

**SKRIPSI**

**TOPIC MODELING DOKUMEN SKRIPSI PRODI  
PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER  
UNIVERSITAS NEGERI JAKARTA MENGGUNAKAN  
METODE *LATENT DIRICHLET ALLOCATION***



**EKI NUGRAHA**

**1512617010**

*Mencerdaskan dan  
Memartabatkan Bangsa*

**PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER  
FAKULTAS TEKNIK  
UNIVERSITAS NEGERI JAKARTA**

**2022**

# HALAMAN PENGESAHAN

Nama: Eki Nugraha

NIM: 1512617010

Telah menyelesaikan skripsi sebagai syarat untuk memperoleh gelar sarjana.

Dengan Judul:

**TOPIC MODELING DOKUMEN SKRIPSI PRODI PENDIDIKAN TEKNIK  
INFORMATIKA DAN KOMPUTER UNIVERSITAS NEGERI JAKARTA  
MENGUNAKAN METODE LATENT DIRICHLET ALLOCATION**

NAMA DOSEN	TANDA TANGAN	TANGGAL
Dr. Widodo, M.Kom. (Dosen Pembimbing I)		12 Agustus 2022
Murien Nugraheni, S.T., M.Cs. (Dosen Pembimbing II)		12 Agustus 2022

Telah disetujui oleh :

NAMA DOSEN	TANDA TANGAN	TANGGAL
M.Ficky Duskarnaen, S.T., M.Sc (Ketua Penguji)		10 Agustus 2022
Diat Nurhidayat, M.T.I (Dosen Penguji I)		12 Agustus 2022
Ali Idrus, S.Kom., M.Kom (Dosen Penguji II)		11 Agustus 2022

Tanggal Ujian : 15 Agustus 2022

## LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa:

1. Skripsi ini merupakan karya asli dan belum pernah diajukan untuk mendapatkan gelar akademik sarjana, baik di Universitas Negeri Jakarta maupun di Perguruan Tinggi lain.
2. Skripsi ini belum dipublikasikan, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan dalam daftar pustaka.
3. Pernyataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Universitas Negeri Jakarta.

Jakarta, 26 Juli 2022

Yang membuat pernyataan



*Mencerdaskan dan  
Memartabatkan Bangsa*

Eki Nugraha

No.Reg. 1512617010



KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN  
UNIVERSITAS NEGERI JAKARTA  
UPT PERPUSTAKAAN

Jalan Rawamangun Muka Jakarta 13220  
Telepon/Faksimili: 021-4894221  
Laman: [lib.unj.ac.id](http://lib.unj.ac.id)

**LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI  
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademika Universitas Negeri Jakarta, yang bertanda tangan di bawah ini, saya:

Nama : Eki Nugraha  
NIM : 1512617010  
Fakultas/Prodi : Pendidikan Teknik Informatika Dan Komputer  
Alamat email : eki.nugraha98@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada UPT Perpustakaan Universitas Negeri Jakarta, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah:

Skripsi     Tesis     Disertasi     Lain-lain (.....)

yang berjudul :

*Topic Modeling* Dokumen Skripsi Prodi Pendidikan Teknik Informatika Dan Komputer  
Universitas Negeri Jakarta Menggunakan Metode *Latent Dirichlet Allocation*

Dengan Hak Bebas Royalti Non-Eksklusif ini UPT Perpustakaan Universitas Negeri Jakarta berhak menyimpan, mengalihmediakan, mengelolanya dalam bentuk pangkalan data (*database*), mendistribusikannya, dan menampilkan/mempublikasikannya di internet atau media lain secara *fulltext* untuk kepentingan akademis tanpa perlu meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan Universitas Negeri Jakarta, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 15 Agustus 2022

(Eki Nugraha)

## KATA PENGANTAR

Segala puji dan syukur penulis haturkan kehadiran Allah Swt. Yang telah memberikan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini. Shalawat serta salam tak lupa terucapkan kepada junjungan Nabi Muhammad Saw yang menjadi panutan umat manusia sepanjang zaman sehingga penulis dapat menyelesaikan penyusunan skripsi yang berjudul “*Topic Modeling Dokumen Skripsi Prodi Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta Menggunakan Metode Latent Dirichlet Allocation*” yang merupakan salah satu syarat meraih gelar Sarjana Program Studi Pendidikan Teknik Informatika dan Komputer , Fakultas Teknik, Universitas Negeri Jakarta.

Selama penyusunan skripsi ini, penulis menyadari bahwa tidak terlepas dari dukungan, bantuan, dan doa dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih kepada:

1. Bapak Dr. Widodo, M.Kom. selaku Koordinator Program Studi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta sekaligus Dosen Pembimbing I yang telah memberikan bimbingan, arahan, serta masukan - masukan dalam penyusunan, sehingga penulis dapat menyelesaikan skripsi dengan baik;
2. Ibu Murien Nugraheni, S.T., M.Cs, selaku Dosen Pembimbing II yang telah memberikan bimbingan, arahan, serta masukan-masukan dalam penyusunan, sehingga penulis dapat menyelesaikan skripsi dengan baik;
3. Seluruh Bapak dan Ibu dosen Pendidikan Teknik Informatika dan Komputer yang telah memeberikan banyak ilmu selama kuliah di Universitas Negeri Jakarta;
4. Kedua Orang Tua yang telah memberikan dukungan doa, semangat, dan motivasi kepada penulis selama penyusunan skripsi;
5. Semua teman-teman Angkatan 2017 Prodi Pendidikan Teknik Informatika dan Komputer yang membantu dan terlibat secara langsung maupun tidak langsung dalam penulisan skripsi ini.

Penulis menyadari bahwa skripsi ini mempunyai keterbatasan, karena kemampuan dan pengetahuan yang dimiliki penulis dalam penyusunan skripsi ini. Oleh karena itu, penulis memohon maaf apabila terdapat kesalahan dan kekurangan dalam penyusunan skripsi ini. Penulis mengharapkan kritik dan saran untuk hasil yang lebih baik di kemudian hari. Semoga penelitian dan penyusunan skripsi ini dapat bermanfaat bagi pembaca pada umumnya dan bagi penulis khususnya.

Jakarta, 26 Juli 2022

Penulis,



Eki Nugraha

1512617010



*Mencerdaskan dan  
Memartabatkan Bangsa*

## ABSTRAK

**Eki Nugraha**, *Topic Modeling* Dokumen Skripsi Prodi Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta Menggunakan Metode *Latent Dirichlet Allocation*. Dosen Pembimbing: Dr. Widodo, S.Kom, M.Kom, Murien Nugraheni, S.T, M.Cs. Program Studi Pendidikan Teknik Informatika & Komputer. Fakultas Teknik. Universitas Negeri Jakarta 2022.

Skripsi atau tugas akhir menjadi suatu syarat untuk memperoleh gelar Sarjana Strata-1 di perguruan tinggi di Indonesia. Pada program studi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta mahasiswa mengambil judul dan tema skripsi yang beragam, tidak semua tema skripsi yang dibuat sesuai dengan konsentrasi jurusan masing-masing. Dengan semakin bervariasinya dokumen skripsi maka dilakukan pemodelan topik skripsi dengan menggunakan metode *Latent Dirichlet Allocation* untuk mengetahui komposisi topik skripsi yang dapat digunakan sebagai referensi pada penelitian selanjutnya. Bahan penelitian yang digunakan berasal dari *repository* admin Prodi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta berupa *softcopy* tahun lulus 2017-2022 sebanyak 329 dokumen skripsi. Dokumen skripsi diproses melalui beberapa tahapan yaitu *pre-processing* yang terdiri dari *casefolding*, *stopwords*, normalisasi, *stemming*, setelah itu dilakukan pemodelan LDA dengan menggunakan *library Gensim*, pembobotan kata dengan TF-IDF, pengujian topik menggunakan *coherence score* dan *perplexity*. Sehingga didapatkan 10 topik yang sering dibahas yaitu topik ke-1 mengenai Pengembangan media pembelajaran sebanyak 50 dokumen, topik ke-2 Evaluasi pembelajaran sebanyak 42 dokumen, topik ke-3 Pengembangan sistem informasi berbasis *website* sebanyak 42 dokumen, topik ke-4 Teknik komputer dan jaringan sebanyak 36 dokumen, topik ke-5 Pengembangan sistem informasi Universitas Negeri Jakarta berbasis *website* menggunakan *user interface*, *user experience* ,& *backend* sebanyak 36 dokumen, topik ke-6 Pengembangan dan evaluasi media pembelajaran *e-learning* menggunakan algoritma sebanyak 34 dokumen, topik ke-7 *data mining* sebanyak 28 dokumen, topik ke-8 Sistem *requirement* pengembangan perangkat lunak sebanyak 26 dokumen, topik ke-9 Standar pencapaian sumber daya Pendidikan sebanyak 14 dokumen, topik ke-10 Evaluasi pengembangan *website*, multimedia dan jaringan sebanyak 21 dokumen. Pengujian menggunakan *coherence score* menghasilkan 0.2614 dan *perplexity* -5.9895.

**Kata Kunci:** Skripsi, Pemodelan Topik, LDA, TF-IDF, *Coherence Score*, *Perplexity*

## ABSTRACT

**Eki Nugraha**, Topic of Thesis Document Modeling In Informatics and Computer Engineering Education, State University of Jakarta Using Latent Dirichlet Allocation Method. Supervisor: Dr. Widodo, S. Kom, M. Kom, Murien Nugraheni, S.T, M.Cs. Informatics & Computer Engineering Education Study Program. Faculty of Engineering. Jakarta State University 2022.

Thesis or final project is a requirement to obtain a bachelor's degree at a university in Indonesia. In the Informatics and Computer Engineering Education study program, State University of Jakarta, students take various thesis titles and themes, not all thesis themes are made according to the concentration of their respective majors. With the variety of thesis documents, the thesis topic modeling is carried out using the Latent Dirichlet Allocation method to determine the composition of the thesis topic which can be used as a reference in further research. The research material used comes from the admin repository of the Informatics and Computer Engineering Education Study Program, State University of Jakarta in the form of softcopy in the 2017-2022 graduation year as many as 329 thesis documents. The script document is processed through several stages, namely pre-processing which consists of casefolding, stopwords, normalization, stemming, after which LDA modeling is carried out using the Gensim library, word weighting with TF-IDF, topic testing using coherence scores and perplexity. So that there are 10 topics that are often discussed, namely the 1st topic regarding the development of learning media as many as 50 documents, the 2nd topic Evaluation of learning as many as 42 documents, the 3rd topic Development of website-based information systems as many as 42 documents, the 4th topic Computer engineering and a network of 36 documents, topic 5 Development of a website-based information system at the State University of Jakarta using a user interface, user experience, & backend as many as 36 documents, topic 6 Development and evaluation of e learning learning media using algorithms as many as 34 documents, topic 6 7 data mining 28 documents, the 8th topic System requirements software development 26 documents, the 9th topic Education resource achievement standards 14 documents, the 10th topic Evaluation of website, multimedia and network development 21 documents. Tests using a coherence score resulted in 0.2614 and a perplexity of -5.9895.

*Memartabatkan Bangsa*

**Keywords:** Thesis, Topic Modeling, LDA, TF-IDF, Coherence Score, Perplexity

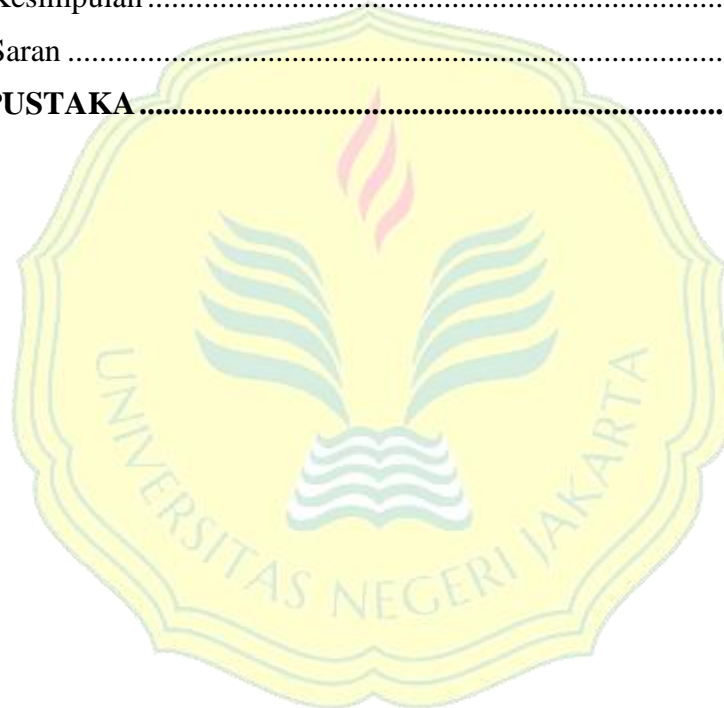


# DAFTAR ISI

<b>HALAMAN PENGESAHAN</b> .....	<b>i</b>
<b>LEMBAR PERNYATAAN</b> .....	<b>ii</b>
<b>KATA PENGANTAR</b> .....	<b>iii</b>
<b>ABSTRAK</b> .....	<b>v</b>
<b>ABSTRACT</b> .....	<b>vi</b>
<b>DAFTAR ISI</b> .....	<b>vii</b>
<b>DAFTAR GAMBAR</b> .....	<b>x</b>
<b>DAFTAR TABEL</b> .....	<b>xii</b>
<b>DAFTAR LAMPIRAN</b> .....	<b>xiii</b>
<b>BAB I PENDAHULUAN</b> .....	<b>1</b>
1.1    Latar Belakang Masalah .....	1
1.2    Identifikasi Masalah.....	4
1.3    Pembatasan Masalah.....	5
1.4    Perumusan Masalah .....	5
1.5    Tujuan Penelitian .....	5
1.6    Manfaat Penelitian .....	5
<b>BAB II KAJIAN PUSTAKA</b> .....	<b>7</b>
2.1    Kerangka Teoritik .....	7
2.1.1    Skripsi .....	7
2.1.2 <i>Data Mining</i> .....	8
2.1.3    Tahap-Tahap <i>Data Mining</i> .....	9
2.1.4 <i>Text Mining</i> .....	10
2.1.5 <i>Pre-Processing</i> .....	11
2.1.6 <i>Python</i> .....	12
2.1.7 <i>Bag Of Words</i> .....	13
2.1.8 <i>Coherence Score</i> .....	13
2.1.9 <i>Topic Modeling</i> .....	14
2.1.10 <i>Latent Dirichlet Allocation (LDA)</i> .....	14
2.1.11    Pembobotan TF-IDF .....	18
2.1.12 <i>Perplexity</i> .....	19
2.2    Penelitian yang Relevan.....	19

2.3	Kerangka Berpikir.....	25
<b>BAB III METODOLOGI PENELITIAN .....</b>		<b>27</b>
3.1.	Tempat, Waktu dan Subjek Penelitian.....	27
3.2.	Alat dan Bahan Penelitian.....	27
3.2.1	Alat.....	27
3.2.2	Bahan.....	28
3.3.	Diagram Alir Penelitian .....	29
3.4.	Teknik dan Prosedur Pengumpulan Data.....	29
3.5.	Proses Labeling Data .....	30
3.6.	<i>Pre-Processing</i> .....	30
3.6.1	Case Folding .....	31
3.6.2	Tokenizing.....	32
3.6.3	Stopwords.....	33
3.6.4	Lemmatization.....	35
3.6.5	Stemming .....	36
3.7.	Penerapan <i>Latent Dirichlet Allocation (LDA)</i> .....	36
3.8.	Visualisasi Data .....	37
3.9.	Teknik Analisis Data .....	39
3.9.1	Coherence Score.....	39
3.9.2	Perplexity .....	40
<b>BAB IV HASIL PENELITIAN.....</b>		<b>41</b>
4.1	Deskripsi Hasil Penelitian.....	41
4.1.2.	<i>Pre-Processing</i> .....	42
4.1.2.1	Case Folding .....	43
4.1.2.2	Tokenizing .....	45
4.1.2.3	Stopwords .....	48
4.1.2.4	Lemmatization .....	50
4.1.2.5	Stemming .....	52
4.1.3.	<i>Coherence Score</i> .....	54
4.1.4.	<i>Latent Dirichlet Allocation (LDA)</i> .....	55
4.1.4.1	Pembuatan kamus dan pembentukan <i>corpus</i> .....	55
4.1.4.2	Pemodelan LDA .....	56
4.1.4.3	Visualisasi LDA.....	59
4.1.5.	Perhitungan Jumlah Kata .....	70

4.1.6.	Perhitungan Jumlah Topik .....	72
4.1.7.	Pembobotan TF-IDF .....	72
4.2	Analisis Data Penelitian .....	74
4.2.1	<i>Coherence Score</i> .....	75
4.2.1	<i>Perplexity</i> .....	77
4.3	Pembahasan .....	78
4.4	Aplikasi Hasil Penelitian .....	78
<b>BAB V</b>	<b>PENUTUP</b> .....	<b>80</b>
5.1	Kesimpulan .....	80
5.2	Saran .....	80
<b>DAFTAR PUSTAKA</b>	.....	<b>82</b>



*Mencerdaskan dan  
Memartabatkan Bangsa*

## DAFTAR GAMBAR

Gambar 2.1 Tahap-tahap <i>Data Mining</i> (Han, 2012 :45).....	9
Gambar 2.2 <i>Topic Modeling</i> (Blei dkk., 2012 :78).....	15
Gambar 2.3 Inversi dengan <i>LDA</i> (Blei dkk., 2012: 79).....	15
Gambar 2.4 Representasi <i>LDA</i> (Blei dkk, 2012 :81).....	17
Gambar 2.5 Kerangka Berpikir Penelitian.....	26
Gambar 3.1 Diagram Alir Penelitian.....	29
Gambar 3.2 Model Representasi <i>LDA</i> (Blei dkk, 2012 :81).....	37
Gambar 3.3 Ilustrasi Hasil <i>PyLDAvis</i> .....	38
Gambar 3.4 <i>Word Cloud</i> Topik ke-1.....	38
Gambar 3.5 <i>Coherence</i> Dengan Limit 1-10.....	40
Gambar 4.1 Jumlah Dokumen Skripsi Setiap Tahun.....	42
Gambar 4.2 <i>Topic Coherence Measures</i> (Roder dkk., 2015 : 406).....	54
Gambar 4.3 <i>PyLDAvis</i> Topik ke -1.....	59
Gambar 4.4 <i>Wordcloud</i> Topik ke -1.....	60
Gambar 4.5 <i>PyLDAvis</i> Topik ke -2.....	60
Gambar 4.6 <i>Wordcloud</i> Topik ke -2.....	61
Gambar 4.7 <i>PyLDAvis</i> Topik ke -3.....	61
Gambar 4.8 <i>Wordcloud</i> Topik ke -3.....	62
Gambar 4.9 <i>PyLDAvis</i> Topik ke -4.....	62
Gambar 4.10 <i>Wordcloud</i> Topik ke -4.....	63
Gambar 4.11 <i>PyLDAvis</i> Topik ke -5.....	63
Gambar 4.12 <i>Wordcloud</i> Topik ke -5.....	64
Gambar 4.13 <i>PyLDAvis</i> Topik Ke-6.....	64
Gambar 4.14 <i>Wordcloud</i> Topik ke -6.....	65
Gambar 4.15 <i>PyLDAvis</i> Topik Ke-7.....	65
Gambar 4.16 <i>Wordcloud</i> Topik ke -7.....	66
Gambar 4.17 <i>PyLDAvis</i> Topik Ke-8.....	66
Gambar 4.18 <i>Wordcloud</i> Topik ke -8.....	67
Gambar 4.19 <i>PyLDAvis</i> Topik Ke-9.....	67
Gambar 4.20 <i>Wordcloud</i> Topik ke -9.....	68
Gambar 4.21 <i>Wordcloud</i> Topik ke -10.....	68
Gambar 4.22 <i>Wordcloud</i> Topik ke -10.....	69
Gambar 4.23 <i>Coherence Score</i> limit 1-10.....	75

Gambar 4.24 Coherence Score limit 1-20..... 76  
Gambar 4.25 Coherence Score limit 1-30..... 76



*Mencerdaskan dan  
Memartabatkan Bangsa*

## DAFTAR TABEL

Tabel 2.1 Perbedaan Penelitian Kuantitatif dan Kualitatif.....	8
Tabel 2.2 Penelitian Relevan.....	21
Tabel 3.1 Daftar Perangkat Keras .....	27
Tabel 3.2 Daftar Perangkat Lunak .....	28
Tabel 3.3 Proses Labeling Data .....	30
Tabel 3.4 Proses <i>Case Folding</i> .....	31
Tabel 3.5 Proses <i>Tokenizing</i> .....	32
Tabel 3.6 Proses <i>Stopwords</i> .....	33
Tabel 3.7 Proses <i>Lemmatization</i> .....	35
Tabel 3.8 Proses <i>Stemming</i> .....	36
Tabel 3.9 Perhitungan TF dan IDF .....	39
Tabel 3.10 Contoh <i>Coherence Score</i> .....	40
Tabel 3.11 Contoh <i>Perplexity</i> .....	40
Tabel 4.1 Proses <i>Case Folding</i> .....	43
Tabel 4.2 Proses <i>Tokenizing</i> .....	45
Tabel 4.3 Proses <i>Stopwords</i> .....	48
Tabel 4.4 Proses <i>Lemmatization</i> .....	50
Tabel 4.5 Proses <i>Stemming</i> .....	52
Tabel 4.6 <i>coherence score</i> 10 topik .....	55
Tabel 4.7 Pembentukan kamus LDA .....	55
Tabel 4.8 Pembentukan indeks <i>Corpus</i> LDA .....	56
Tabel 4.9 Model LDA 10 Topik .....	56
Tabel 4.10 Pemodelan 10 Topik .....	69
Tabel 4.11 Perhitungan Jumlah Kata .....	70
Tabel 4.12 Perhitungan Jumlah Topik .....	72
Tabel 4.13 Hasil TF.....	73
Tabel 4.14 Hasil pembobotan TF-IDF .....	73
Tabel 4.15 Sampel Hasil TF-IDF kata “ <i>algoritma</i> ” .....	74
Tabel 4.16 <i>Coherence Score limit</i> 1-10.....	75
Tabel 4.17 <i>Coherence Score limit</i> 1-20.....	75
Tabel 4.18 <i>Coherence Score limit</i> 1-30.....	76
Tabel 4.19 Nilai <i>Perplexity</i> .....	78
Tabel 4.20 <i>Coherence Score</i> dan <i>Perplexity</i> .....	78

## DAFTAR LAMPIRAN

Lampiran 1. Surat Tugas Dosen Pembimbing Skripsi .....	87
Lampiran 2. Surat Bimbingan Dosen Pembimbing I .....	88
Lampiran 3. Surat Bimbingan Dosen Pembimbing II .....	89
Lampiran 4. Surat Izin Penelitian .....	90
Lampiran 5. Surat Pernyataan Persetujuan Pendaftaran Sidang Skripsi Dosen Pembimbing I .....	91
Lampiran 6. Surat Pernyataan Persetujuan Pendaftaran Sidang Skripsi Dosen Pembimbing II .....	92
Lampiran 7. Data Skripsi Tahun 2017-2022 .....	93
Lampiran 8. <i>Script Topic Modeling</i> .....	98



*Mencerdaskan dan  
Memartabatkan Bangsa*