BAB V
KESIMPULAN DAN SARAN

A. Kesimpulan

Kesimpulan yang didapatkan setelah melakukan serangkaian penelitian dan pengujian adalah:

1. Hasil *clustering* dengan menggunakan pembobotan TF-IDF yang diintegrasikan ke *k-Means clustering* memiliki tingkat akurasi yang cukup tinggi (di atas 50%).
2. Rumus perhitungan TF-IDF standar (Robertson, 2004) tidak cocok diintegrasikan pada *k-Means clustering*, sehingga harus diberikan rumus normalisasi TF-IDF.
3. Penentuan titik awal *centroid* tiap *cluster* berpengaruh terhadap hasil *clustering*.
4. Nilai *threshold* tidak memberikan akibat yang signifikan pada hasil *clustering* kecuali dipasang nilai terlalu tinggi (di atas 75%)

B. Saran

Saran yang dapat diberikan bagi peneliti dalam bidang sejenis, yang ingin melanjutkan atau mengembangkan penelitian yang sudah diselesaikan ini adalah:

1. Mengembangkan aplikasi ini untuk tipe dokumen lain seperti .doc/.pdf/.html.

2. Meningkatkan waktu kinerja sistem yang dirasakan masih memakan waktu cukup banyak.
3. Menemukan suatu algoritma untuk mencari titik centroid awal yang terbaik untuk korpus dokumen tertentu.

Daftar Pustaka

Abbasi, Rabeeh and Steffen Staab, 2009, *RichVSM: enRiched Vector Space Model for Folksonomies*, Inform*a*tion Systems and Semantic Web Research Group, Institute for Computer Science, University of Koblenz – Landau, Koblenz.

Abual-Rub, Mohammed Said, Rosni Abdullah dan Nur'aini Abdul Rashid, 2007, *A Modified Vector Space Model for Protein Retrieval*, school of computer sciences, universiti Sains Malaysia, Penang, Malaysia.

Arthur, David and Sergei Vassilvitskii, 2006, *How Slow is the k-Means Method*, Stanford University, Stanford, CA.

Ben-David, Shai, David Pal, and Hans Ulrich Simon, 2009, *Stability of k-Means clustering*, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.

Bhatia, Sanjiv K., 2004, *Adaptive K-Means clustering*, Department of Mathematics & Computer Science, University of Missouri – St. Louis.

Buckland, Michael and Fredric Gey, 1994, *The Relation between Recall and Precision,* Journal of the American Society for Information Science (1986-1998); Jan 1994; 45, 1; ABI/IN*FORM* Global pg. 12.

Chen, Ja-Shen, Russel KH Ching, Yi-Shen Lin, 2004, *An Extended Study Of The K-Means Algorithm For Data clustering And Its Applications*, Journal of the Operational Research Society (2004) 55: 976-987.

Chen, Yiheng, Bing Qin, Ting Liu, Yuanchao Liu, Sheng Li, 2010, *The Comparison of SOM and k-Means for Text clustering*, Computer and In*form*ation Science, Vol. 3.

Cummins, Ronan and Colm O'Riordan, *Evolving Local and Global Weighting Schemes in Information Retrieval*, 2006, Springer Science+Business Media, LLC 2006

Ding, Jiarui, Jinhong Shi, Fang-Xiang Wu, 2009, *Quality Assesment of Tandem Mess Spectra By Using A Weighted k-Means*, Clin Proteom (2009) 5:15-22; DOI10.1007/s12014-009-9025-4.

Dominich S., 2008, *The Modern Algebra of Information Retrieval*, Springer. ISBN 3540776583.

Douglas Steinley, 2006, *k-Means clustering: A Half Century Analysis*, British Journal of Mathematical & Statistical Psychology, Academic Research Library.

Elkan, Charles; 2005, *Deriving TF-IDF as a Fisher Kernel*, Department of Computer Science and Engineering, University of California, San Diego.

Frahling, Gereon and Sohler, 2005, *A Fast K-Means Implementation Using Coresets*, Department of Computer Science, University of Paderborn, Paderborn

Gehanno, Jean-Francois, Laetitia Rollin, Tony Le Jean, 2009, *Precision and Recall of Search Strategies for Identifying Studies on Return-to-Work in Medline*, J Occup Rehabil (2009) 19:223–230; DOI 10.1007/s10926-009-9177-0.

Gong, Zhiguo and Qian Liu, 2009, *Improving Keyword Based Web Image Search with Visual Feature Distribution and Term Expansion*, Knowl Inf Syst (2009) 20:63–79; DOI 10.1007/s10115-008-0151-5.

Intan, Rolly dan Andrew Defeng, 2006, *Subject Based Search Engine Menggunakan TF-IDF dan Jaccard's Coefficient*, Universitas Kristen Petra, Surabaya.

Jin, Ruoming, Anjan Goswani, Gagan Agrawal, 2006, *Fast and Exact Out-of-Core And Distributed K-Means clustering*, Knowledge In*form*ation System (2006) 10(1): 17–40; DOI 10.1007/s10115-005-0210-0.

Kanungo, Tapas, David M. Mount, Nathan S. Nethanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, 2002, *An Efficient k-Means clustering Algorithm: Analysis and Implementation*, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Kathuria, Ashish, 2010, *Classifying the User Intent of Web Queries Using K-Means clustering*, Emerald Group Publishing Limited.

Khatatneh, Khalaf, M. Wedyan, Mohamed Alham, Basem Alrifai, 2005, *Using New Data Structure to Implement Documents Vectors in Vector Space Model in Information Retrieval System*, Prince Abdu Allah Bin Ghazi for IT, Al-Balqa Applied University Salt, Jordan

Kogan, Jacob, Marc Teboulle, Charles Nicholas, 2006, *Data Driven Similarity Measures for k-Means Like clustering Algorithm*, In*form*ation Retrieval, 8, 331–349, 2005, © 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

Le Wang, Yan Jia, dan Weihong Han, 2007, *Instant Message clustering Based on Extended Vector Space Model*, Computer School, National University of Defense Technology, Changsha, China.

Likas, Aristidis, Nikos Vlassis, dan Jacob J. Verbeek, 2002, *The Global K-Means clustering Algorithm*, Department of Computer Science, University of Ioannina, Ioannina, Greece.

Lloyd., S. P. (1982). *Least squares quantization in PCM*. IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

Mahdavi, Mehrdad and Hassan Abolhassani; 2008, *Harmony k-Means Algorithm for Document clustering*, Springer Science+Business Media, LLC 2008.

Manning, Christoper D., Prabhakar Raghavan, Hinrich Schutze; 2008, *Introduction to Information Retrieval*, © Cambridge University Press, ISBN: 978-0-521-86571-5.

McJunkin, Monica Cahill, 1995, *Precision and Recall in Title Keyword Searches,* In*form*ation Technology and Libraries; Sep 1995; 14, 3; ABI/IN*FORM* Global pg. 161.

Ming, Mark and Tso Chiang, 2010, *Intelligent Choice of the Number of Cluster in k-Means clustering: An Experimental Study with Different Cluster Spreads*, Journal of Classification 27:3-40.

Modha, Dharmendra S. and W. Scott Spangler, 2003, *Feature Weighting in k-Means clustering*, Kluwer Academic Publishers.

Moulin, Christophe, Cecile Barat, and Christophe Ducottet, 2010, *Fusion of TF.IDF weighted bag of visual features for image classification*, CBMI 2010.

Pham, D. T., S. S. Dimov, and C. D. Nguyen, 2004, *An Incremental k-Means Algorithm*, Proceedings of the Institution of Mechanical Engineers; Jul 2004; 218, 7; ProQuest Science Journals.

Pham, D. T., S. S. Dimov, and C. D. Nguyen, 2004, *A Two Phase k-Means Algorithm For Large Datasets*, 2004, Proceedings of the Institution of Mechanical Engineers; Proquest Science Journals.

Pham, D. T., S. S. Dimov, and C. D. Nguyen, 2005, *Selection of K in k-Means clustering*, Proceedings of the Institution of Mechanical Engineers; Proquest Science Journals

Price, Simon, Sebastian Spiegler, Peter A. Flach; 2010, *SubSift: a Novel Application of the Vector Space Model to Support the Academic Research Process*, Institute for Learning and Research Technology, University of Bristol, Bristol.

Ramos, Juan, 2010, *Using TF-IDF to Determine Word Relevance in Document Queries*, Department of Computer Science, Rutgers University, Piscataway.

Recupero, Diego R., 2007, *A New Unsupervised Method for Document clustering By Using Wordnet Lexical And Conceptual Relations.*, Inf Retrieval (2007) 10:563–579; DOI 10.1007/s10791-007-9035-7; © Springer Science+Business Media, LLC 2007.

Rezgui, Yacine, 2007, *Text Based Domain Ontology Using TF-IDF and Metric Cluster Techniques*, The Knowledge Engineering Review, Vol. 22:4, 379–403. 2007, Cambridge University Press, doi:10.1017/S0269888907001130.

Robertson, Stephen, 2004, *Understanding Inverse Document Frequency: On Theoritical Arguments for IDF*, Journal of Documentation; 2004; 60, 5; ABI/IN*FORM* Global.

Savoy, Jacques, 2007, *Searching Strategies for the Bulgarian Language*, Springer Science + business Media, LLC 2007.

Schlieder, Torsten and Holger Meuss, 2002, *Querying and Ranking XML Desktop*, Journal of the American Society for In*form*ation Science and Technology; Apr 2002; 53, 6; ABI/IN*FORM* Global pg. 489.

Setodji, Messan, and R Dennis Cook, 2004, *K-Means Inverse Regression*, Technometrics; Nov 2004; 46, 4; ABI/IN*FORM* Global pg. 421.

Steinley, Douglas and Laurence Hubert, 2008, *Order-Constrained Solutions In k-Means clustering: Even Better Than Being Globally*

Optimal, PSYCHOMETRIKA – Vol. 73, No. 4, 647-664; December 2008; DOI: 10.1007/s11336-008-9058-z.

Steinley, Douglas and Michael J. Brusco, 2007, *Initializing K-Means Batch clustering: A Critical Evaluation Of Several Techniques*, Journal of Classification 24:99-121 (2007); DOI: 10.1007/s00357-007-0003-0.

Su, Louise T., 1994, *The Relevance of Recall and Precision in User Evaluation*, Journal of the American Society for In*form*ation Science (1986-1998); Apr 1994; 45, 3; ABI/IN*FORM* Global pg. 207.

Tarpey, Thaddeus, 2007, *A Parametric k-Means Algorithm*, © Springer Verlag 2007, Computational Statistic 22: 71-89.

Takano, Kosuke, Xing Chen, Keisuke Masuda, 2009, *A Framework for a Feedback Process to Analyze and Personalize A Document Vector Space in a Feature Extraction Model*, Inf Technol Manag (2009) 10:151–176; DOI 10.1007/s10799-009-0055-4.

Umran, Munzir and Taufik F. Abidin, 2009, *Pengelompokan Dokumen Menggunakan K-Means dan singular value decomposition: Studi Kasus menggunakan Data Blog*, Data Mining and In*form*ation Retrieval Research Group, Universitas Syiah Kuala, Banda Aceh

Walters, William H., 2009, *Google Scholar Search Performance: Comparative Recall and Precision*, Portal: Libraries and the Academy, Vol. 9, No. 1 (2009), pp. 5–24. Copyright © 2009 by The Johns Hopkins University Press, Baltimore, MD 21218.

Wang, Ye-Yi dan Alex Acero, 2007, *Maximum Entropy Model Parameterization with TF*IDF Weighted Vector Space Model*, Microsoft Research.

Wang, Zheng, Qing Wang, Ding-Wei Wang, 2009, *Bayesian Network Based Business Information Retrieval Model*, Knowl Inf Syst (2009) 20:63–79; DOI 10.1007/s10115-008-0151-5.

Zhang, Tong and Frank J. Oles, 2001, *Text Categorization Based on Regularized Linear Classification Methods,* In*form*ation Retrieval; Apr 2001; 4, 1; ABI/IN*FORM* Global pg. 5.

SKPL

SPESIFIKASI KEBUTUHAN PERANGKAT LUNAK

APLIKASI CLUSTERING DOKUMEN TEKS DENGAN MENGGUNAKAN METODE K-MEANS CLUSTERING DAN PEMBOBOTAN TF-IDF (ClustKT)

Dipersiapkan oleh:

Deddy Wijaya Suliantoro / 105301466

Program Studi Magister Teknik Informatika

Program Pasca Sarjana

Universitas Atma Jaya Yogyakarta

DAFTAR PERUBAHAN

Revisi	Deskripsi
A	
B	
C	
D	
E	
F	

INDEX TGL	-	A	B	C	D	E	F	G
Ditulis oleh								
Diperiksa oleh								
Disetujui oleh								

DAFTAR HALAMAN PERUBAHAN

Halaman	Revisi	Halaman	Revisi

DAFTAR ISI

DAFTAR GAMBAR

1. Pendahuluan

1.1 Tujuan

Tujuan dari dokumen spesifikasi kebutuhan perangkat lunak ini merupakan dokumen spesifikasi kebutuhan perangkat lunak ClustKT (Aplikasi clustering dokumen teks dengan menggunakan metode k-Means dan pembobotan TF-IDF) untuk mendefinisikan kebutuhan perangkat lunak yang meliputi antarmuka eksternal (antarmuka antara sistem dengan perangkat lunak dan perangkat keras, dan pengguna), performansi (kemampuan perangkat lunak dari segi kecepatan, tempat penyimpanan yang dibutuhkan, serta keakuratan), dan atribut tambahan yang dimiliki sistem, serta mendefinisikan fungsi perangkat lunak. SKPL-SIGJ ini juga mendefinisikan batasan perancangan perangkat lunak, karakteristik program, serta asumsi dan ketergantungan perangkat lunak ini.

1.2 Lingkup Masalah

Perangkat lunak ClustKT dikembangkan dengan tujuan untuk :

1. Menangani pengelolaan data korpus
2. Menangani pengelolaan data stopwords
3. Menangani proses clustering dokumen teks

1.3 Definisi dan Akronim

Daftar definisi akronim dan singkatan :

Keyword/Phrase	Definisi
SKPL	Dokumen SKPL ini berisi tentang spesifikasi kebutuhan dari pengembangan perangkat lunak.
ClustKT	Sistem Clustering Dokumen Teks Menggunakan Metode k-Means dan

	pembobotan TF-IDF
Clustering	Proses pengelompokan data ke dalam beberapa cluster (kelompok) tertentu.
korpus	Data dokumen teks yang berada dalam database sistem yang digunakan untuk melakukan proses clustering
stopwords	Daftar kata-kata (token) yang tidak diikutkan dalam perhitungan di proses clustering

1.4 Deskripsi Umum

Secara umum dokumen SKPL ini terbagi 4 bagian utama. Bagian pertama berisi penjelasan mengenai dokumen SKPL tersebut yang mencakup tujuan pembuatan SKPL, ruang lingkup masalah dalam pengembangan perangkat lunak, definisi, referensi, dan deskripsi umum tentang dokumen SKPL ini.

Bagian kedua berisi penjelasan umum tentang perangkat lunak ClustKT yang akan dikembangkan mencakup perspektif produk yang, fungsi produk perangkat lunak, karakterisitik pengguna, batasan dalam penggunaan perangkat lunak dan asumsi yang terpakai dalam pengembangan perangat lunak ClustKT.

Bagian ketiga berisi penjelasan secara lebih rinci tentang kebutuhan perangkat lunak ClustKT yang akan dikembangkan. Pada bagian terakhir atau bagian keempat berisi tentang spesifikasi kebutuhan data.

2. Deskripsi Kebutuhan

2.1 Perspektif Produk

ClustKT merupakan perangkat lunak yang digunakan untuk melakukan proses pengelompokan dokumen-dokumen teks yang tergabung dalam korpus ke dalam sejumlah cluster yang sudah ditentukan terlebih dahulu.

Perangkat lunak ClustKT berjalan pada platform Windows yang memiliki .NET Framework versi 3.5 ke atas. Bahasa pemrograman yang digunakan dalam pembangunan ClustKT adalah Visual Basic .NET dengan menggunakan tools Visual Studio 2008.

Pengguna akan berinteraksi dengan sistem melalui antarmuka GUI (Graphical User Interface).

2.1.1 Antarmuka Pemakai

Pengguna berinteraksi dengan antarmuka yang ditampilkan dalam bentuk form-form yang merupakan aplikasi desktop.

2.1.2 Antarmuka Perangkat Keras

Piranti perangkat keras yang dibutuhkan oleh perangkat lunak ClustKT adalah sebagai berikut:

1. PC (Personal Computer)
2. Mouse
3. Keyboard

2.1.3 Antarmuka Perangkat Lunak

Perangkat lunak yang dibutuhkan dalam mengoperasikan perangkat lunak ClustKT adalah:

1. Sistem Operasi Windows sebagai sistem operasi komputer
2. .NET Framework 4.0 sebagai pustaka pemrograman yang dibutuhkan dalam menjalankan perangkat lunak ClustKT

3. Microsoft SQL Server 2005 sebagai DBMS dari perangkat lunak ClustKT.

2.2 Fungsi Produk

Fungsi produk perangkat lunak ClustKT dibagi menjadi 3 bagian besar:

1. Fungsi kelola stopwords **(SKPL-ClustKT-01)**
   Merupakan fungsi yang digunakan untuk mengelola kata-kata dalam stopwords.
   a. Fungsi penambahan kata dalam stopwords **(SKPL-ClustKT-01-01)**
      Merupakan fungsi untuk menambah kata-kata ke dalam daftar stopwords.
   b. Fungsi penghapusan kata dalam stopwords **(SKPL-ClustKT-01-02)**
      Merupakan fungsi untuk menghapus kata tertentu dalam daftar stopwords.
2. Fungsi kelola korpus **(SKPL-ClustKT-02)**
   Merupakan fungsi yang digunakan dalam mengelola dokumen-dokumen dalam korpus.
   a. Fungsi penambahan dokumen **(SKPL-ClustKT-02-01)**
      Fungsi ini digunakan dalam menambahkan dokumen ke dalam korpus.
   b. Fungsi penghapusan dokumen **(SKPL-ClustKT-02-02)**
      Fungsi ini digunakan dalam menghapus dokumen tertentu maupun semua dokumen dari dalam korpus.
   c. Fungsi pengecekan path dokumen **(SKPL-ClustKT-02-03)**
      Fungsi ini digunakan dalam mengecek path yang tercatat apakah masih valid atau tidak.
3. Fungsi melihat isi dokumen **(SKPL-ClustKT-03)**
   Fungsi ini digunakan untuk melihat isi dokumen teks pada path tertentu.

4. Fungsi pembangunan indeks dokumen **(SKPL-ClustKT-04)**
   Fungsi ini digunakan untuk membangun indeks dokumen dan melakukan perhitungan pembobotan TF-IDF.
5. Fungsi clustering dokumen **(SKPL-ClustKT-05)**
   Merupakan fungsi yang digunakan untuk melakukan proses clustering terhadap dokumen-dokumen dalam korpus.

2.3 Karakteristik Pengguna

Karakteristik pengguna yang menggunakan perangkat lunak ClustKT yang dibangun yaitu :

a. Mampu mengoperasikan komputer pada level dasar (Menyalakan, mematikan, menggunakan aplikasi)
b. Memahami sistem komputer windows
c. Memahami konsep clustering
d. Mengerti proses pengelolaan data

2.4 Batasan – batasan

Batasan yang ditetapkan dalam pengembangan perangkat lunak ClustKT ini adalah:

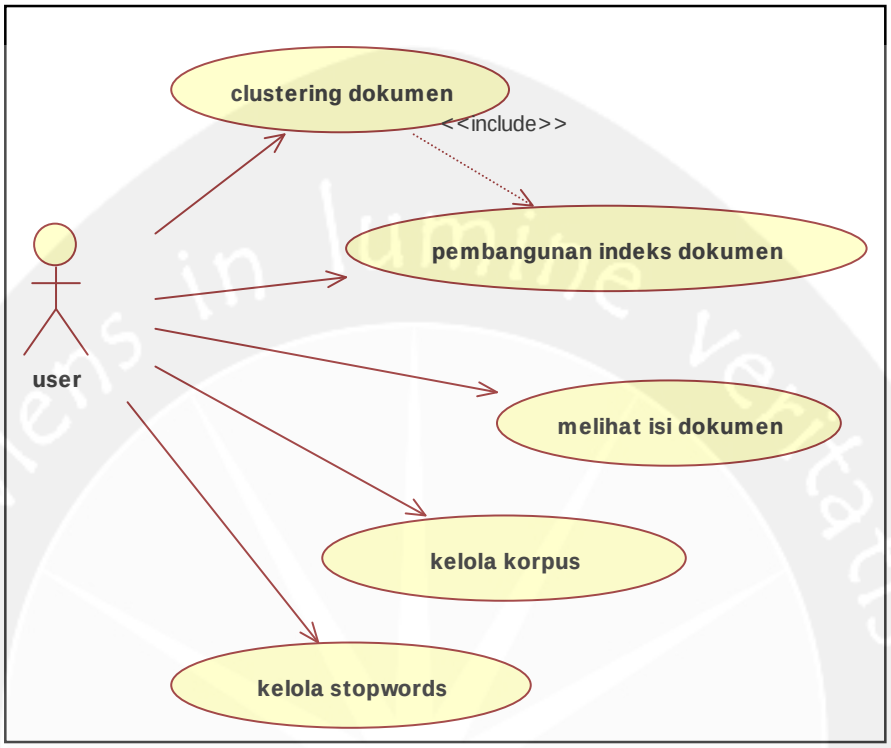
a. Tujuan perangkat lunak ClustKT ini adalah sebagai instrumen penelitian untuk meneliti akurasi metode k-Means yang dipadukan dengan pembobotan TF-IDF dan membantu user dalam proses clustering dokumen teks
b. Keterbatasan perangkat keras akan ditentukan kemudian setelah aplikasi ini berjalan (sesuai dengan kebutuhan)

2.5 Asumsi dan Ketergantungan

Sistem ini dapat dijalankan ada perangkat desktop yang menggunakan sistem operasi versi windows XP atau Vista atau Seven yang dilengkapi dengan .NET Framework 3.5 atau lebih tinggi.

3. Kebutuhan Fungsionalitas Perangkat Lunak

3.1 Use Case Diagram



Gambar 1 Use Case Diagram

3.2 Use Case Spesification

3.2.1 Use Case Specification : Kelola Stopwords

Brief Description	Use case ini digunakan oleh aktor untuk melakukan pengelolaan kata-kata dalam stopwords list.
Primary Actor	User
Supporting Actor	-
Basic Flow	1. Use case ini dimulai ketika aktor memilih untuk mengelola data stopwords. 2. Sistem memberikan pilihan untuk menambah kata atau menghapus kata. 3. Aktor memilih untuk menambah kata ke dalam stopwords list A-1 Aktor memilih untuk menghapus kata dari stopwords list.

	4. Aktor menginputkan kata baru 5. Sistem mengecek kata yang telah diinputkan. E-1 Kata yang diinputkan sudah ada E-2 Kata yang diinputkan mengandung spasi 6. Sistem menyimpan kata baru ke dalam database. 7. Use case selesai
Alternative Flow	A-1 Aktor memilih untuk menghapus kata dari stopwords list. 1. Aktor memilih kata yang ingin dihapus dan menekan tombol hapus. E-3 Aktor belum memilih kata dan menekan tombol hapus. 2. Sistem menghapus kata dari database dan memberi informasi ke aktor. 3. Berlanjut ke Basic Flow langkah 7
Error Flow	E-1 Kata yang diinputkan sudah ada 1. Sistem memberi peringatan bahwa kata yang dimasukkan sudah ada dalam stopwords list. 2. Kembali ke Basic Flow langkah yang ke-3 E-2 Kata yang diinputkan mengandung spasi 1. Sistem memberi peringatan bahwa kata yang dimasukkan mengandung spasi. 2. Kembali ke Basic Flow langkah yang ke-3 E-3 Aktor belum memilih kata dan menekan tombol hapus. 1. Sistem memberi peringatan bahwa Aktor belum memilih kata yang mau dihapus. 2. Kembali ke Basic Flow langkah yang ke-3
Pre-Conditions	-
Post-Conditions	Data kata dalam stopwords list bertambah atau berkurang.

3.2.2 Use Case Spesification : Kelola Korpus

Brief Description	Use case ini digunakan oleh aktor untuk melakukan pengelolaan dokumen dalam korpus
Primary Actor	User
Supporting Actor	-
Basic Flow	1. Use case ini dimulai ketika aktor memilih untuk mengelola data korpus. 2. Aktor memilih untuk menambah dokumen ke dalam korpus

	A-1 Aktor memilih untuk menghapus dokumen tertentu dari korpus A-2 Aktor memilih untuk menghapus seluruh dokumen dari korpus. A-3 Aktor memilih untuk mengecek validitas tiap dokumen dalam korpus. 3. Sistem membuka form input dokumen baru 4. Aktor memilih dokumen yang ingin ditambahkan ke dalam list yang disediakan sistem. 5. Aktor menyetujui memasukkan dokumen dalam list ke korpus A-4 Aktor membatalkan proses input dokumen. 6. Sistem mengecek dokumen-dokumen yang telah diinputkan. E-1 Dokumen yang diinputkan sudah terdaftar dalam database 7. Sistem menyimpan dokumen baru ke dalam database. 8. Use case selesai
Alternative Flow	A-1 Aktor memilih untuk menghapus dokumen tertentu dari korpus 1. Aktor memilih dokumen yang ingin dihapus dan menekan tombol hapus. E-2 Aktor belum memilih dokumen dan menekan tombol hapus. 2. Sistem menghapus dokumen dari database dan memberi informasi ke aktor. 3. Berlanjut ke Basic Flow langkah 8 A-2 Aktor memilih untuk menghapus seluruh dokumen dari korpus. 1. Aktor memilih untuk menghapus semua dokumen dalam korpus 2. Sistem menghapus semua dokumen dari database dan memberi informasi ke aktor. 3. Berlanjut ke Basic Flow langkah 8 A-3 Aktor memilih untuk mengecek validitas tiap dokumen dalam korpus. 1. Aktor memilih untuk mengecek validitas tiap dokumen dalam korpus 2. Sistem mengecek validitas masing-masing dokumen dan jika ada yang tidak valid, sistem akan memberi informasi ke aktor

	dan kemudian menghapus dokumen tersebut. 3. Berlanjut ke Basic Flow langkah 8
Error Flow	E-1 Dokumen yang diinputkan sudah terdaftar dalam database 1. Sistem memberi peringatan bahwa ada dokumen yang dimasukkan sudah ada dalam korpus. 2. Kembali ke Basic Flow langkah yang ke-2 E-2 Aktor belum memilih dokumen dan menekan tombol hapus 1. Sistem memberi peringatan bahwa Aktor belum memilih dokumen yang mau dihapus. 2. Kembali ke Basic Flow langkah yang ke-3
Pre-Conditions	-
Post-Conditions	Data dokumen dalam korpus bertambah atau berkurang.

3.2.3 Use Case Spesification : Melihat Isi Dokumen

Brief Description	Use case ini digunakan oleh aktor untuk melihat isi dokumen dari path tertentu.
Primary Actor	User
Supporting Actor	-
Basic Flow	1. Use case ini dimulai ketika aktor membuka aplikasi. A-1 Use case dimulai setelah Use case clustering selesai 2. Aktor memilih melakukan double-click terhadap salah satu dokumen dalam daftar dokumen. 3. Sistem membuka form untuk membuka isi dokumen 4. Aktor menutup form isi dokumen A-2 Aktor memilih dokumen lain dari daftar 5. Use case selesai
Alternative Flow	A-1 Use case dimulai setelah Use case clustering selesai 1. Berlanjut ke Basic Flow langkah 2 A-2 Aktor memilih dokumen lain dari daftar 1. Sistem mengubah isi dari form isi dokumen yang terbuka dengan isi dokumen yang baru dipilih 2. Berlanjut ke Basic Flow langkah 4

Error Flow	-
Pre-Conditions	Ada dokumen dalam korpus
Post-Conditions	-

3.2.4 **Use Case Spesification :** Pembangunan Indeks Dokumen

Brief Description	Use case ini digunakan oleh aktor untuk membangun indeks dari dokumen yang diperlukan dalam use case clustering.
Primary Actor	User
Supporting Actor	-
Basic Flow	1. Use case ini dimulai ketika aktor memilih untuk melakukan pengindeksan dokumen 2. Sistem memberi konfirmasi untuk melakukan proses pengindeksan E-1 Belum ada dokumen dalam korpus 3. Sistem melakukan pengindeksan dokumen 4. Sistem memberi informasi bahwa proses pengindeksan selesai. 5. Use case selesai
Alternative Flow	-
Error Flow	E-1 Belum ada dokumen dalam korpus 1. Sistem menginformasikan bahwa belum ada dokumen dalam korpus 2. Berlanjut ke Basic Flow langkah 5
Pre-Conditions	Ada dokumen dalam korpus
Post-Conditions	indeks dokumen terbangun

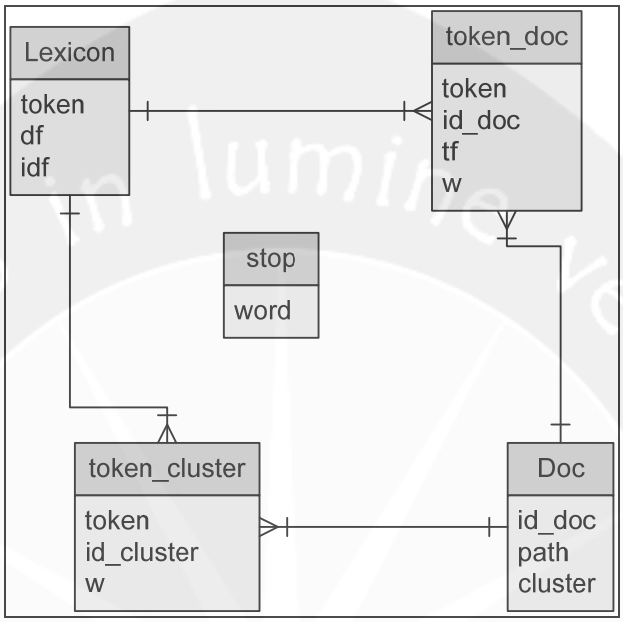
3.2.5 Use Case Spesification : Clustering

Brief Description	Use case ini digunakan oleh aktor untuk melakukan clustering dari sejumlah dokumen dalam korpus yang sudah dibangun indeksnya.
Primary Actor	User
Supporting Actor	-
Basic Flow	1. Use case ini dimulai ketika aktor memilih untuk melakukan clustering dokumen 2. Sistem menunjukkan form clustering 3. Aktor memasukkan jumlah cluster yang diinginkan dan nilai threshold yang diinginkan. 4. Sistem melakukan proses clustering dokumen E-1 Jumlah dokumen tidak mencukupi untuk dilakukan clustering

	E-2 Nilai cluster dan threshold yang dimasukkan tidak valid E-3 Belum ada indeks dari dokumen. 5. Sistem memberi informasi bahwa proses clustering selesai. 6. Sistem menunjukkan hasil clustering. 7. Use case selesai
Alternative Flow	-
Error Flow	E-1 Jumlah dokumen tidak mencukupi untuk dilakukan clustering 1. Sistem menginformasikan bahwa jumlah dokumen dalam korpus tidak mencukupi untuk dilakukan proses clustering 2. Berlanjut ke basic flow langkah 3 E-2 Nilai cluster dan threshold yang dimasukkan tidak valid 1. Sistem menginformasikan bahwa nilai threshold atau cluster tidak valid 2. Berlanjut ke basic flow langkah 3 E-3 Belum ada indeks dari dokumen 1. Sistem menginformasikan bahwa belum ada indeks dari dokumen dalam korpus 2. Berlanjut ke basic flow langkah 7
Pre-Conditions	Ada indeks dokumen yang sudah dibangun
Post-Conditions	setiap dokumen dalam korpus terbagi ke sejumlah cluster yang diinputkan

4. Analisa Kebutuhan Data

4.1 ERD (Entitiy Relationship Diagram)



Gambar 2 Entity Relationship Diagram

4.2 Data Definition

4.2.1 Data doc

4.2.1.1 Elemen data id_doc

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk id dari dokumen dalam korpus	Numeric	0-9	-	-	integer

4.2.1.2 Elemen data path

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk path dari dokumen yang ada dalam korpus	Text	semua karakter kecuali whitespace	-	-	varchar (200)

4.2.1.3 Elemen data cluster

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk informasi cluster dari dokumen yang bersangkutan	Numeric	0-9	-	-	integer

4.2.2 Data lexicon

4.2.2.1 Elemen data token

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk data token unik yang ada	Text	a-z	-	-	char (15)

4.2.2.2 Elemen data df

Representasi	Domain	Range	Format	Presisi	Struktur data
Nilai df dari tiap token	Numeric	0-9	-	-	integer

4.2.2.3 Elemen data idf

Representasi	Domain	Range	Format	Presisi	Struktur data
Nilai idf dari tiap token	Numeric	0-9	-	-	double

4.2.3 Data stop

4.2.3.1 Elemen data word

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk kata dari setiap stopword. (primary key)	Text	a-z	-	-	varchar (15)

4.2.4 Data token_doc

4.2.4.1 Elemen data token

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk data token yang ada	Text	a-z	-	-	char (15)

4.2.4.2 Elemen data id_doc

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk id dari dokumen dalam korpus dimana token tersebut muncul	Numeric	0-9	-	-	integer

4.2.4.3 Elemen data tf

Representasi	Domain	Range	Format	Presisi	Struktur data
untuk nilai tf dari token dan dokumen tertentu	Numeric	0-9	-	-	integer

4.2.4.4 Elemen data w

Representasi	Domain	Range	Format	Presisi	Struktur data
untuk nilai bobot hubungan dari token dan dokumen tertentu	Numeric	0-9	-	-	double

4.2.5 Data token_cluster

4.2.5.1 Elemen data token

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk data token yang ada	Text	a-z	-	-	char (15)

4.2.5.2 Elemen data id_cluster

Representasi	Domain	Range	Format	Presisi	Struktur data
Untuk id dari cluster dimana token tersebut muncul	Numeric	0-9	-	-	integer

4.2.5.3 Elemen data w

Representasi	Domain	Range	Format	Presisi	Struktur data
untuk rata-rata nilai bobot hubungan dari token dan dokumen tertentu di cluster tertentu	Numeric	0-9	-	-	double

DPPL

DESKRIPSI PERANCANGAN PERANGKAT LUNAK

APLIKASI CLUSTERING DOKUMEN TEKS DENGAN MENGGUNAKAN METODE K-MEANS CLUSTERING DAN PEMBOBOTAN TF-IDF (ClustKT)

Dipersiapkan oleh:

Deddy Wijaya Suliantoro / 105301466

Program Studi Magister Teknik Informatika

Program Pasca Sarjana

Universitas Atma Jaya Yogyakarta

DAFTAR PERUBAHAN

Revisi	Deskripsi
A	
B	
C	
D	
E	
F	

INDEX TGL	-	A	B	C	D	E	F	G
Ditulis oleh								
Diperiksa oleh								
Disetujui oleh								

DAFTAR HALAMAN PERUBAHAN

Halaman	Revisi	Halaman	Revisi

DAFTAR ISI

DAFTAR GAMBAR

1. Pendahuluan

1.1 Tujuan

Dokumen Deskripsi Perancangan Perangkat Lunak (DPPL) bertujuan untuk mendefinisikan perancangan perangkat lunak ClustKT yang akan dikembangkan. Dokumen DPPL tersebut digunakan oleh pengembang perangkat lunak sebagai acuan untuk implementasi pada tahap selanjutnya.

1.2 Ruang Lingkup

Perangkat lunak ClustKT dikembangkan dengan tujuan untuk :

1. Menangani pengelolaan data korpus
2. Menangani pengelolaan data stopwords
3. Menangani proses clustering dokumen teks

1.3 Definisi dan Akronim

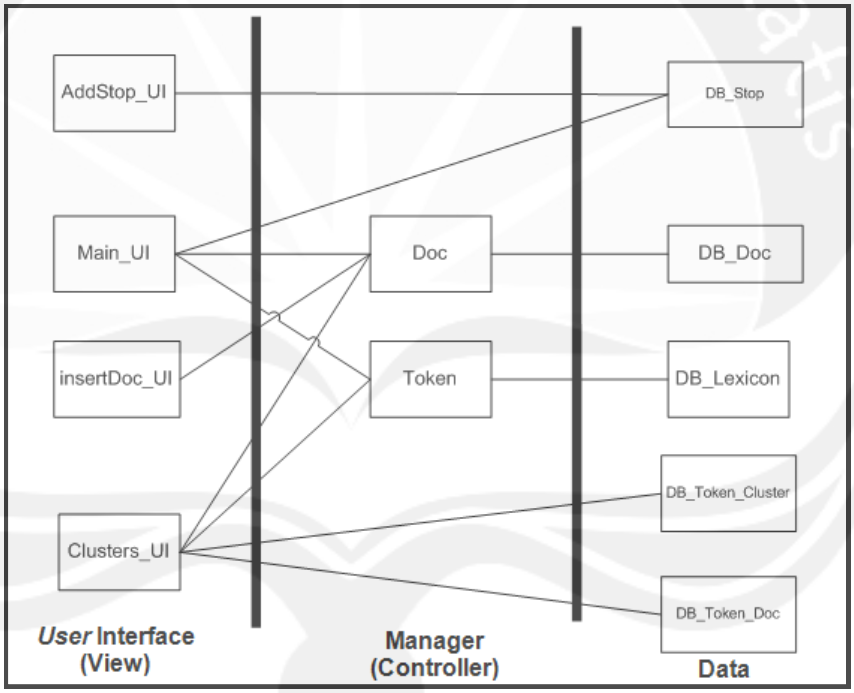
Daftar definisi akronim dan singkatan :

Keyword/Phrase	Definisi
DPPL	Deskripsi Perancangan Perangkat Lunak disebut juga Software Design Description (SDD) merupakan deskripsi dari perancangan perangkat lunak yang akan dikembangkan. Dokumen ini merupakan lanjutan dari SKPL.
ClustKT	Sistem Clustering Dokumen Teks Menggunakan Metode k-Means dan pembobotan TF-IDF
Clustering	Proses pengelompokan data ke dalam beberapa cluster (kelompok)

	tertentu.
korpus	Data dokumen teks yang berada dalam database sistem yang digunakan untuk melakukan proses clustering
stopwords	Daftar kata-kata (token) yang tidak diikutkan dalam perhitungan di proses clustering

2. Perancangan Sistem

2.1 Perancangan Arsitektur



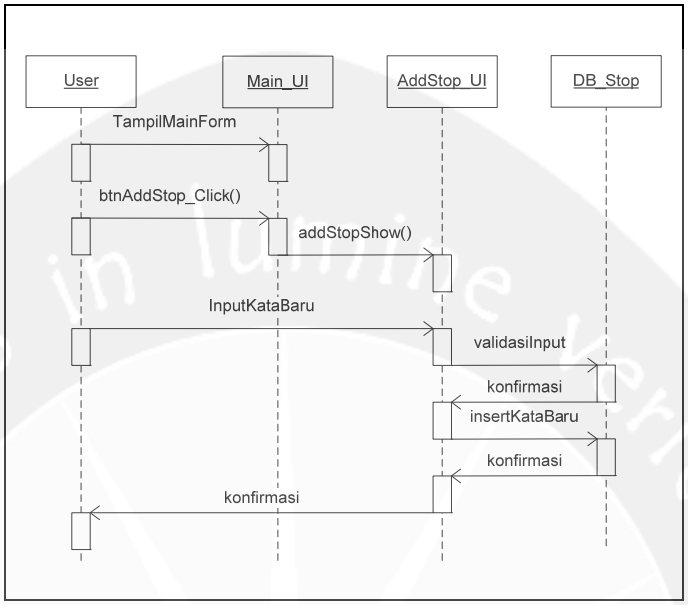
Gambar 1 Perancangan Arsitektur

2.2 Perancangan Rinci

2.2.1 Sequence Diagram

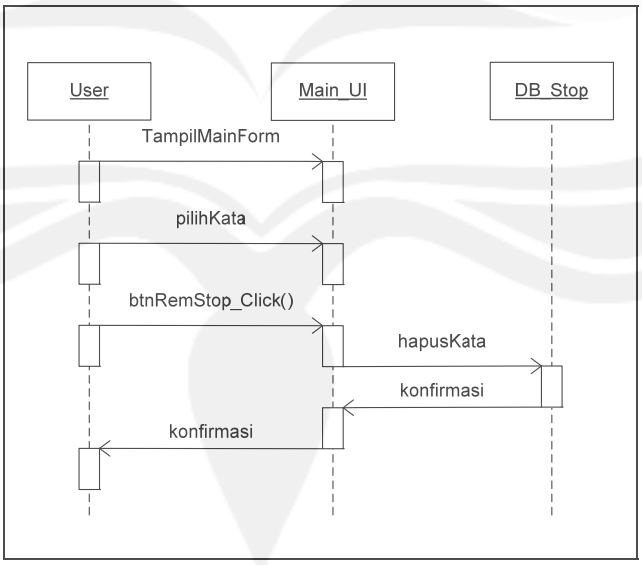
2.2.1.1 Kelola Stopwords

2.2.1.1.1 Menambah Kata ke dalam Stopwords



Gambar 2 Sequence Diagram : Penambahan stopwords

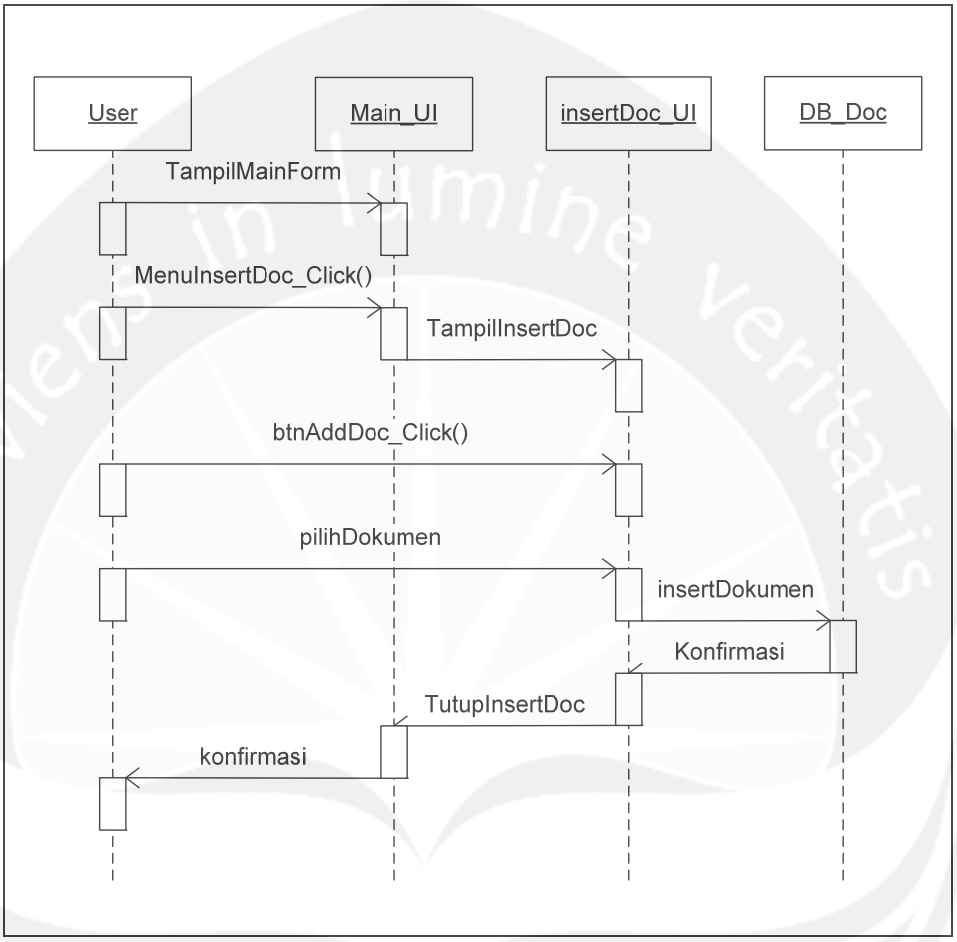
2.2.1.1.2 Menghapus Kata dari Stopwords



Gambar 3 Sequence Diagram : Menghapus kata dari daftar stopwords

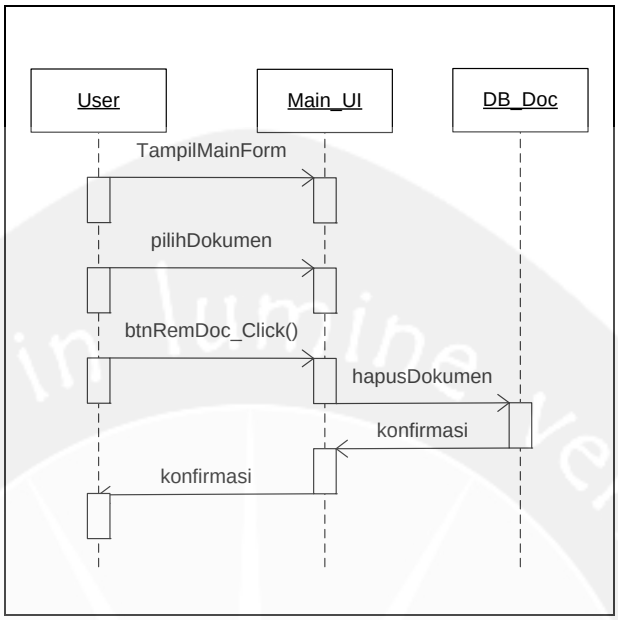
2.2.1.2 Kelola Korpus

2.2.1.2.1 Penambahan Dokumen



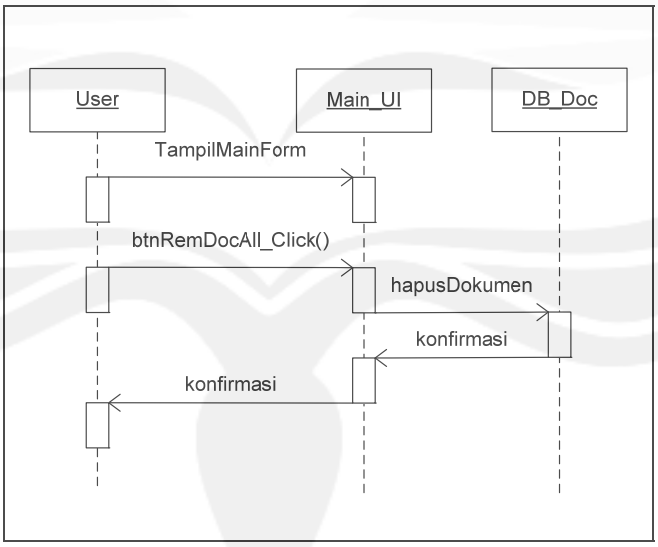
Gambar 4 Sequence Diagram : Penambahan Dokumen

2.2.1.2.2 Penghapusan Dokumen Tertentu



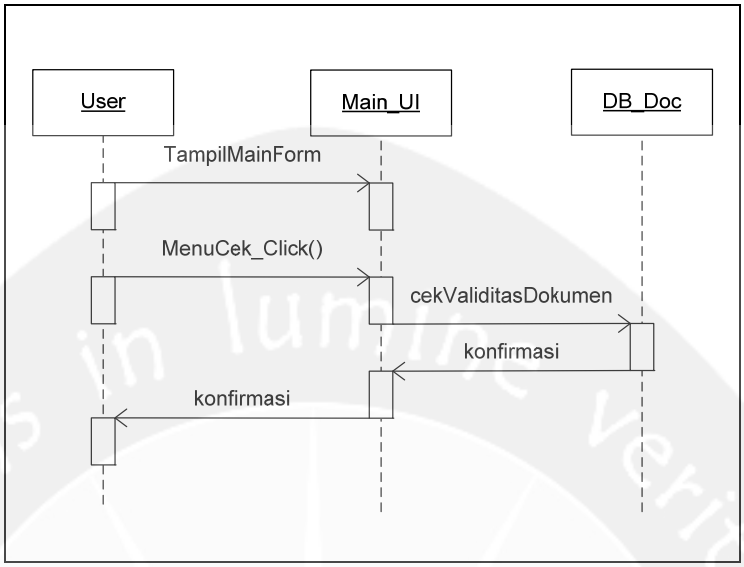
Gambar 5 Sequence Diagram : Penghapusan Dokumen Tertentu

2.2.1.2.3 Penghapusan Seluruh Dokumen



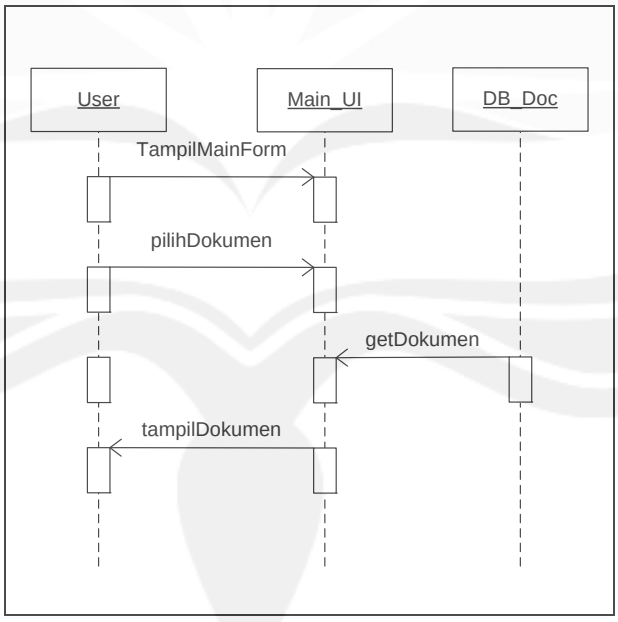
Gambar 6 Sequence Diagram : Penghapusan Seluruh Dokumen

2.2.1.2.4 Pengecekan Path Dokumen



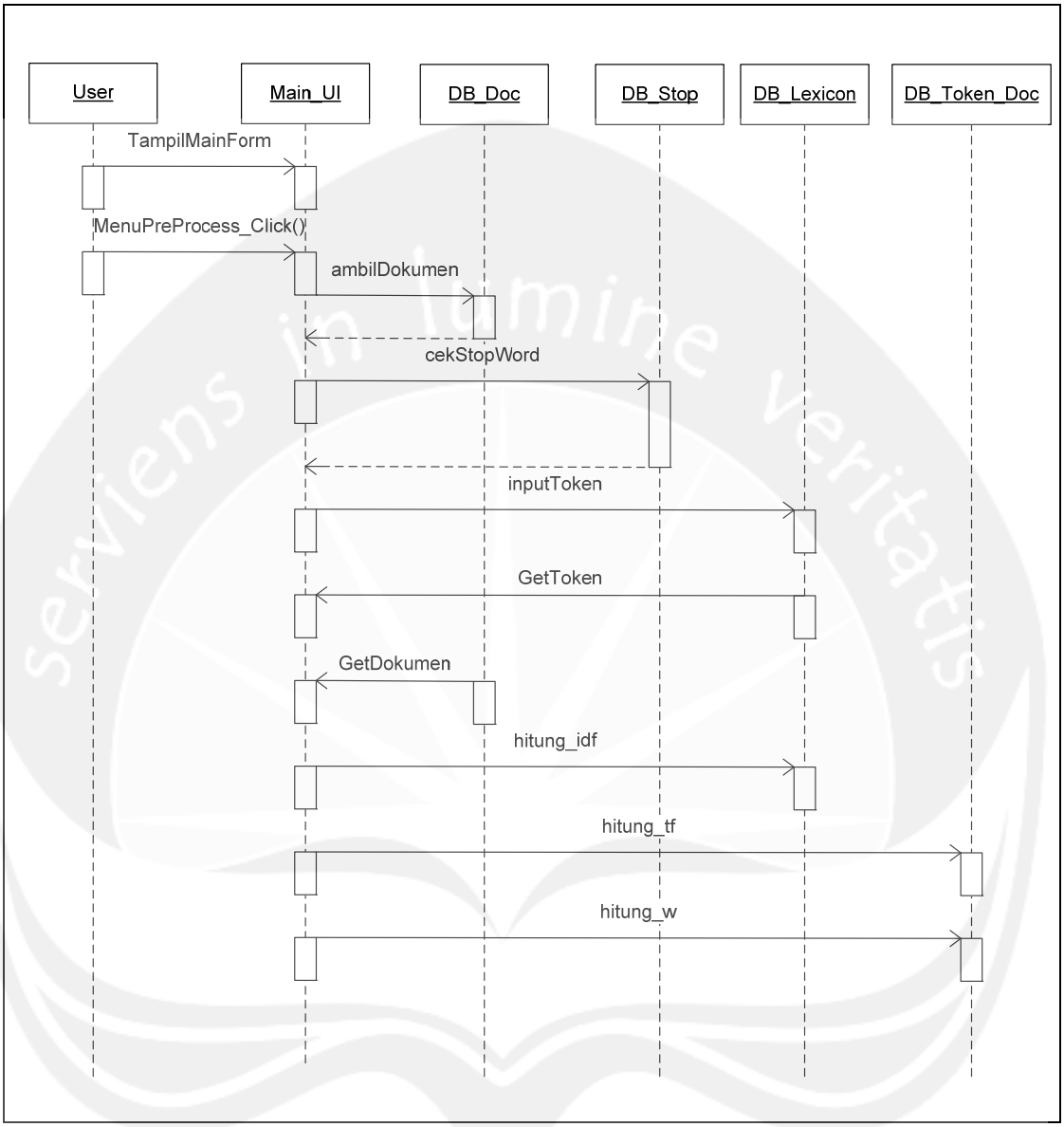
Gambar 7 Sequence Diagram : Pengecekan Path Dokumen

2.2.1.3 Melihat Isi Dokumen



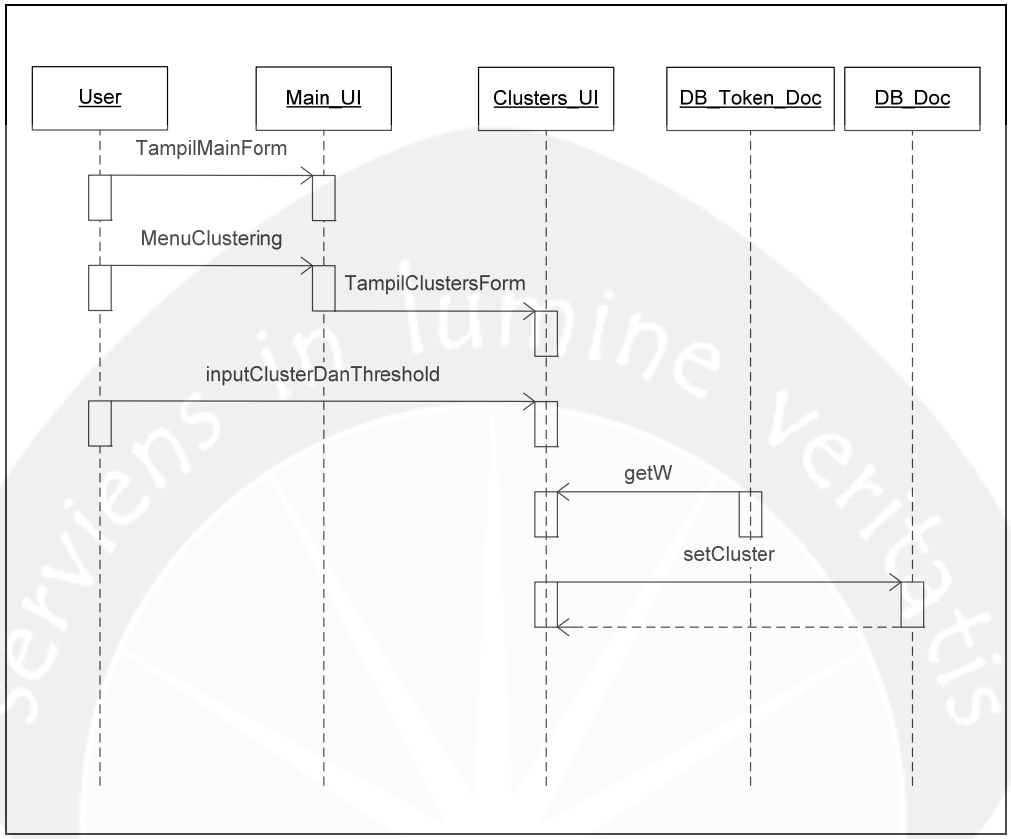
Gambar 8 Sequence Diagram : Melihat Isi Dokumen

2.2.1.4 Pembangunan Indeks Dokumen



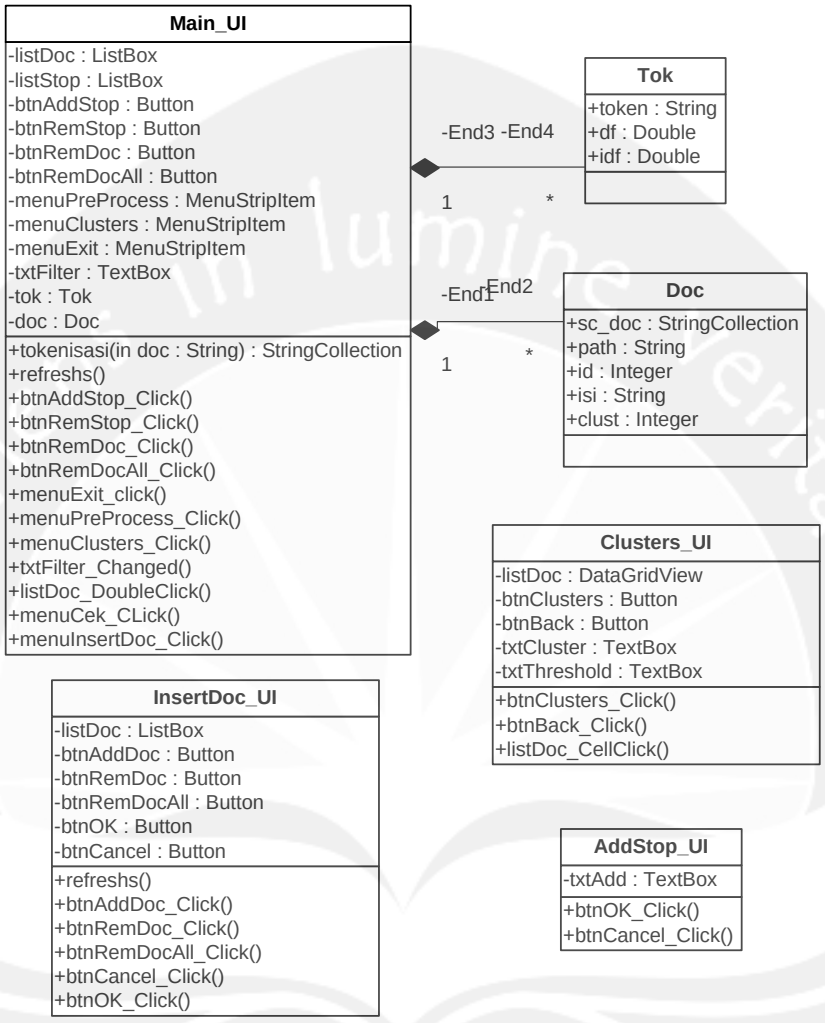
Gambar 9 Sequence Diagram : Pembangunan Indeks Dokumen

2.2.1.5 Clustering Dokumen



Gambar 10 Sequence Diagram : Clsutering Dokumen

2.2.2 Class Diagram



Gambar 11 Class Diagram

2.2.3 Deskripsi Kelas

2.2.3.1 Spesific Design Class Main_UI

Main_UI	
+tokenisasi(string) : StringCollection Operasi ini digunakan untuk melakukan proses tokenisasi terhadap sebuah string (isi dokumen).	

+refreshs() Operasi ini akan memperbarui daftar dokumen dan daftar kata pada stopwords list setiap kali dijalankan. +btnAddStop_Click() Operasi ini akan membuka sebuah dialogbox untuk memasukkan satu kata baru ke dalam stopwords list. +btnRemStop_Click() Operasi ini akan menghapus kata yang terpilih dari daftar stopwords +btnRemDoc_Click() Operasi ini akan menghapus dokumen yang terpilih dari daftar dokumen dalam korpus. +btnRemDocAll_Click() Operasi ini akan menghapus semua dokumen dalam korpus +menuExit_Click() Operasi ini akan menutup aplikasi +menuPreProcess_Click() Operasi ini akan menjalankan proses pembangunan indeks dokumen dari korpus yang ada. +menuClusters_Click() Operasi ini akan membuka form Clusters untuk melakukan proses clustering +txt_Filter_Changed() Operasi ini akan memfilter daftar stopwords berdasarkan karakter yang tertulis pada txtFilter +listDoc_DoubleClick() Operasi ini akan membuka form yang menampilkan isi dari dokumen yang terpilih dari daftar. +menuCek_Click() Operasi ini akan melakukan pengecekan terhadap semua path dokumen yang ada apakah valid atau tidak.

2.2.3.2 Spesific Design Class InsertDoc_UI

InsertDoc_UI	
+refreshs() Operasi ini akan memperbarui daftar dokumen pada listBox. +btnAddDoc_Click() Operasi ini akan membuka sebuah OpenFileDialog dimana user dapat memilih dokumen dari harddisk +btnRemDoc_Click() Operasi ini akan menghapus dokumen tertentu yang dipilih user +btnRemDocAll_Click()	

Operasi ini akan menghapus seluruh dokumen dalam daftar / listBox +btnOK_Click() Operasi ini akan memasukkan semua dokumen dalam daftar ke database, kecuali dokumen yang sudah ada atau yang tidak valid. +btnCancel_Click() Operasi ini akan membatalkan proses input dokumen dan kembali ke menu utama.

2.2.3.3 Spesific Design Class Clusters_UI

Clusters_UI	
+btnClusters_Click() Operasi ini akan memulai proses clustering terhadapa dokumen dalam korpus +btnBack_Click() Operasi ini akan menutup form dan membuka kembali form utama +listDoc_CellClick() Operasi ini akan membuka dokumen tertentu yang dipilih user dari daftar.	

2.2.3.4 Spesific Design Class AddStop_UI

AddStop_UI	
+btnOK_Click() Operasi ini akan memasukkan kata baru ke dalam daftar stopwords ke database, kecuali kata sudah ada atau yang tidak valid. +btnCancel_Click() Operasi ini akan membatalkan proses input kata dan kembali ke menu utama.	

3. Perancangan Data

3.1 Dekomposisi Data

3.1.1 Deskripsi Entitas doc

Nama	Tipe	Panjang	Keterangan
id_doc	integer	11	id dokumen, primary key
path	varchar	200	path dokumen tersimpan
cluster	integer	2	letak cluster dari dokumen

3.1.2 Deskripsi Entitas lexicon

Nama	Tipe	Panjang	Keterangan
token	char	15	token unik, primary key
df	integer	11	nilai df tiap token
idf	double	6,4	nilai idf tiap token

3.1.3 Deskripsi Entitas stop

Nama	Tipe	Panjang	Keterangan
word	char	15	token unik, primary key

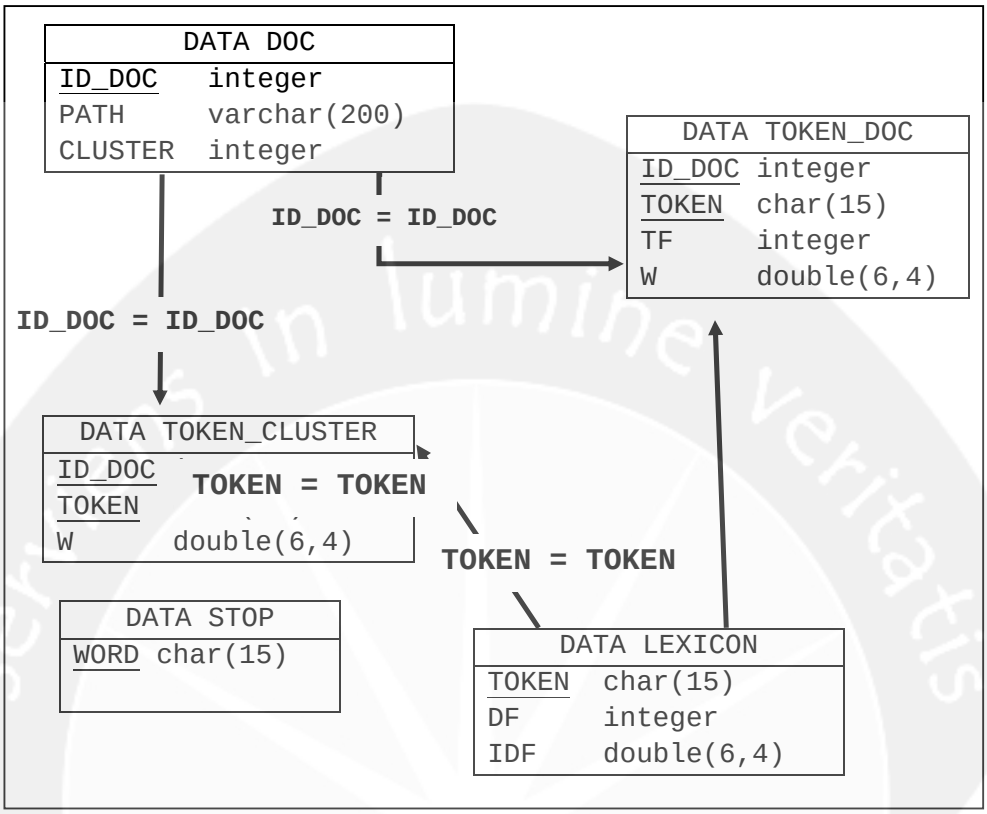
3.1.4 Deskripsi Entitas token_doc

Nama	Tipe	Panjang	Keterangan
token	char	15	token unik, primary key
id_doc	integer	11	id dokumen, primary key
tf	integer	11	nilai tf tiap token dan dokumen tertentu
w	double	6,4	nilai w tiap token dan dokumen tertentu

3.1.5 Deskripsi Entitas token_cluster

Nama	Tipe	Panjang	Keterangan
token	char	15	token unik, primary key
id_cluster	integer	11	id cluster, primary key
w	double	6,4	nilai rata-rata w tiap token dan cluster tertentu

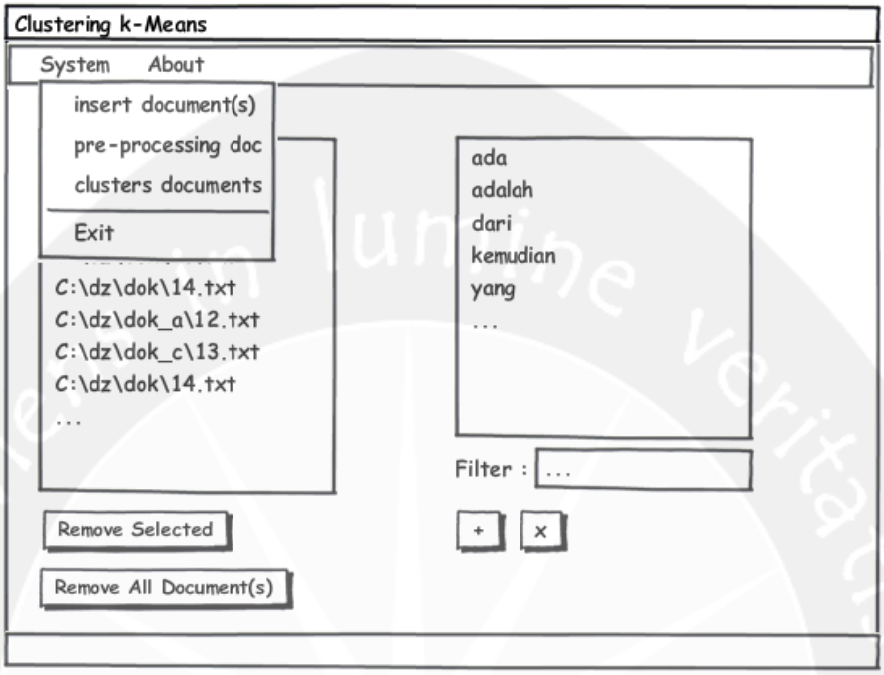
3.2 Physical Data Model



Gambar 12 Physical Data Model

4. Perancangan Antarmuka

4.1 Form Main



Gambar 13 Rancangan form Main

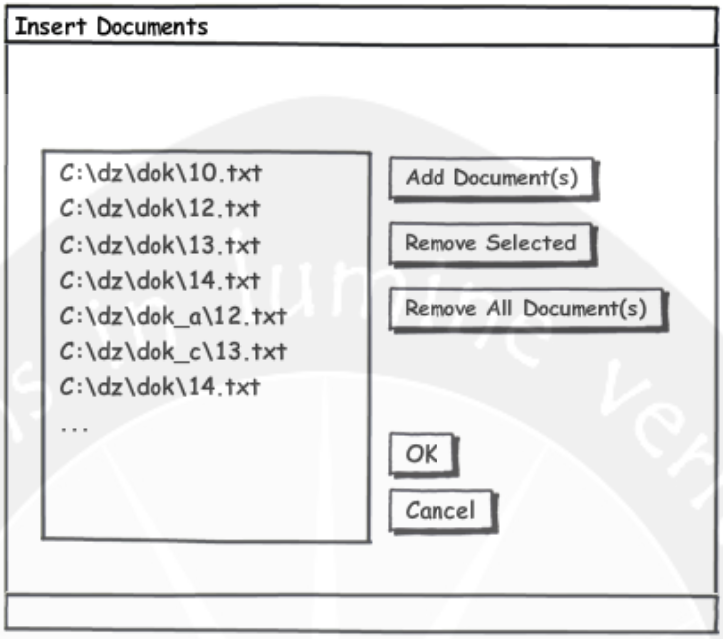
Form Main nantinya adalah Form utama dan yang pertama dijalankan ketika aplikasi clustering ini dijalankan. Form ini memuat dua listbox yang masing-masing berisi daftar korpus dan daftar stopwords yang digunakan oleh aplikasi.

Form ini juga memiliki menu untuk menuju Form dengan fungsi lain seperti form insertDoc dan form Clusters seperti ditunjukkan oleh gambar 4.

4.2 Form InsertDoc

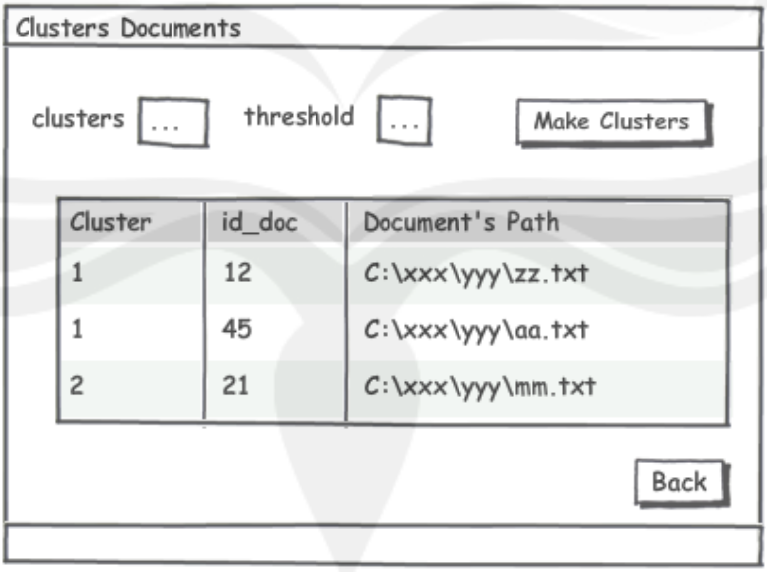
Form InsertDoc ini digunakan untuk memasukkan sejumlah dokumen ke dalam database sebagai korpus dari aplikasi. Pertama-tama, user memilih file txt yang dia pilih ke dalam daftar yang tersedia di form, kemudian saat user sudah selesai memilih, maka form ini akan ditutup dan kembali ke form Main.

Pada saat yang bersamaan, file-file yang terdaftar di listbox akan dimasukkan ke database.



Gambar 14 Rancangan Form InsertDoc

4.3 Form Clusters



Gambar 15 Rancangan Form Clusters

Gambar 6 menunjukkan rancangan antarmuka untuk form Clusters dimana user nantinya harus memasukkan jumlah cluster yang diinginkan dan nilai threshold untuk batasan iterasi.

Setelah user menekan tombol “make clusters”, maka sistem akan melakukan proses clustering terhadap dokumen yang ada di dalam korpus dan setelah selesai proses, maka sistem akan menampilkan hasilnya dalam tabel yang disediakan.

4.4 Form ShowDoc

Form ini hanya berisi satu RichTextBox yang digunakan untuk menampilkan isi dari dokumen teks yang terpilih. Form ini ditampilkan saat user memilih dokumen tertentu dari daftar korpus.