

di dalam dokumen yang diberikan menunjukkan seberapa penting kata tersebut di dalam dokumen. Untuk menghitung bobot setiap *term* di dokumen, digunakan persamaan 2.1 [23]:

$$IDF = \log\left(\frac{D}{df}\right) W_{d,t} = tf_{d,t} \cdot IDF \cdot t \quad (2.1)$$

Dimana:

d: dokumen ke-d

t: kata ke-t dari kata kunci

W: bobot dokumen ke-d terhadap kata ke-t

D: Total dokumen

df: banyak dokumen yang mengandung kata yang dicari

tf: banyak kata yang dicari pada sebuah dokumen

2.2.7 *Cosine Similarity*

Cosine Similarity digunakan untuk melakukan perhitungan kesamaan dari dokumen. Rumus *Cosine Similarity* seperti yang ditunjukkan pada persamaan 2.2 [24]:

$$\cos a = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (2.2)$$

Keterangan:

A = Vektor A, yang akan dibandingkan kemiripan

B = Vektor B, yang akan dibandingkan kemiripan

A.B = dot product antara vektor A dan vektor B

|A| = Panjang vektor A

|B| = Panjang vektor B

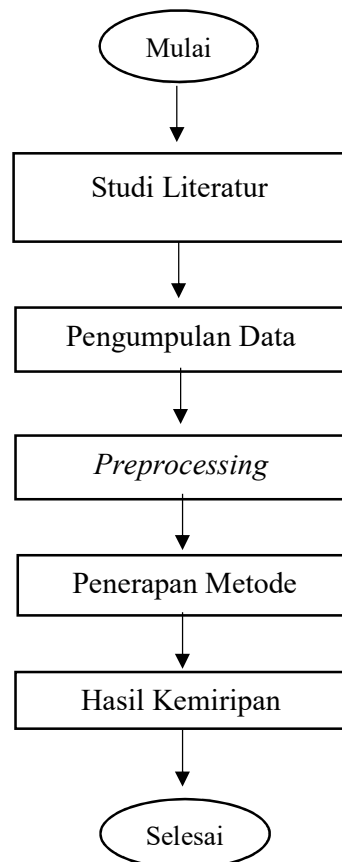
|A||B| = cross product antara |A| dan |B|

Metode yang digunakan pada penelitian ini adalah dengan cara membandingkan kesamaan atau similaritas antara dokumen abstrak yang akan dicek dengan dokumen abstrak yang lain. Kemudian dicari jumlah similaritas tertinggi antara dokumen abstrak yang di cek dengan dokumen yang abstrak lain yang dimiliki. Apabila total similaritas yang didapatkan adalah nol (0) maka dokumen yang diolah tidak memiliki kesamaan dan jika nilai yang didapatkan maksimal adalah 1 maka dokumen tersebut memiliki kemiripan.

BAB III METODE PENELITIAN

3.1 Tahapan Penelitian

Penelitian ini menggunakan penelitian kualitatif akan mencoba mengimplementasikan metode *Word2vec* untuk mendeteksi kemiripan dokumen dengan tujuan mengetahui kemiripan antara satu dokumen dengan dokumen yang dimiliki. Tahapan yang dilakukan dalam penelitian adalah :



Gambar 3. 1 Tahapan penelitian

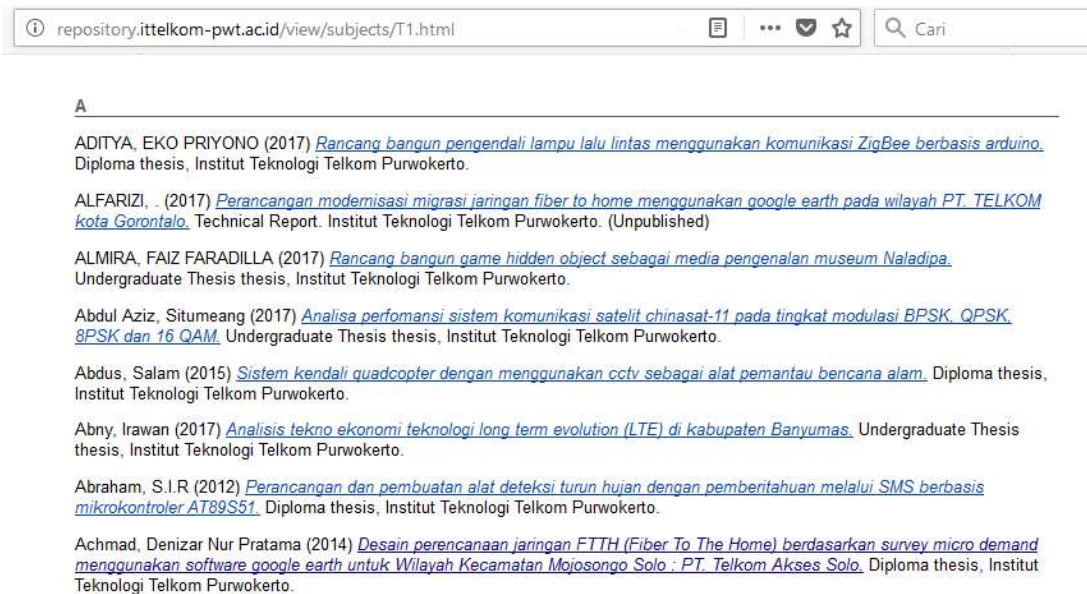
Tahapan pelaksanaan penelitian diatas menjelaskan tahapan demi tahapan penelitian yang akan dilakukan selama melaksanakan penelitian. Berikut penjelasan tahapan-tahapan tersebut adalah :

3.1.1 Studi Literatur

Pada penelitian ini dibutuhkan referensi sebagai dasar dalam melakukan tahapan penelitian, penulis menggunakan jurnal terdahulu yang masih berhubungan dengan pengolahan teks sebagai referensi. Selain jurnal terdahulu penulis juga menggunakan buku untuk memperluas pengetahuan dan mempelajari cara kerja dari metode yang akan digunakan.

3.1.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dokumen abstrak tugas akhir yang berekstensi .docx, .txt, dan .pdf. dokumen didapat dari perpustakaan berupa *file* tugas akhir Institut Teknologi Telkom Purwokerto dokumen tersebut akan menjadi bahan untuk membandingkan kemiripan antar dokumen.



Gambar 3. 2 Website repository perpustakaan

Gambar 3.2 merupakan *website repository* Institut Teknologi Telkom Purwokerto. Dalam *website* tersebut terdapat kumpulan data skripsi mahasiswa dan penelitian yang dilakukan oleh dosen maupun mahasiswa. Penulis mengumpulkan sebanyak 116 abstrak dari tahun 2012 sampai 2017

Tabel 3. 1 Abstrak Skripsi

Abstrak 1	Abstrak 2
<p>Perkembangan dibidang teknologi komunikasi memiliki peran sangat penting di dalam mendukung komunikasi berbasis broadband service yang berkualitas. Untuk memenuhi kebutuhan masyarakat khususnya pengguna jasa telekomunikasi dari kalangan rumah tangga dihadirkanlah teknologi jaringan FTTH (Fiber to the home). Pada Tugas akhir ini dilakukan perencanaan pembangunan jaringan FTTH di wilayah Solo Baru, Kecamatan Grogol, Solo dengan metode survey menggunakan aplikasi Google earth dan KML CSV converter. Desain jaringan FTTH ini dibuat supaya dapat menjadi solusi teknologi akses dari sentral sampai pada pelanggan. Tahap awal yang dilakukan adalah melakukan survey homepass dan mendapatkan data seperti alamat rumah, kriteria rumah, dan status berlangganan layanan Telkom atau tidak. Data survey dimasukkan kedalam aplikasi Google earth untuk dipetakan calon pelanggan dan KML CSV converter untuk pemberian nama pelanggan sesuai dengan ketentuan. Kemudian menentukan perangkat dan material yang dibutuhkan seperti ODC, ODP, Tiang, Kabel Feeder, Kabel Distribusi, Kabel Drop agar distribusi jaringan dari sentral bisa sampai pada pelanggan secara efisien dan ideal. Kemudian menghitung Link Budget dengan 3 kondisi one stage 1:32, two stage 1:2 => 1:4 dan 1:2 =>1:16 pada Boundary 5 dan membandingkan diantara ketiga kondisi tersebut untuk mengetahui konfigurasi yang paling tepat. Dan jarak yang dihitung adalah jarak paling terdekat, menengah, dan terjauh dari STO sampai dengan ONT. Dari perbandingan jaringan tersebut dapat disimpulkan dengan melihat rata-rata yang dihasilkan pada skenario 1:32 menghasilkan rata-rata redaman sebesar 24.97 dB, pada skenario 1:2/1:16 menghasilkan rata-rata redaman sebesar 21.96 dB, dan pada skenario 1:4/1:8 menghasilkan rata-rata redaman sebesar 21.14 dB. Maka pada daerah Solo baru lebih cocok menggunakan skenario 1:4/1:8 karena redaman yang dihasilkan lebih kecil yaitu 21.14 dB dibandingkan dengan menggunakan skenario 1:32 dan 1:2/1:16.</p>	<p>Perkembangan teknologi dari tahun ke tahun semakin berkembang, begitu juga dengan kebutuhan komunikasi yang semakin meningkat, mulai dari layanan <i>voice</i>, data, dan <i>video</i>. Pada setiap layanan diperlukan kualitas yang lebih baik dalam mengakses layanan tersebut. Serta diperlukan jaringan yang handal yang mampu memberikan performansi yang baik. Dengan mengganti jaringan telekomunikasi tembaga saat ini dengan jaringan telekomunikasi fiber optik diharapkan dapat meningkatkan kualitas layanan yang lebih baik. Dengan melakukan perencanaan jaringan <i>Fiber To The Home</i> (FTTH) diharapkan dapat menjadikan solusi untuk menggantikan jaringan tembaga ke jaringan fiber optik. Saat ini untuk daerah klaten utara belum menggunakan jaringan FTTH dan untuk layanan pada daerah Klaten Utara masih menggunakan satu layanan yaitu layanan telepon, belum menggunakan layanan <i>video</i> dan data. Untuk dapat menggunakan layanan <i>video</i> dan data, maka dilakukannya perencanaan jaringan FTTH pada daerah Klaten Utara. Pada perencanaan jaringan FTTH untuk daerah Klaten Utara dilakukannya <i>survey</i> untuk mendapatkan data-data seperti tipe rumah, di huni atau tidak, berlangganan telkom atau tidak, dan alamat rumah didaerah tersebut. Untuk dapat mengetahui skenario mana yang cocok di daerah tersebut, maka pada desain perencanaan jaringan FTTH dibuat dengan 3 skenario <i>one stage</i> 1:32, <i>two stage</i> 1:2/1:16, dan 1:4/1:8, serta dibuat <i>link budget</i> untuk jarak terdekat, menengah, dan terjauh dari <i>Optical Distribution Frame</i> (ODF)-<i>Optical Network Termination</i> (ONT), serta dibuat <i>Bill Of Quantity</i> (BOQ). Dari perbandingan jaringan tersebut dapat disimpulkan dengan melihat rata-rata redaman yang dihasilkan pada skenario 1:32 menghasilkan rata-rata redaman sebesar 22.32 dB, pada 1:2/1:16 menghasilkan rata-rata redaman sebesar 22.47 dB, dan pada skenario 1:4/1:8</p>

Abstrak 1	Abstrak 2
<p>Dan dari ketiga skenario tersebut tidak ada yang melebihi batas maksimal toleransi redaman yaitu 28 dB. Tabel BoQ (Bill of Quantity) yang dihasilkan pada Boundary 5 adalah Kabel Feeder 1.632 kilometer, Kabel Distribusi 5.9391 kilometer, Kabel Drop 62.5 Kilometer, ODC 1 Buah, ODP 78 Buah, konektor 6875 Buah, dan Tiang 22 buah.</p>	<p>menghasilkan rata-rata redaman sebesar 22.28 dB, maka pada daerah klaten utara lebih cocok menggunakan skenario 1:4/1:8 karena redaman yang dihasilkan lebih kecil yaitu sebesar 22.28 dB dibanding dengan skenario 1:32 dan 1:2/1:16 dan dari ketiga skenario tersebut tidak ada yang melebihi batas maksimal toleransi redaman yaitu sebesar 28 dB.</p>

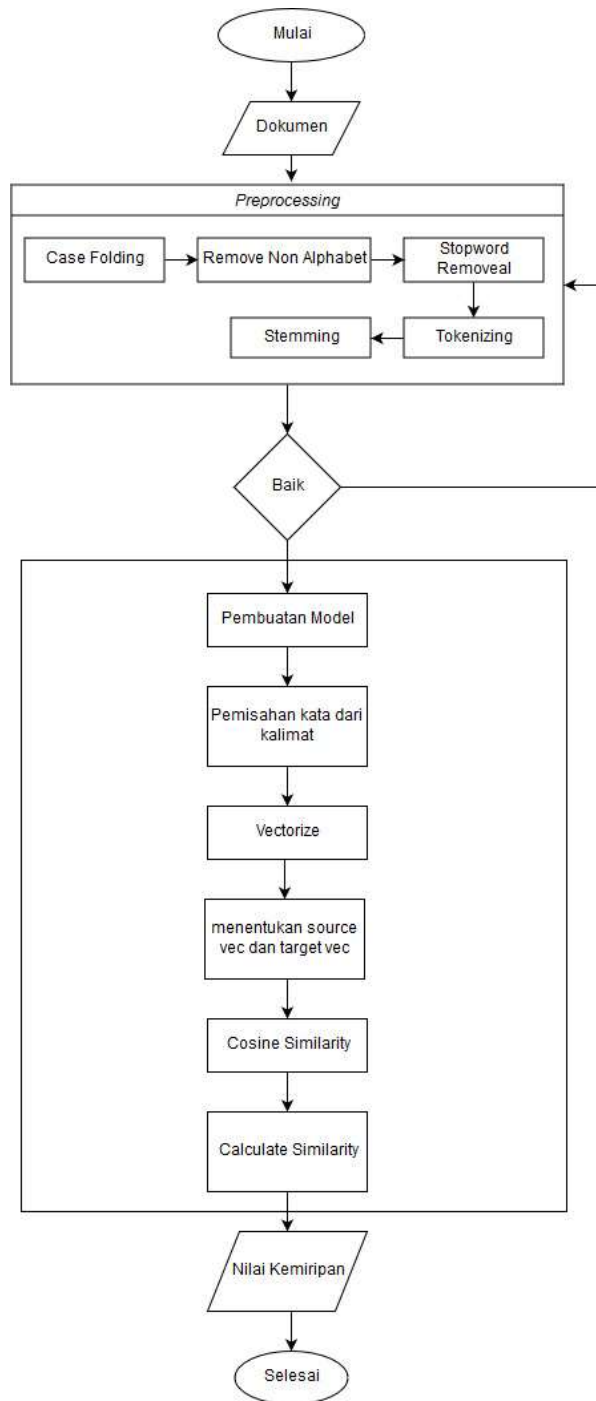
Pada tabel 3.2 Merupakan contoh data abstrak skripsi yang didapatkan dari *website repository* Institut Teknologi Telkom Purwokerto. Data tersebut nantinya akan di olah oleh sistem kemiripan dokumen yang akan dibuat oleh penulis.

3.1.3 *Preprocessing*

Tahapan yang dilakukan setelah mendapatkan data berupa dokumen abstrak yang akan digunakan untuk penelitian yaitu *preprocessing* untuk mengurangi noise pada proses deteksi kemiripan dokumen. Proses dalam tahapan ini adalah proses berupa merubah karakter menjadi bentuk yang sama (*case folding*), menghilangkan kata yang tidak perlu (*remove stopword*), memisahkan kata (*tokenizing*), dan menghilangkan kata imbuhan (*stemming*).

3.1.4 Penerapan Metode

Data yang sudah diolah pada tahap *preprocessing* selanjutnya disimpan menjadi satu penyimpanan *dataset* dokumen pembandingan yang nantinya dokumen uji akan dicek kemiripan dengan dokumen pembandingan dan sudah dilakukan *preprocessing*. Penulis menggunakan bahasa pemrograman *python* untuk melakukan percobaan. Metode *Word2vec* digunakan untuk megubah kata-kata menjadi *vector* dengan tujuan untuk mencari nilai kedekatan *vector* antar kata. Urutan proses *Word2vec* dapat dilihat pada gambar 3.3 dibawah ini.



Gambar 3. 3 Gambaran pemodelan *Word2vec*

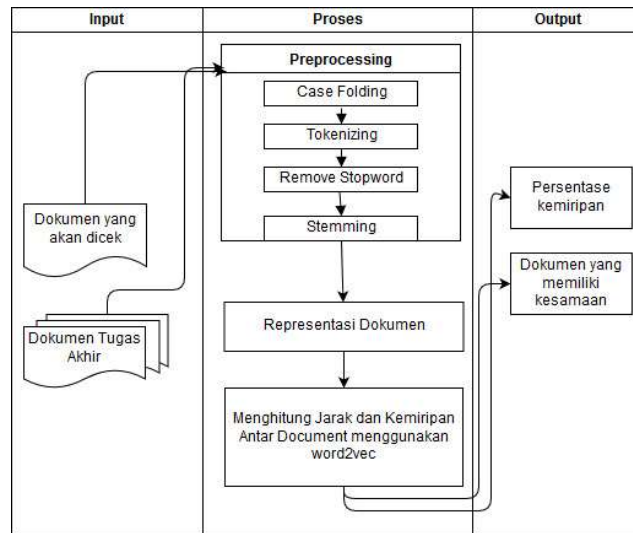
Pada gambar 3.3 yaitu sistem membaca seluruh korpus yang sudah dilakukan *preprocessing* dimana data yang dibaca berupa kata-kata pada suatu kalimat yang telah diubah kedalam bentuk array. Selanjutnya yaitu pembuatan model *Word2vec*

sebagai data tes. Setelah proses pembuatan model selesai, maka sistem menghasilkan vektor-vektor dari setiap kata dari data korpus. *Library* yang digunakan dalam percobaan yaitu *Gensim*. Dalam *Word2vec*, setiap satu kata bisa memiliki lebih dari satu *vector* hal ini dikarenakan setiap kata pada sebuah kalimat memiliki konteks yang berbeda.

3.1.5 Hasil Kemiripan

Setelah dilakukan proses perhitungan *vector* dengan *Word2vec*, maka sistem akan membaca *vector* dari dokumen uji dan dokumen pembanding sebanyak 116 dokumen lalu melakukan perhitungan kemiripan dengan menggunakan rumus *Cosine Similarity*. Apabila total similaritas kemiripan yang didapatkan adalah nol (0) maka dokumen yang diolah tidak memiliki kesamaan dan jika nilai yang didapatkan maksimal adalah 1 maka dokumen tersebut memiliki kemiripan.

3.2 Gambaran Umum Sistem



Gambar 3. 4 Gambaran Umum Sistem

Pada gambar 3.4 menjelaskan alur program yang akan dibuat. Dokumen uji dan dokumen pembanding dilakukan proses preprocessing selanjutnya akan menampilkan isi teks yang sudah di preprocessing. Kemudian akan dihitung nilai kemiripan menggunakan cosine similarity dan *Word2vec*. Hasil kemiripan akan muncul dari yang tertinggi ke terendah beserta nama dokumen yang memiliki kemiripan.

BAB IV

HASIL PENELITIAN DAN ANALISIS

4.1 Pengumpulan Data

Pada penelitian ini pengumpulan data yang dilakukan dengan cara mengunduh dari *website* <http://repository.itelkom-pwt.ac.id>. Data yang digunakan berupa teks abstrak berbahasa Indonesia yang terdapat pada tahun 2012 sampai 2017. Total data yang digunakan oleh penulis yaitu 116 abstrak. Data tersebut akan digunakan sebagai dokumen pembanding yang nantinya dokumen uji akan dideteksi kemiripannya dengan dokumen pembanding tersebut.

4.2 *Preprocessing*

Setelah mendapatkan dokumen pembanding berupa abstrak dan dokumen uji terlebih dahulu harus melalui tahapan *preprocessing* sebelum dideteksi kemiripan. *Preprocessing* yaitu memperbaiki teks yang ada dalam dokumen, pada tahap *preprocessing* terdapat beberapa proses yang dilakukan diantaranya yaitu *case folding*, *remove stopword*, *remove alphabet*, *tokenizing*, *stemming*. Proses tersebut dilakukan secara langsung dalam tahapan *preprocessing*. Berikut adalah contoh dokumen uji yaitu abstrak37 (Q) dan dokumen pembanding yaitu abstrak 33 (D) yang telah dilakukan *preprocessing* dan *tokenizing*.

Tabel 4. 1 Sebelum dan Sesudah *Preprocessing*

Dokumen Q	<i>Preprocessing</i>	<i>Tokenizing</i>
Meningkatnya pengguna internet tidak diimbangi dengan sistem keamanan yang tinggi. Data-data yang ada di dalam jaringan masih dapat diakses oleh orang lain yang berada dalam jaringan yang sama. VPN yang menawarkan teknologi enkripsi data sehingga data dari	tingkat guna internet imbang sistem keman tinggi data jaring akses orang ada jaring sama vpn tawar teknologi enkripsi data data kirim pasti aman tuju beri efisiensi kirim data perlu buah manajemen trafik guna beberapa teori antri beda teliti laku simulasi amat	[['tingkat', 'guna', 'internet', 'imbang', 'sistem', 'keman', 'tinggi', 'data', 'jaring', 'akses', 'orang', 'ada', 'jaring', 'sama', 'vpn', 'tawar', 'teknologi', 'enkripsi', 'data', 'data', 'kirim', 'pasti', 'aman', 'tju', 'beri', 'efisiensi', 'kirim', 'data', 'perlu', 'buah', 'manajemen', 'trafik', 'guna', 'beberapa', 'teori', 'antri',

Dokumen Q	<i>Preprocessing</i>	<i>Tokenizing</i>
<p>pengirim dipastikan aman sampai ke tujuan. Untuk memberikan efisiensi dalam pengiriman data, diperlukan sebuah manajemen trafik, seperti menggunakan beberapa teori antrian yang berbeda. Dalam penelitian ini dilakukan simulasi dan pengamatan terhadap parameter aplikasi VPN menggunakan 3 skenario dengan masing-masing skenario menggunakan teori antrian yang berbeda, yaitu First In First Out (FIFO), Priority Queuing (PQ), dan Weighted Fair Queuing (WFQ). Pada penelitian ini, jaringan VPN diterapkan pada teknologi WAN. Penelitian ini bertujuan untuk mengetahui teori antrian mana yang cocok diterapkan pada jaringan VPN. Simulasi dilakukan menggunakan Software Opnet Modeler 14.5. Jenis layanan yang digunakan yaitu Video conference, VoIP, FTP, dan HTTP. Parameter yang diamati yaitu: Delay, Delay Variation, Traffic Dropped, dan Throughput. Dari hasil simulasi terlihat bahwa teori antrian WFQ paling cocok digunakan pada</p>	<p>parameter aplikasi vpn guna skenario masing-masing skenario guna teori antri beda first in first out fifo priority queuing pq weighted fair queuing wfq teliti jaring vpn terap teknologi wan teliti tuju tahu teori antri mana cocok terap jaring vpn simulasi laku guna software opnet modeler jenis layan guna video conference voip ftp http parameter amat delay delay variation traffic dropped throughput hasil simulasi lihat teori antri wfq paling cocok guna vpn milik traffic dropped paling kecil besar paket throughput paling tinggi besar bit delay delay variation teori antri wfq layan video conference besar teori antri pq nilai besar layan voip teori antri wfq milik delay delay variation besar teori antri pq nilai besar kata kunci vpn fifo pq wfq opnet modeler</p>	<p>'beda', 'teliti', 'laku', 'simulasi', 'amat', 'parameter', 'aplikasi', 'vpn', 'guna', 'skenario', 'masing-masing', 'skenario', 'guna', 'teori', 'antri', 'beda', 'first', 'in', 'first', 'out', 'fifo', 'priority', 'queuing', 'pq', 'weighted', 'fair', 'queuing', 'wfq', 'teliti', 'jaring', 'vpn', 'terap', 'teknologi', 'wan', 'teliti', 'taju', 'tahu', 'teori', 'antri', 'mana', 'cocok', 'terap', 'jaring', 'vpn', 'simulasi', 'laku', 'guna', 'software', 'opnet', 'modeler', 'jenis', 'layan', 'guna', 'video', 'conference', 'voip', 'ftp', 'http', 'parameter', 'amat', 'delay', 'delay', 'variation', 'traffic', 'dropped', 'throughput', 'hasil', 'simulasi', 'lihat', 'teori', 'antri', 'wfq', 'paling', 'cocok', 'guna', 'vpn', 'milik', 'traffic', 'dropped', 'paling', 'kecil', 'besar', 'paket', 'throughput', 'paling', 'tinggi', 'besar', 'bit', 'delay', 'delay', 'variation', 'teori', 'antri', 'wfq', 'layan', 'video', 'conference', 'besar', 'teori', 'antri', 'pq', 'nilai', 'besar', 'layan', 'voip', 'teori', 'antri', 'wfq', 'milik', 'delay', 'delay', 'variation', 'besar', 'teori', 'antri', 'pq', 'nilai', 'besar', 'kata', 'kunci', 'vpn', 'fifo', 'pq', 'wfq', 'opnet', 'modeler']]</p>

Dokumen Q	<i>Preprocessing</i>	<i>Tokenizing</i>
<p>VPN karena memiliki traffic dropped yang paling kecil yaitu sebesar 847.97 paket/s dan throughput yang paling tinggi yaitu sebesar 6834565 bit/s. Delay dan delay variation teori antrian WFQ untuk layanan video conference yaitu sebesar 6.08 s dan 1.596 s, sedangkan pada teori antrian PQ nilainya sebesar 5.848 s dan 0.78 s. Untuk layanan VoIP, teori antrian WFQ memiliki delay dan delay variation sebesar 60.29881 s dan $2.71 \cdot 10^{-7}$ s sedangkan pada teori antrian PQ nilainya sebesar 60.298 s dan $2.71 \cdot 10^{-7}$ s.</p> <p>Kata kunci: VPN, FIFO, PQ, WFQ, Opnet Modeler 14.5</p>		

Tabel 4. 2 Dokumen pembanding sebelum dan sesudah *preprocessing*

Dokumen D	<i>Preprocessing</i>	<i>Tokenizing</i>
<p>Perkembangan teknologi telekomunikasi dimulai dari pemanfaatan teknologi Voice over Internet Protocol (VoIP) hingga teknologi satelit yang memungkinkan penggunaanya dapat</p>	<p>'kembang teknologi telekomunikasi mulai manfaat teknologi voice over internet protocol voip hingga teknologi satelit mungkin guna laku komunikasi kapan jaring broadband alami banyak</p>	<p>['kembang', 'teknologi', 'telekomunikasi', 'mulai', 'manfaat', 'teknologi', 'voice', 'over', 'internet', 'protocol', 'voip', 'hingga', 'teknologi', 'satelit', 'mungkin', 'guna', 'laku', 'komunikasi', 'kapan', 'jaring', 'broadband', 'alami',</p>

Dokumen D	<i>Preprocessing</i>	<i>Tokenizing</i>
<p>melakukan komunikasi dimana saja dan kapan saja. Pada jaringan broadband juga mengalami banyak perubahan dari masa ke masa demi menjaga kualitas pelayanan pada pelanggan operator seluler. Salah satunya adalah Universal Mobile Telecommunication Service (UMTS) yang merupakan generasi ketiga (3G) dari teknologi telekomunikasi bergerak. UMTS menyediakan berbagai layanan aplikasi seperti email, web browsing, video conferencing, dan voice. Untuk memberikan efisiensi dalam pengiriman data, diperlukan sebuah manajemen trafik, salah satunya menggunakan beberapa teori antrian yang berbeda. Dalam penelitian ini akan dilakukan pengamatan terhadap parameter layanan UMTS menggunakan beberapa teori antrian yang berbeda, yaitu First In First Out, Priority Queuing, dan Wighted</p>	<p>ubah masa masa jaga kualitas layan langgan operator seluler salah satu universal mobile telecommunication service umts rupa generasi tiga teknologi telekomunikasi gerak umts sedia bagai layan aplikasi email web browsing video conferencing voice beri efisiensi kirim data perlu buah manajemen trafik salah satu guna beberapa teori antri beda teliti laku amat parameter layananumts guna beberapa teori antri beda first in first out priority queuing wighted fair queuing simulasi laku guna software opnet modeler hasil simulasi lihat teori antrianwfg hasil rata delay kecil layan video conference besar ms delay variation kecil video conference dapat teori antri fifo besar ms layan voip delay delay variation kecil dapat teori antri fifo besar ms ms packet loss kecil dapat teori antri fifo besar lihat delay delay variation packet loss hasil teori antri fifo rupa teori antri</p>	<p>'banyak', 'ubah', 'masa', 'masa', 'jaga', 'kualitas', 'layan', 'langgan', 'operator', 'seluler', 'salah', 'satu', 'universal', 'mobile', 'telecommunication', 'service', 'umts', 'rupa', 'generasi', 'tiga', 'teknologi', 'telekomunikasi', 'gerak', 'umts', 'sedia', 'bagai', 'layan', 'aplikasi', 'email', 'web', 'browsing', 'video', 'conferencing', 'voice', 'beri', 'efisiensi', 'kirim', 'data', 'perlu', 'buah', 'manajemen', 'trafik', 'salah', 'satu', 'guna', 'beberapa', 'teori', 'antri', 'beda', 'teliti', 'laku', 'amat', 'parameter', 'layanumts', 'guna', 'beberapa', 'teori', 'antri', 'beda', 'first', 'in', 'first', 'out', 'priority', 'queuing', 'wighted', 'fair', 'queuing', 'simulasi', 'laku', 'guna', 'software', 'opnet', 'modeler', 'hasil', 'simulasi', 'lihat', 'teori', 'antrianwfg', 'hasil', 'rata', 'delay', 'kecil', 'layan', 'video', 'conference', 'besar', 'ms', 'delay', 'variation', 'kecil', 'video', 'conference', 'dapat', 'teori', 'antri', 'fifo', 'besar', 'ms', 'ms', 'packet', 'loss', 'kecil', 'dapat', 'teori', 'antri', 'fifo', 'besar', 'ms', 'ms', 'packet', 'loss', 'kecil', 'dapat', 'teori', 'antri',</p>

Dokumen D	<i>Preprocessing</i>	<i>Tokenizing</i>
<p>Fair Queuing. Simulasi dilakukan dengan menggunakan software OPNET Modeler 14.5. Dari Hasil Simulasi dapat dilihat bahwa pada teori antrian WFQ menghasilkan rata-rata delay terkecil pada layanan Video Conference yaitu sebesar 220.045 ms. Sedangkan untuk delay variation terkecil pada Video Conference terdapat pada teori antrian FIFO yaitu sebesar 0.0455 ms. Sedangkan untuk layanan VoIP, delay dan delay variation terkecil terdapat pada teori antrian FIFO yaitu sebesar 228.336 ms dan 147.871 ms. Dan untuk packet loss terkecil terdapat pada teori antrian FIFO sebesar 2.04028%. sehingga dilihat dari delay, delay variation, dan packet loss yang telah dihasilkan, maka teori antrian FIFO merupakan teori antrian yang sangat cocok untuk digunakan dalam jaringan UMTS.</p> <p>Kata kunci: UMTS, OPNET, Video Conference, VoIP, FTP, Teori Antrian</p>	<p>sangat cocok guna jaring umts kata kunci umts opnet video conference voip ftp teori antri'</p>	<p>'fifo', 'besar', 'lihat', 'delay', 'delay', 'variation', 'packet', 'loss', 'hasil', 'teori', 'antri', 'fifo', 'rupa', 'teori', 'antri', 'sangat', 'cocok', 'guna', 'jaring', 'umts', 'kata', 'kunci', 'umts', 'opnet', 'video', 'conference', 'voip', 'ftp', 'teori', 'antri']]</p>

Berdasarkan tabel 4.1 dan tabel 4.2 menunjukkan perbedaan teks dokumen sebelum dilakukan *preprocessing* dan sesudah dilakukan *preprocessing*. Pada penelitian ini *preprocessing* yang dilakukan adalah *case folding* yaitu merubah semua teks menjadi huruf kecil. Kemudian proses *removenonalphabet* yang digunakan untuk menghilangkan angka dan symbol yang terdapat pada teks. Lalu proses *stemming* untuk mencari kata dasar dilakukan dengan menghilangkan semua imbuhan dari kata, baik itu awalan, sisipan, maupun akhiran. Selanjutnya stopword removal yaitu menghilangkan kata-kata yang tidak memiliki arti seperti dan, itu, juga, dan sebagainya. Penelitian ini menggunakan kamus sastra untuk menghilangkan stopwords. Proses *preprocessing* terakhir adalah proses *tokenizing* untuk memotong kata-suatu kata menjadi token-token, modul yang digunakan untuk *tokenizing* pada penelitian ini menggunakan modul *spacy* khusus berbahasa Indonesia.

4.3 Penerapan Metode

Setelah dilakukan tahapan *preprocessing* dokumen testing dan dataset selanjutnya adalah deteksi kemiripan similarity dari dokumen testing dengan masing-masing dokumen yang ada dalam dataset. Pada penelitian ini penulis melakukan deteksi kemiripan menggunakan algoritma *Cosine Similarity* dan menggunakan metode *Word2vec*. Penulis juga melakukan pembobotan menggunakan metode TF-IDF guna mengetahui perbedaan hasil yang didapatkan dari masing masing pembobotan.

4.3.1 Pembobotan TF-IDF dan *Cosine Similarity*

Dokumen uji dan dokumen pembanding berupa dataset abstrak yang sudah dilakukan *preprocessing* selanjutnya dilakukan pembobotan kata/*term* untuk menghitung frekuensi kemunculan setiap kata dokumen testing dalam masing-masing dokumen yang ada dalam dataset. Proses TF-IDF menggunakan bantuan model dari *sklearn.feature_extraction.text*. Setelah melakukan proses TF-IDF menggunakan TF-IDF *Vectorizer*, dokumen uji Q dan dokumen pembanding D terdapat 128 *term* yang unik. dalam percobaan yang dilakukan *term* dari dokumen uji Q dan dokumen pembanding D diurutkan berdasarkan abjad. Berikut merupakan *term* dari kedua dokumen:

['ada', 'akses', 'alami', 'aman', 'amat', 'antri', 'antrianwfq', 'aplikasi', 'bagai', 'banyak', 'beberapa', 'beda', 'beri', 'besar', 'bit', 'broadband', 'browsing', 'buah', 'cocok', 'conference', 'conferencing', 'dapat', 'data', 'delay', 'dropped', 'efisiensi', 'email', 'enkripsi', 'fair', 'fifo', 'first', 'ftp', 'generasi', 'gerak', 'guna', 'hasil', 'hingga', 'http', 'imbang', 'in', 'internet', 'jaga', 'jaring', 'jenis', 'kapan', 'kata', 'kecil', 'keman', 'kembang', 'kirim', 'komunikasi', 'kualitas', 'kunci', 'laku', 'langgan', 'layan', 'layananumts', 'lihat', 'loss', 'mana', 'manajemen', 'manfaat', 'masa', 'masing', 'milik', 'mobile', 'modeler', 'ms', 'mulai', 'mungkin', 'nilai', 'operator', 'opnet', 'orang', 'out', 'over', 'packet', 'paket', 'paling', 'parameter', 'pasti', 'perlu', 'pq', 'priority', 'protocol', 'queuing', 'rata', 'rupa', 'salah', 'sama', 'sangat', 'satelit', 'satu', 'sedia', 'seluler', 'service', 'simulasi', 'sistem', 'skenario', 'software', 'tahu', 'tawar', 'teknologi', 'telecommunication', 'telekomunikasi', 'teliti', 'teori', 'terap', 'throughput', 'tiga', 'tinggi', 'tingkat', 'traffic', 'trafik', 'tujuan', 'ubah', 'umts', 'universal', 'variation', 'video', 'voice', 'voip', 'vpn', 'wan', 'web', 'weighted', 'wfq', 'wighted']

Berdasarkan *term* dari dokumen uji (Q) dan dokumen pembanding (D) selanjutnya dihitung banyaknya kemunculan kata dari masing-masing dokumen (TF) kemudian hasil dari banyaknya kemunculan dihitung nilai IDF dengan rumus $\text{Log}(N/DF)$. Hasil tersebut kemudian dikalikan dan mendapatkan hasil nilai TF-IDF dari masing masing kata. Hasil dari pembobotan TF-IDF sebagai berikut:

[[0.05368445	0.05368445	0	0.05368445	0.07639385	0.30557542
0	0.03819693	0	0	0.03819693	0.07639385
0.03819693	0.22918156	0.05368445	0	0	0.03819693
0.07639385	0.07639385	0	0	0.15278771	0.22918156
0.1073689	0.03819693	0	0.05368445	0.03819693	0.07639385
0.07639385	0.03819693	0	0	0.26737849	0.03819693
.....
0.32210669	0.05368445	0	0.05368445	0.26842224	0]
[0	0	0.05647734	0	0.04018409	0.32147275
0.05647734	0.04018409	0.05647734	0.05647734	0.08036819	0.08036819
0.04018409	0.16073638	0	0.05647734	0.05647734	0.04018409
0.04018409	0.12055228	0.05647734	0.16943203	0.04018409	0.24110456
0	0.04018409	0.05647734	0	0.04018409	0.16073638
0.08036819	0.04018409	0.05647734	0.05647734	0.20092047	0.12055228
.....
0.16073638	0.11295468	0.12055228	0	0	0.05647734]

Hasil pembobotan TF-IDF dari dokumen uji (Q) dan dokumen pembanding (D) kemudian dihitung nilai similarity menggunakan algoritma *Cosine Similarity*

dan didapatkan nilai similarity sebesar 0.56484648. Tabel 4.3 merupakan hasil dari dokumen uji (Q) dengan seluruh dokumen pembanding yaitu sebanyak 116 abstrak:

Tabel 4. 3 Nilai Kemiripan Dokumen Uji dengan Dokumen Pembanding dengan TF-IDF

Dokumen Pembanding	Nilai Kemiripan	Dokumen pembanding	Nilai kemiripan
Abstrak 1	0.05505323	Abstrak 59	0.13406274
Abstrak 2	0.02744747	Abstrak 60	0.13337578
Abstrak 3	0.07559086	Abstrak 61	0.11494524
Abstrak 4	0.06745039	Abstrak 62	0.06635504
Abstrak 5	0.06745039	Abstrak 63	0.09110803
Abstrak 6	0.05591735	Abstrak 64	0.08209712
Abstrak 7	0.06145965	Abstrak 65	0.06546887
Abstrak 8	0.10752431	Abstrak 66	0.03632408
Abstrak 9	0.10437786	Abstrak 67	0.05810611
Abstrak 10	0.08962029	Abstrak 68	0.0822703
Abstrak 11	0.05817915	Abstrak 69	0.28144386
Abstrak 12	0.03847128	Abstrak 70	0.06497782
Abstrak 13	0.07776414	Abstrak 71	0.04663605
Abstrak 14	0.0645286	Abstrak 72	0.21707739
Abstrak 15	0.07490983	Abstrak 73	0.02785215
Abstrak 16	0.06272732	Abstrak 74	0.12002718
Abstrak 17	0.05733868	Abstrak 75	0.04182541
Abstrak 18	0.19104265	Abstrak 76	0.06432293
Abstrak 19	0.11785182	Abstrak 77	0.11118873
Abstrak 20	0.0593272	Abstrak 78	0.02794283
Abstrak 21	0.05374373	Abstrak 79	0.0808007
Abstrak 22	0.10761659	Abstrak 80	0.03192188
Abstrak 23	0.10968077	Abstrak 81	0.15431798
Abstrak 24	0.067611	Abstrak 82	0.10741905
Abstrak 25	0.09168304	Abstrak 83	0.03117623
Abstrak 26	0.13651776	Abstrak 84	0.06973119
Abstrak 27	0.11836027	Abstrak 85	0.04968132
Abstrak 28	0.09500405	Abstrak 86	0.11031348
Abstrak 29	0.12659687	Abstrak 87	0.26208497
Abstrak 30	0.11069516	Abstrak 88	0.04662862
Abstrak 31	0.2118826	Abstrak 89	0.2595726