis.gensim.prepare(model, corpus_tfidf, dictionary)
print(data)
pyLDAvis.save_html(data, 'ldagensimskripsi11.html')
```

```
C:\Users\Dell\Anaconda3\lib\site-packages\pyLDAvis\_prepare.py:257: FutureWarning: Sorting because non-concate
nation axis is not aligned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=True'.

To retain the current behavior and silence the warning, pass sort=False

  return pd.concat([default_term_info] + list(topic_dfs))
```

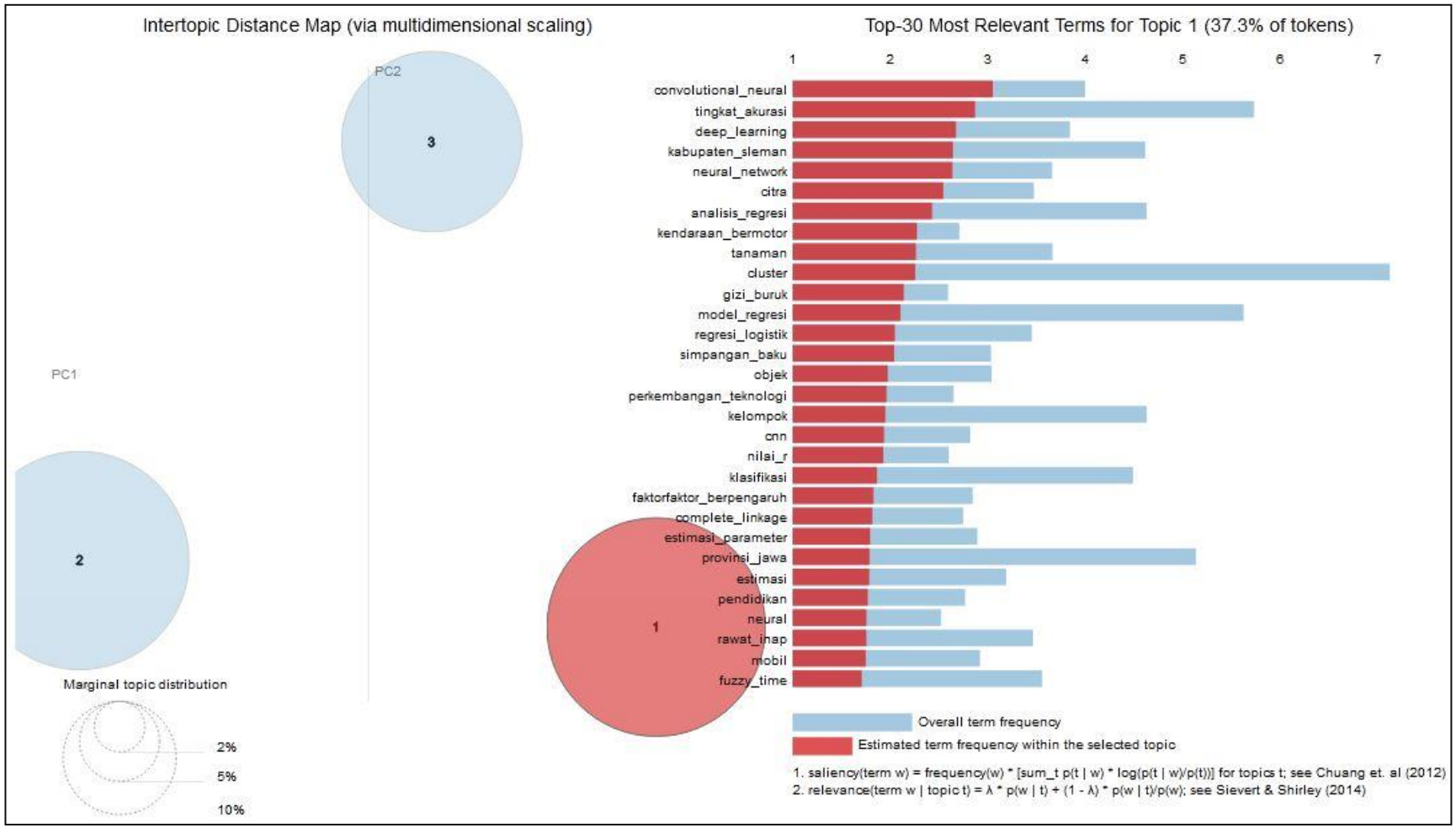
```
PreparedData(topic_coordinates=              x         y  topics  cluster       Freq
topic
0       0.022057 -0.015207       1        1  37.344780
2      -0.025566 -0.009700       2        1  37.246544
1       0.003510  0.024908       3        1  25.408676, topic_info=     Category      Freq                  Te
rm     Total  loglift  logprob
term
1235  Default  2.000000           indeks_harga  2.000000  30.0000  30.0000
751   Default  2.000000           metode_fuzzy  2.000000  29.0000  29.0000
1150  Default  2.000000             data_panel  2.000000  28.0000  28.0000
```

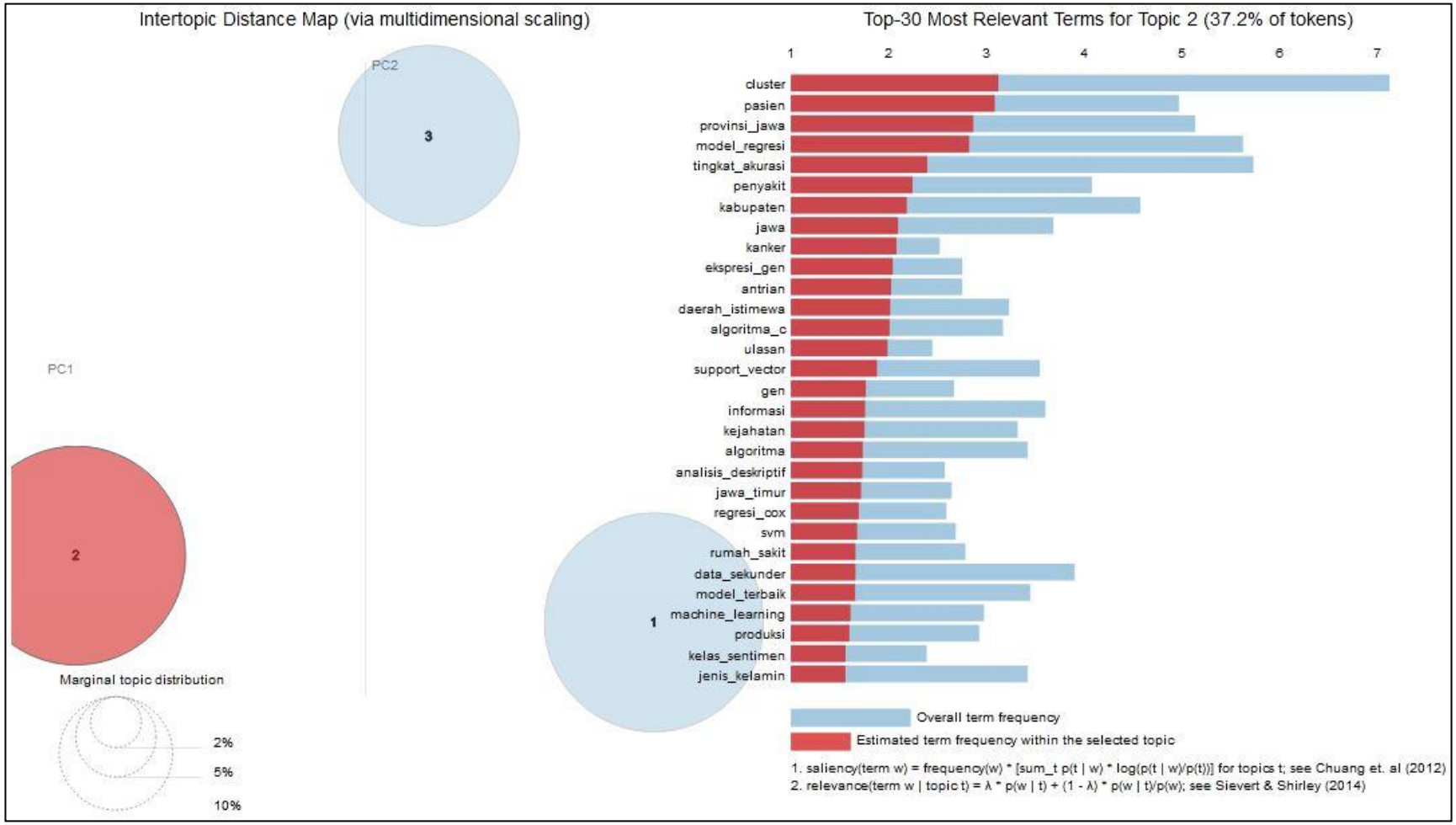
```
import matplotlib.pyplot as plt
from wordcloud import WordCloud as wd

for t in range(model.num_topics):
    plt.figure(figsize=(7,6))
    plt.imshow(wd(background_color="white",max_font_size=50, min_font_size=6).fit_words(dict(model.show_topic(t, 20
    plt.axis("off")
    plt.title("Topic #" + str(t))
    plt.savefig("wcld-topic11-#"+str(t)+".png", facecolor='k', bbox_inches='tight')

plt.show()
```

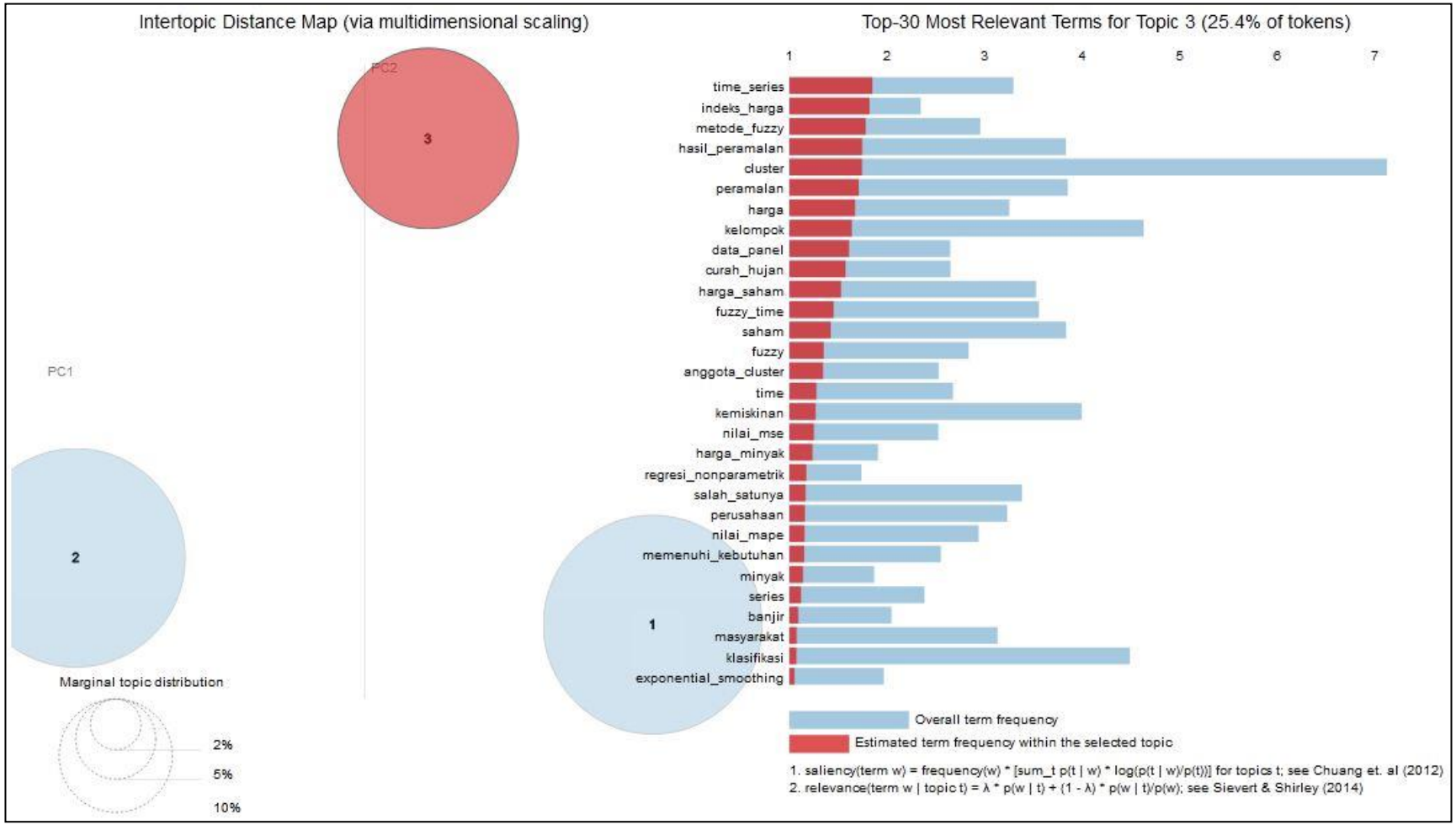
Link Script python lengkap:

http://bit.ly/ScriptLDAella

Lampiran 3 *Output* Visualisasi *Topic Modeling* dengan LDA




Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
2
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (37.2% of tokens)
1
2
3
4
5
6
7
cluster
pasien
provinsi_jawa
model_regresi
tingkat_akurasi
penyakit
kabupaten
jawa
kanker
ekspresi_gen
antrian
daerah_istimewa
algoritma_c
ulasan
support_vector
gen
informasi
kejahatan
algoritma
analisis_deskriptif
jawa_timur
regresi_cox
svm
rumah_sakit
data_sekunder
model_terbaik
machine_learning
produksi
kelas_sentimen
jenis_kelamin
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
2
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (25.4% of tokens)
1
2
3
4
5
6
7
time_series
indeks_harga
metode_fuzzy
hasil_peramalan
cluster
peramalan
harga
kelompok
data_panel
curah_hujan
harga_saham
fuzzy_time
saham
fuzzy
anggota_cluster
time
kemiskinan
nilai_mse
harga_minyak
regresi_nonparametrik
salah_satunya
perusahaan
nilai_mape
memenuhi_kebutuhan
minyak
series
banjir
masyarakat
klasifikasi
exponential_smoothing
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)