DATA MINING

PERTEMUAN Ke-13

Text Mining

Versi : 01

Text Mining

29.1 Text Mining Concepts

29.2 Text Clustering

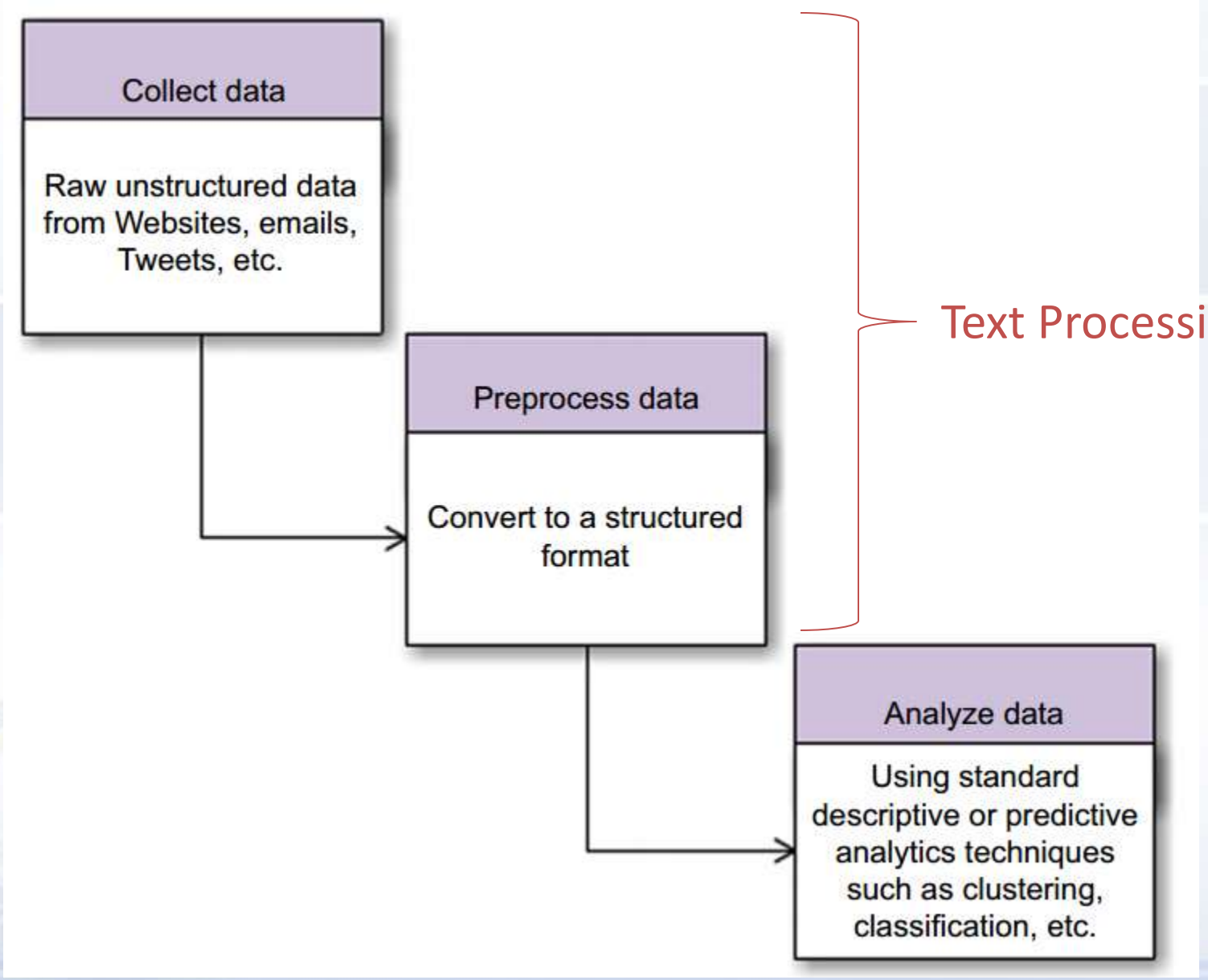
30.1 Text Classification

30.2 Data Mining Law

Konsep Text Mining

- **Cara Kerja Text Mining**
- Langkah mendasar dalam Tect Mining melibatkan konversi teks menjadi data semi-terstruktur
- Setelah mengubah teks tidak terstruktur menjadi data semi-terstruktur, tidak ada kendala untuk menerapkan teknik analisis apa pun, mengklasifikasikan, mengelompokkan, dan memprediksi.
- Teks tidak terstruktur perlu diubah menjadi kumpulan data semi-terstruktur sehingga dapat menemukan pola dan lebih baik lagi, melatih model untuk mendeteksi pola dalam teks baru dan tidak terlihat

I.4


Collect data
Raw unstructured data from Websites, emails, Tweets, etc.
Preprocess data
Convert to a structured format
Analyze data
Using standard descriptive or predictive analytics techniques such as clustering, classification, etc.
Text Processing

Versi : 01

I.5

Proses Data Mining

1. Himpunan Data

(Pemahaman dan Pengolahan Data)

DATA PRE-PROCESSING
Data Cleaning
Data Integration
Data Reduction
Data Transformation
Text Processing

2. Metode Data Mining

(Pilih Metode Sesuai Karakter Data)

Estimation
Prediction
Classification
Clustering
Association

3. Pengetahuan

(Pola/Model/Rumus/ Tree/Rule/Cluster)

4. Evaluation

(Akurasi, AUC, RMSE, Lift Ratio,…)

Word, Token and Tokenization

Document 1	This is a book on data mining.
Document 2	This book describes data mining and text mining using RapidMiner.

- Kata-kata dipisahkan oleh karakter khusus: spasi kosong
- Setiap kata disebut token
- Proses mendiskritisasi kata-kata dalam dokumen disebut tokenisasi
- Untuk tujuan kita di sini, setiap kalimat dapat dianggap sebagai dokumen terpisah, meskipun apa yang dianggap sebagai dokumen individual mungkin bergantung pada konteksnya
- Untuk saat ini, dokumen di sini hanyalah kumpulan token yang berurutan

Ketentuan Matrix

- Kita dapat menerapkan beberapa bentuk struktur pada data mentah ini dengan membuat matriks, dimana:
  - kolomnya terdiri dari semua token yang ditemukan di dua dokumen
  - sel-sel matriks adalah hitungan berapa kali token muncul
- Setiap token sekarang menjadi atribut dalam bahasa data mining standar dan setiap dokumen, contohnya

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

I.8

Term Document Matrix (TDM)

- Basically, unstructured raw data is now transformed into a format that is recognized, not only by the human users as a data table, but more importantly by all the machine learning algorithms which require such tables for training
- This table is called a document vector or term document matrix (TDM) and is the cornerstone of the preprocessing required for text mining

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	Using
Document 1	1/7 = 0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0	0
Document 2	1/10 = 0.1	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

Term Document Matrix (TDM)

- Pada dasarnya, data mentah yang tidak terstruktur kini diubah menjadi format yang dikenali, tidak hanya oleh pengguna manusia sebagai tabel data, namun yang lebih penting lagi oleh semua algoritme pembelajaran mesin yang memerlukan tabel tersebut untuk pelatihan.
- Tabel ini disebut vektor dokumen atau term document matrix (TDM) dan merupakan landasan pra-pemrosesan yang diperlukan untuk penambangan teks.

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	Using
Document 1	1/7 = 0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0	0
Document 2	1/10 = 0.1	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

TF–IDF

- Kita juga dapat memilih untuk menggunakan skor TF-IDF untuk setiap term untuk membuat vektor dokumen
- N adalah jumlah dokumen yang kami coba tambang
- Nk adalah banyaknya dokumen yang mengandung kata kunci k

$$TF = n_k/n \qquad IDF = \log_2 (N/N_k)$$

$$TF - IDF = n_k/n * \log_2 (N/N_k)$$

ExampleSet (2 examples, 0 special attributes, 12 regular attributes)

Row No.	RapidMiner	This	a	and	book	data	describes	is	mining	on	text	using
1	0	0	0.577	0	0	0	0	0.577	0	0.577	0	0
2	0.447	0	0	0.447	0	0	0.447	0	0	0	0.447	0.447

Stopwords

- Dalam dua contoh dokumen teks terdapat kemunculan kata-kata umum seperti “a”, “this”, “and”, dan istilah serupa lainnya
- Jelasnya dalam dokumen-dokumen yang lebih besar kita mengharapkan lebih banyak istilah-istilah seperti itu yang tidak benar-benar menyampaikan makna spesifik
- Sebagian besar kebutuhan tata bahasa seperti artikel, konjungsi, preposisi, dan kata ganti mungkin perlu disaring sebelum kita melakukan analisis tambahan
  - Istilah seperti ini disebut stopwords dan biasanya mencakup sebagian besar artikel, konjungsi, kata ganti, dan preposisi
  - Pemfilteran stopword biasanya merupakan langkah kedua setelah tokenisasi
- Perhatikan bahwa vektor dokumen kita memiliki ukuran yang berkurang secara signifikan setelah menerapkan pemfilteran stopword bahasa Inggris standar

Row No.	RapidMiner	book	data	describes	mining	text	using
1	0	1	1	0	1	0	0
2	1	1	1	1	2	1	1

I.12

Stopwords Bahasa Indonesia

- Lakukan googling dengan keyword: stopwords bahasa Indonesia
- Download stopword bahasa Indonesia dan gunakan di Rapidminer

Stemming

- Kata-kata seperti “dikenali (recognized)”, “dapat dikenali (recognizable)”, atau “pengakuan (recognition)” dalam penggunaan yang berbeda, namun secara kontekstual semuanya mungkin mempunyai arti yang sama, misalnya:
  - “Einstein adalah nama yang terkenal (well-recognized) di bidang fisika”
  - “Fisikawan tersebut menggunakan nama Einstein yang mudah dikenali (recognizable)”
  - “Hanya sedikit fisikawan lain yang memiliki pengenalan (recognize) nama seperti yang dimiliki Einstein”
  - Kata dasar dari semua kata yang disorot ini adalah “kenali(recognize)”
- Dengan mereduksi istilah-istilah dalam dokumen ke bentuk dasarnya, kita dapat menyederhanakan konversi teks tidak terstruktur menjadi data terstruktur karena sekarang kita hanya memperhitungkan kemunculan istilah-istilah dasar.
- Proses ini disebut stemming. Teknik stemming yang paling umum untuk text mining dalam bahasa Inggris adalah metode Porter (Porter, 1980)

I.14

A Typical Sequence of Preprocessing Steps to Use in Text Mining

Step	Action	Result
1	Tokenize	Convert each word or term in a document into a distinct attribute
2	Stopword removal	Remove highly common grammatical tokens/words
3	Filtering	Remove other very common tokens
4	Stemming	Trim each token to its most essential minimum
5	n-grams	Combine commonly occurring token pairs or tuples (more than 2)

Urutan Langkah-Langkah Pemrosesan Awal yang Biasa Digunakan dalam Penambangan Teks

1. **Pembersihan Teks (Text Cleaning):** Tahap ini melibatkan penghapusan karakter khusus, tanda baca, dan simbol yang tidak relevan atau tidak diinginkan. Pembersihan juga dapat mencakup langkah-langkah seperti mengonversi teks menjadi huruf kecil semua (lowercasing) atau menghilangkan spasi ekstra.
2. **Tokenisasi:** Proses memecah teks menjadi unit-unit yang lebih kecil yang disebut token. Token bisa berupa kata, frasa, atau karakter tergantung pada kebutuhan analisis.
3. **Penghapusan Stop Words:** Stop words adalah kata-kata umum yang sering muncul dalam teks (seperti "dan", "atau", "di", dll.) yang mungkin tidak memberikan informasi signifikan untuk analisis tertentu. Penghapusan stop words dapat membantu fokus pada kata-kata kunci yang lebih penting.

4. **Stemming atau Lemmatisasi:** Tahapan ini bertujuan untuk mengonversi kata-kata ke bentuk dasarnya agar kata-kata dengan akar yang sama dapat diidentifikasi sebagai satu entitas. Stemming memotong akhiran kata, sementara lemmatisasi mengubah kata-kata ke bentuk dasarnya (lemma).
5. **Vektorisasi:** Proses mengubah teks menjadi representasi vektor numerik. Metode umum termasuk penggunaan model seperti TF-IDF (Term Frequency-Inverse Document Frequency) atau Word Embeddings seperti Word2Vec atau GloVe.
6. **Pengkodean Teks (Text Encoding):** Jika diperlukan, teks kemudian dikodekan ke dalam format numerik untuk digunakan dalam algoritma pembelajaran mesin.
7. **Eksplorasi dan Analisis Lanjutan:** Setelah tahapan pemrosesan awal selesai, data teks siap untuk dieksplorasi lebih lanjut, dan analisis lebih lanjut dapat dilakukan. Ini bisa termasuk pemodelan, klasifikasi, klasterisasi, atau tugas-tugas lainnya sesuai dengan tujuan spesifik dari proyek text mining.

I.17

N-Grams

- There are families of words in the spoken and written language that typically go together
  - The word “Good” is usually followed by either “Morning,” “Afternoon,” “Evening,” “Night,” or in Australia, “Day”
  - Grouping such terms, called **n-grams**, and analyzing them statistically can present new insights
- Search engines use word n-gram models for a variety of applications, such as:
  - Automatic translation, identifying speech patterns, checking misspelling, entity detection, information extraction, among many different use cases

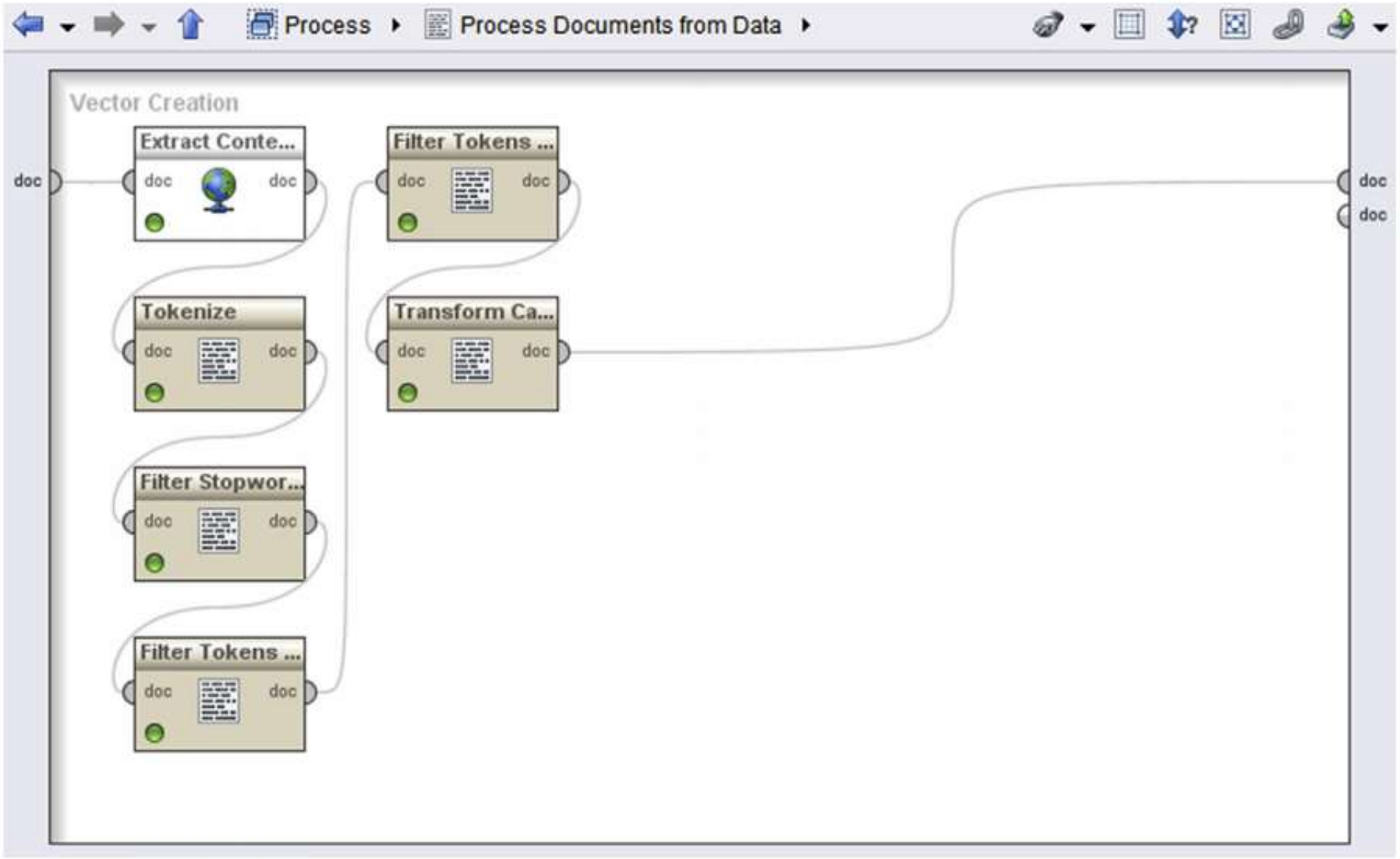
Row...	label	RapidMiner	book	book_data	book_descr...	data	data_mining	describes	describes_data	mining	mining_text	mining_usi...	text_0	text_mining	using	using_RapidMiner
1	text1	0	0.447	0.447	0	0.447	0.447	0	0	0.447	0	0	0	0	0	0
2	text2	0.243	0.243	0	0.243	0.243	0.243	0.243	0.243	0.485	0.243	0.243	0.243	0.243	0.243	0.243

N-Grams

- Ada kelompok kata dalam bahasa lisan dan tulisan yang biasanya menyatu
- Kata “Good” biasanya diikuti dengan “Morning”, “Afternoon”, “Night”, “Evening”, atau di Australia, “Day”.
- Mengelompokkan istilah-istilah tersebut, yang disebut n-gram, dan menganalisisnya secara statistik dapat memberikan wawasan baru
- Mesin pencari menggunakan model kata n-gram untuk berbagai aplikasi, seperti:
- Terjemahan otomatis, mengidentifikasi pola bicara, memeriksa kesalahan ejaan, deteksi entitas, ekstraksi informasi, dan banyak kasus penggunaan lainnya

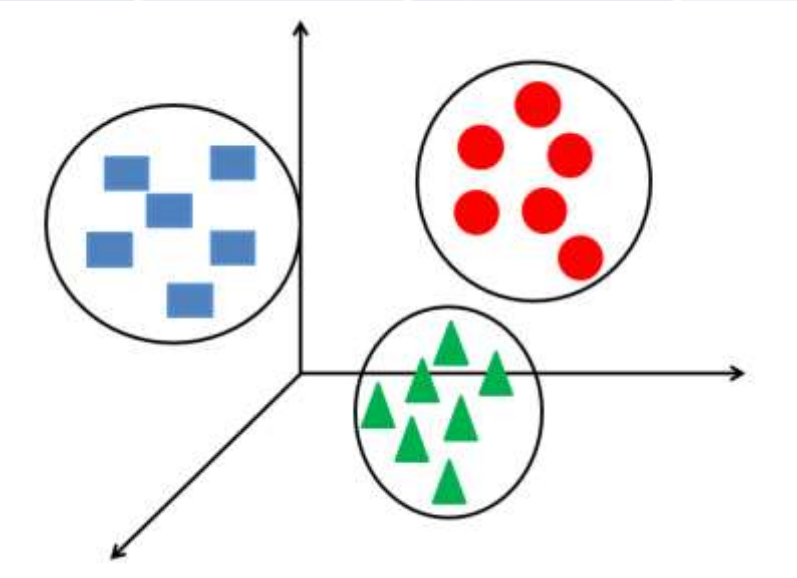
I.19

Rapidminer Process of Text Mining



Versi : 01

29.2 TEXT CLUSTERING



Latihan

- Lakukan eksperimen mengikuti buku Matthew North (Data Mining for the Masses) Chapter 12 (Text Mining), 2012, p 189-215

- Datasets: Federalist Papers

- Pahami alur text mining yang dilakukan dan sesuaikan dengan konsep yang sudah dipelajari

1. Business Understanding

- Motivation:
  - Gillian is a historian, and she has recently curated an exhibit on the Federalist Papers, the essays that were written and published in the late 1700's
  - The essays were published anonymously under the author name 'Publius', and no one really knew at the time if 'Publius' was one individual or many
  - After Alexander Hamilton died in 1804, some notes were discovered that revealed that he (Hamilton), James Madison and John Jay had been the authors of the papers
  - The notes indicated specific authors for some papers, but not for others:
    - John Jay was revealed to be the author for papers 3, 4 and 5
    - James Madison for paper 14
    - Hamilton for paper 17
    - Paper 18 had no author named, but there was evidence that Hamilton and Madison worked on that one together
- Objective:
  - Gillian would like to analyze paper 18's content in the context of the other papers with known authors, to see if she can generate some evidence that the suspected collaboration between Hamilton and Madison is in fact

1. Pemahaman Bisnis

- Motivasi:
  - Gillian adalah seorang sejarawan, dan dia baru-baru ini menjadi kurator sebuah pameran tentang Federalist Papers, esai yang ditulis dan diterbitkan pada akhir tahun 1700-an.
  - Esai-esai tersebut diterbitkan secara anonim dengan nama penulis 'Publius', dan tidak ada yang benar-benar tahu pada saat itu apakah 'Publius' adalah satu atau beberapa individu.
  - Setelah Alexander Hamilton meninggal pada tahun 1804, ditemukan beberapa catatan yang mengungkapkan bahwa dia (Hamilton), James Madison dan John Jay adalah penulis makalah tersebut.

Pemahaman Bisnis

- Catatan tersebut menunjukkan penulis tertentu untuk beberapa makalah, tetapi tidak untuk makalah lainnya:
    - John Jay diturunkan menjadi penulis makalah 3, 4 dan 5
    - James Madison untuk makalah 14
    - Hamilton untuk makalah 17
    - Makalah 18 tidak menyebutkan nama penulisnya, tetapi ada bukti bahwa Hamilton dan Madison mengerjakannya bersama-sama
- Objektif:
- Gillian ingin menganalisis isi makalah 18 dalam konteks makalah lain dengan penulis terkenal, untuk melihat apakah dia dapat menghasilkan beberapa bukti bahwa dugaan kolaborasi antara Hamilton dan Madison sebenarnya adalah sebuah kejahatan.

2. Data Understanding

- The Federalist Papers are available through a number of sources:
  - They have been re-published in book form, they are available on a number of different web sites
  - Their text is archived in many libraries throughout the world

- Gillian's data set is simple (6 dataset):
  - Federalist03_Jay.txt
  - Federalist04_Jay.txt
  - Federalist05_Jay.txt
  - Federalist14_Madison.txt
  - Federalist17_Hamilton.txt
  - Federalist18_Collaboration.txt (*suspected*)

2. Pemahaman Data

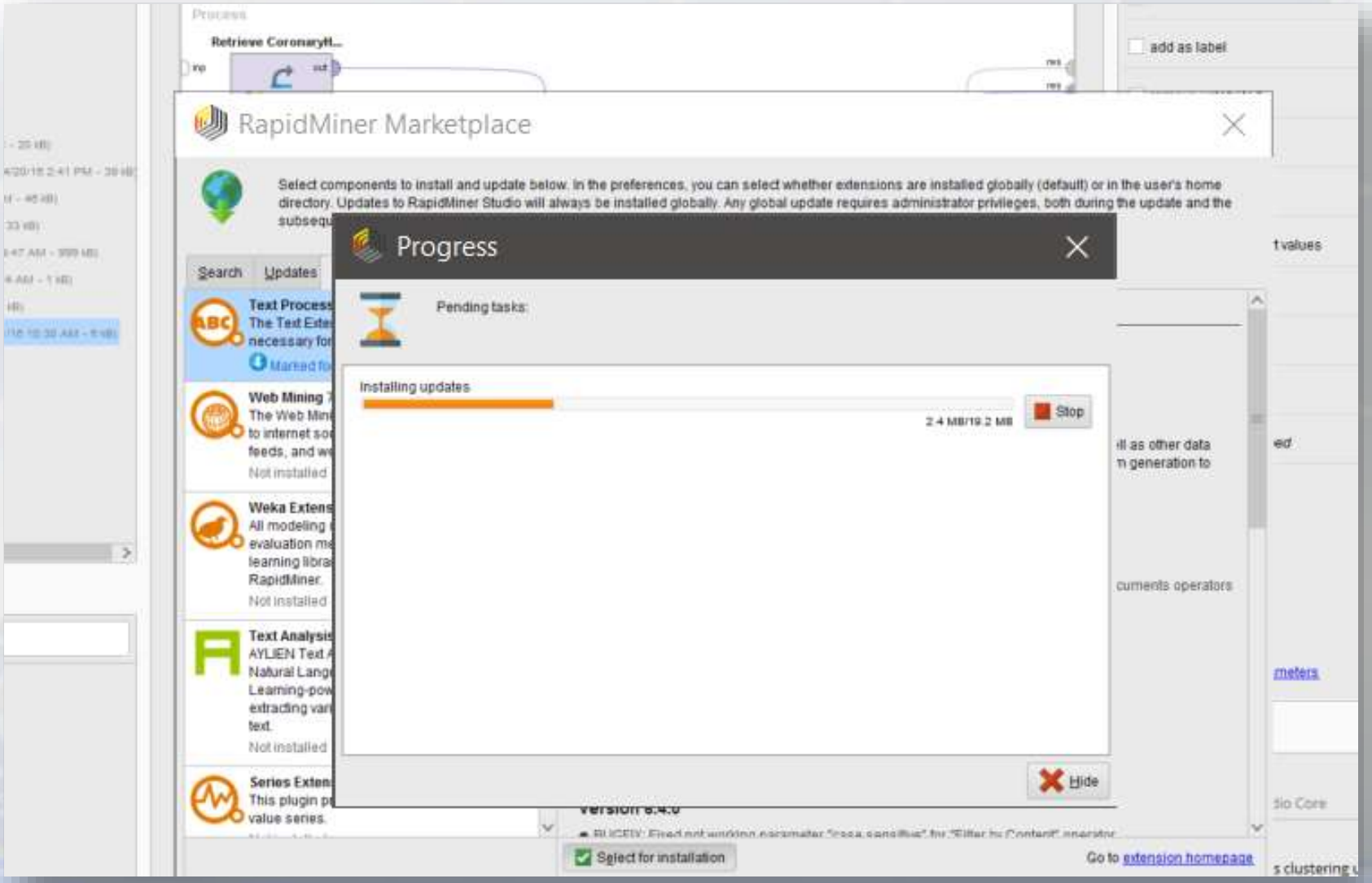
- Federalist Papers tersedia melalui sejumlah sumber:
  - Mereka telah diterbitkan ulang dalam bentuk buku, dan tersedia di sejumlah situs web berbeda
  - Teks mereka diarsipkan di banyak perpustakaan di seluruh dunia

- Kumpulan data Gillian sederhana (6 kumpulan data):
  - Federalis03_Jay.txt
  - Federalis04_Jay.txt
  - Federalis05_Jay.txt
  - Federalis14_Madison.txt
  - Federalis17_Hamilton.txt
  - Federalist18_Collaboration.txt (diduga)

I.27

Modeling

Text Processing Extension Installation



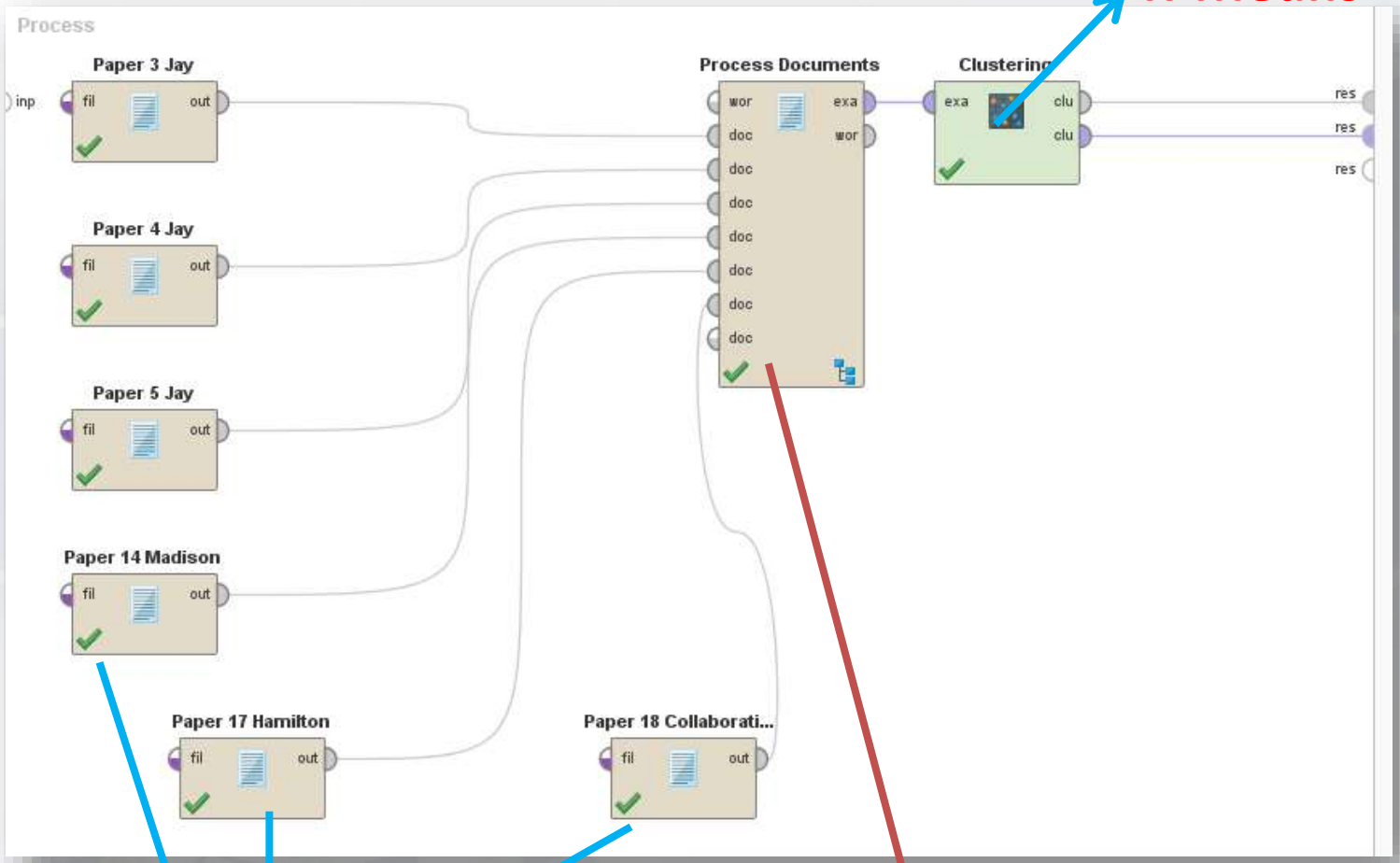
Versi : 01

Modeling

I.28

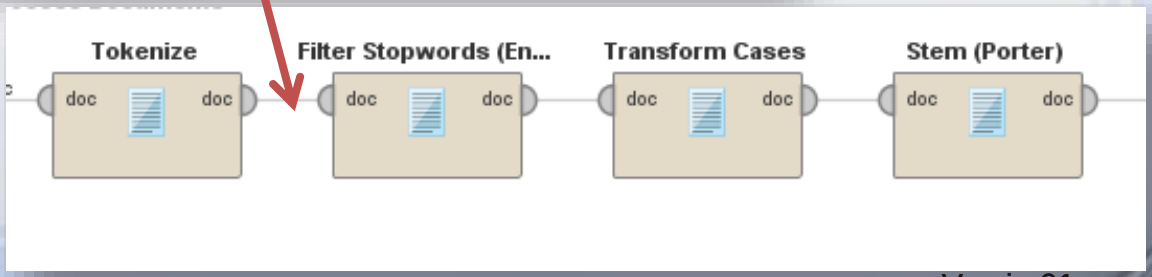
Operator

K-Means



Operator

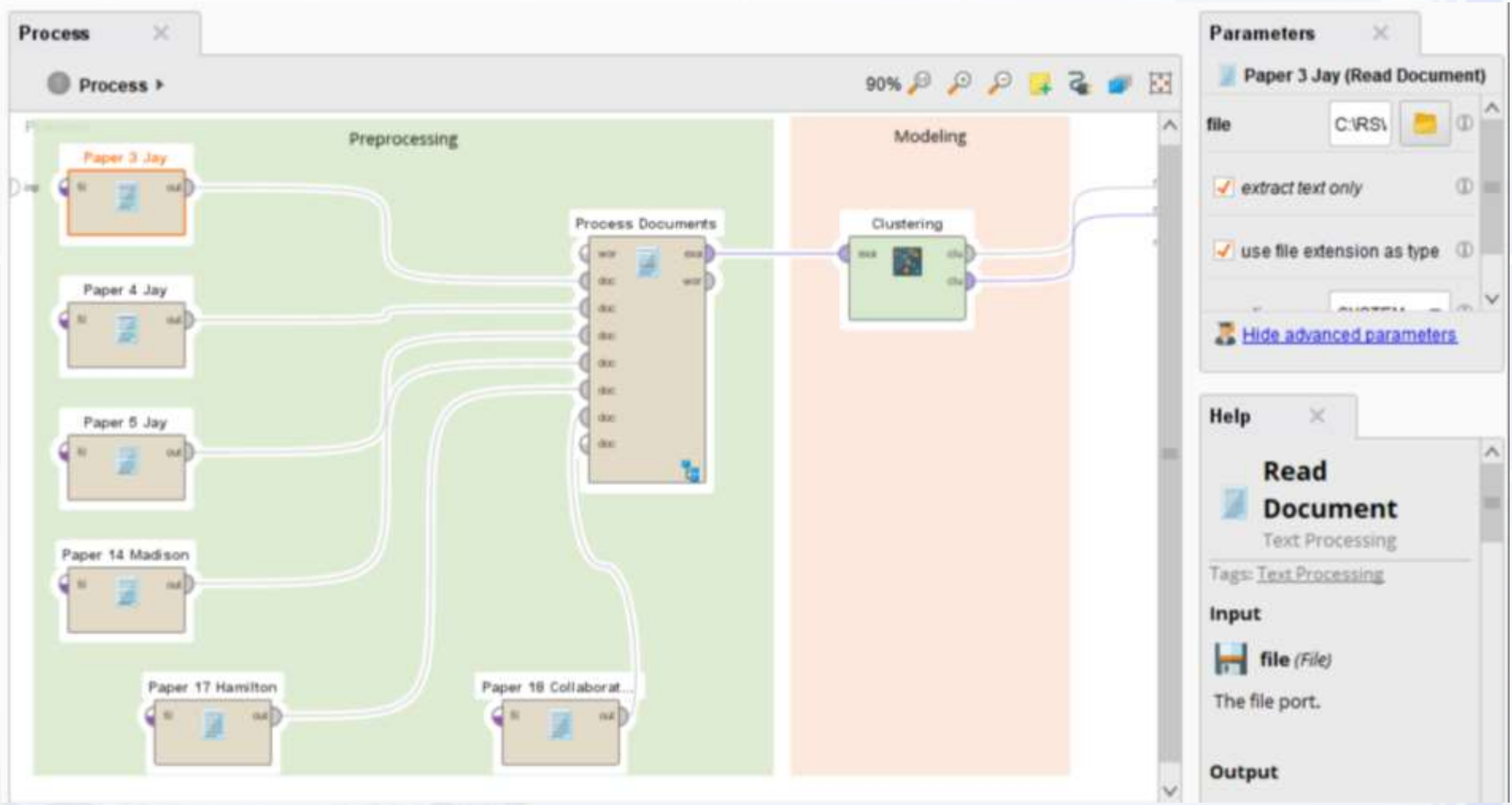
Read Document



Versi : 01

Modelling with Annotation

I.29



Versi : 01

Evaluasi

- Gillian merasa yakin bahwa makalah 18 adalah kolaborasi yang tidak disumbangkan oleh John Jay
- Kosakata dan struktur tata bahasanya sangat berbeda dengan Hamilton dan Madison

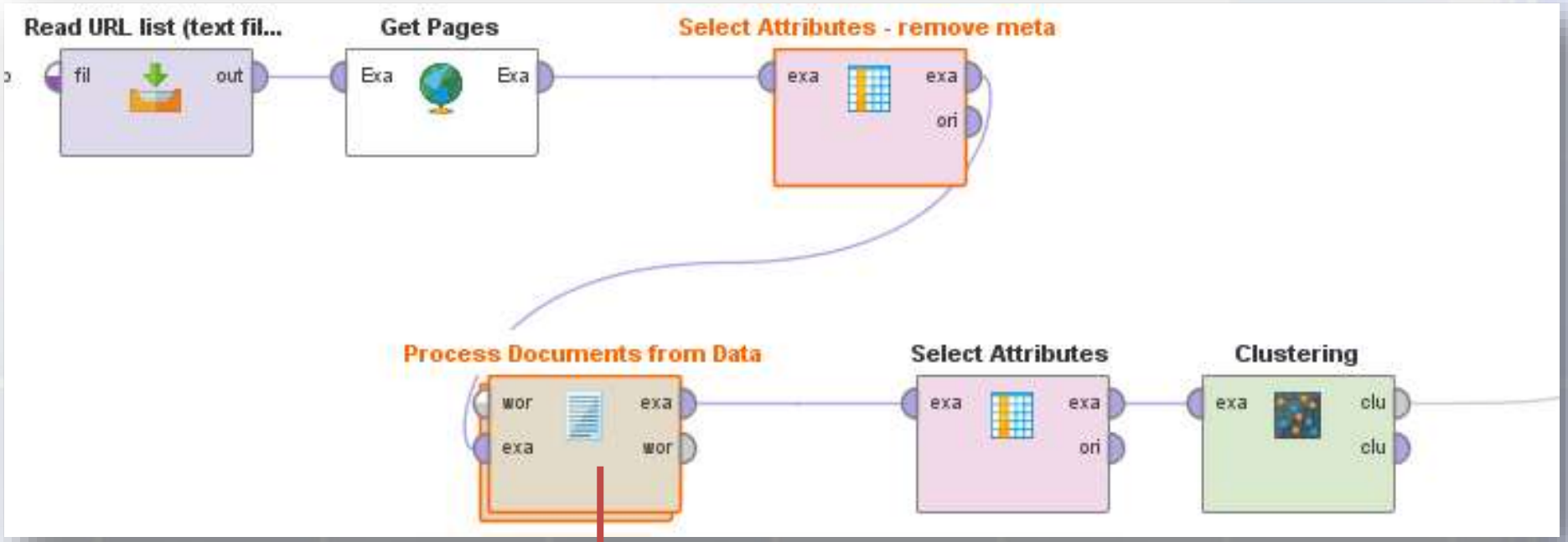
ExampleSet (Process Documents) | Cluster Model (Clustering)

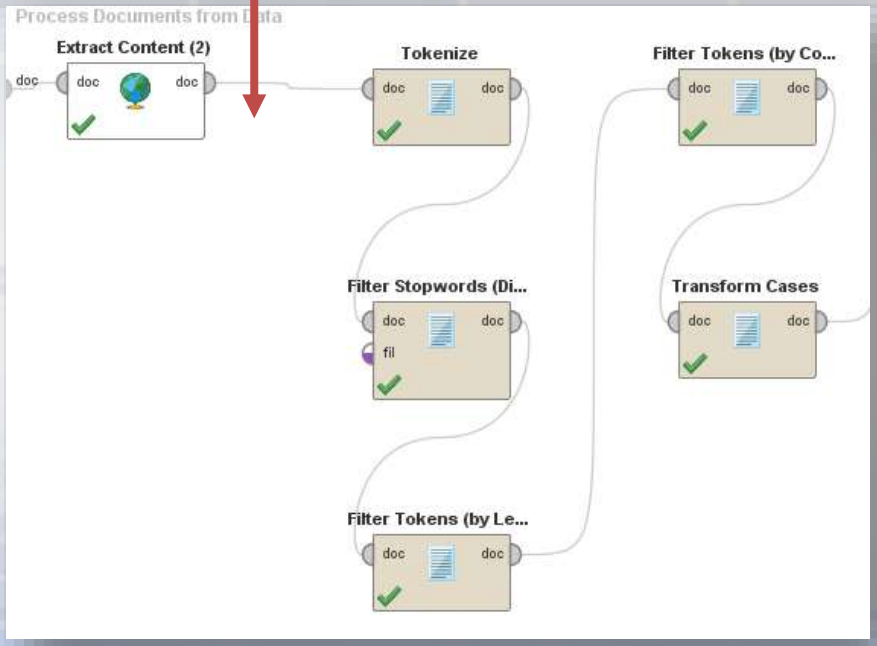
ExampleSet (6 examples, 7 special attributes, 1523 regular attributes) Filter (6 / 6 examples): all

Row No.	id	cluster	file_type	metadata_file	metadata_d...	metadata_p...	metadata_si...
1	1	cluster_0	txt	\Federalist03_Jay.txt	Jun 11, 2015 ...	C:\RSWLectu...	8760
2	2	cluster_0	txt	\Federalist04_Jay.txt	Jun 11, 2015 ...	C:\RSWLectu...	9729
3	3	cluster_0	txt	\Federalist05_Jay.txt	Jun 8, 2015 1...	C:\RSWLectu...	8276
4	4	cluster_1	txt	\Federalist14_Madison.txt	Jun 8, 2015 1...	C:\RSWLectu...	12741
5	5	cluster_1	txt	\Federalist17_Hamilton.txt	Jun 8, 2015 1...	C:\RSWLectu...	9720
6	6	cluster_1	txt	\Federalist18_Collaboration.txt	Jun 8, 2015 1...	C:\RSWLectu...	12962

Latihan

- Lakukan eksperimen mengikuti buku Vijay Kotu (Predictive Analytics and Data Mining) Chapter 9 (Text Mining), Case Study 1: Keyword Clustering, p 284-287
- Datasets (file pages.txt):
  1. https://www.cnnindonesia.com/olahraga
  2. https://www.cnnindonesia.com/ekonomi
- Gunakan stopword Bahasa Indonesia (ada di folder dataset), dengan operator Stopword (Dictionary) dan pilih file stopword-indonesia.txt
- Untuk mempermudah, copy/paste file 09_Text_9.3.1_keyword_clustering_webmining.rmp ke Repository dan kemudian buka di Rapidminer
  - Pilih file pages.txt yang berisis URL pada Read URL

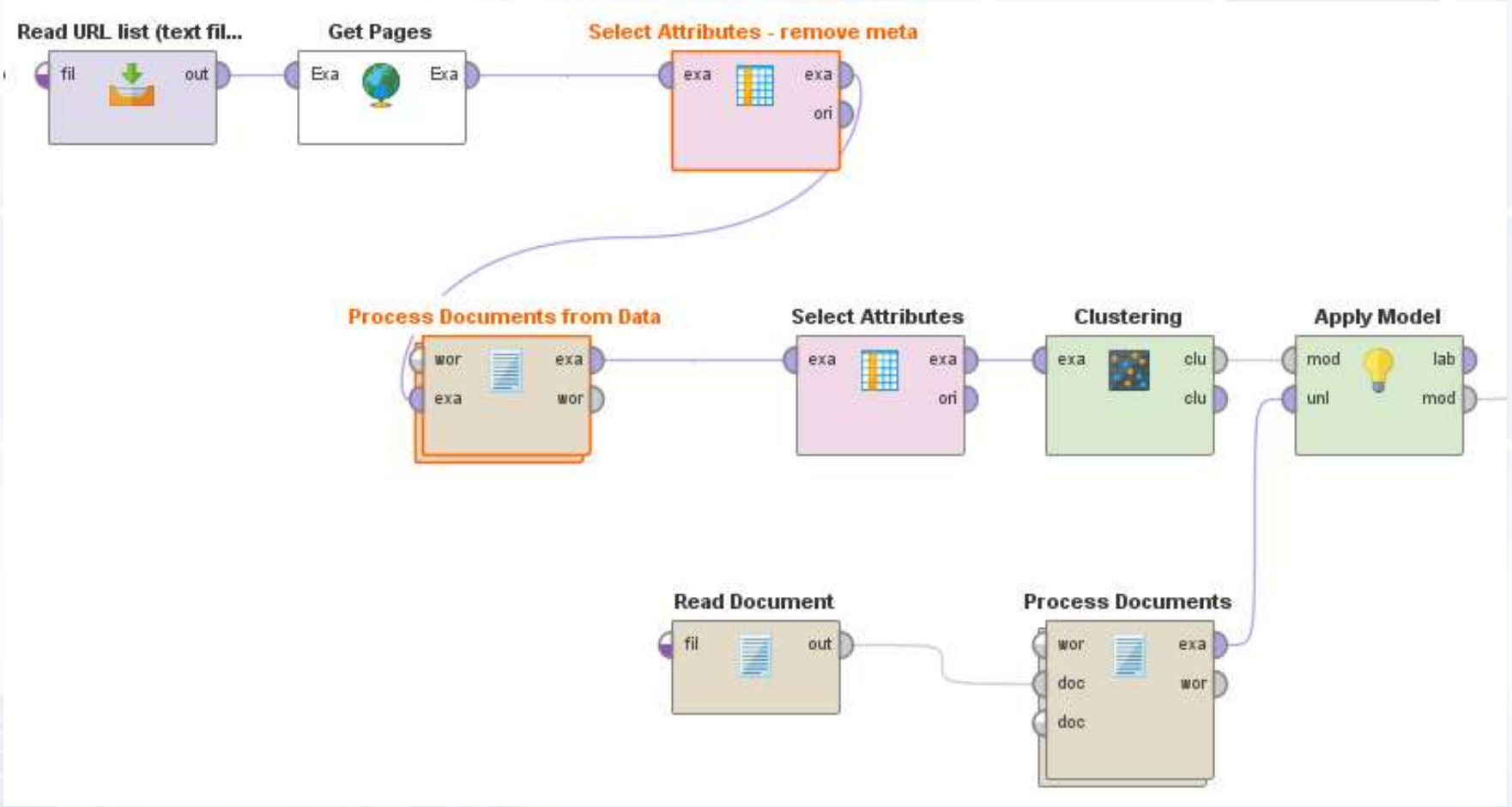

Read URL list (text fil...
fil
out
Get Pages
Exa
Exa
Select Attributes - remove meta
exa
exa
ori
Process Documents from Data
wor
exa
exa
wor
Select Attributes
exa
exa
ori
Clustering
exa
clu
clu


Process Documents from Data
Extract Content (2)
doc
doc
doc
Tokenize
doc
doc
Filter Tokens (by Co...
doc
doc
Filter Stopwords (Di...
doc
doc
fil
Transform Cases
doc
doc
Filter Tokens (by Le...
doc
doc

Versi : 01

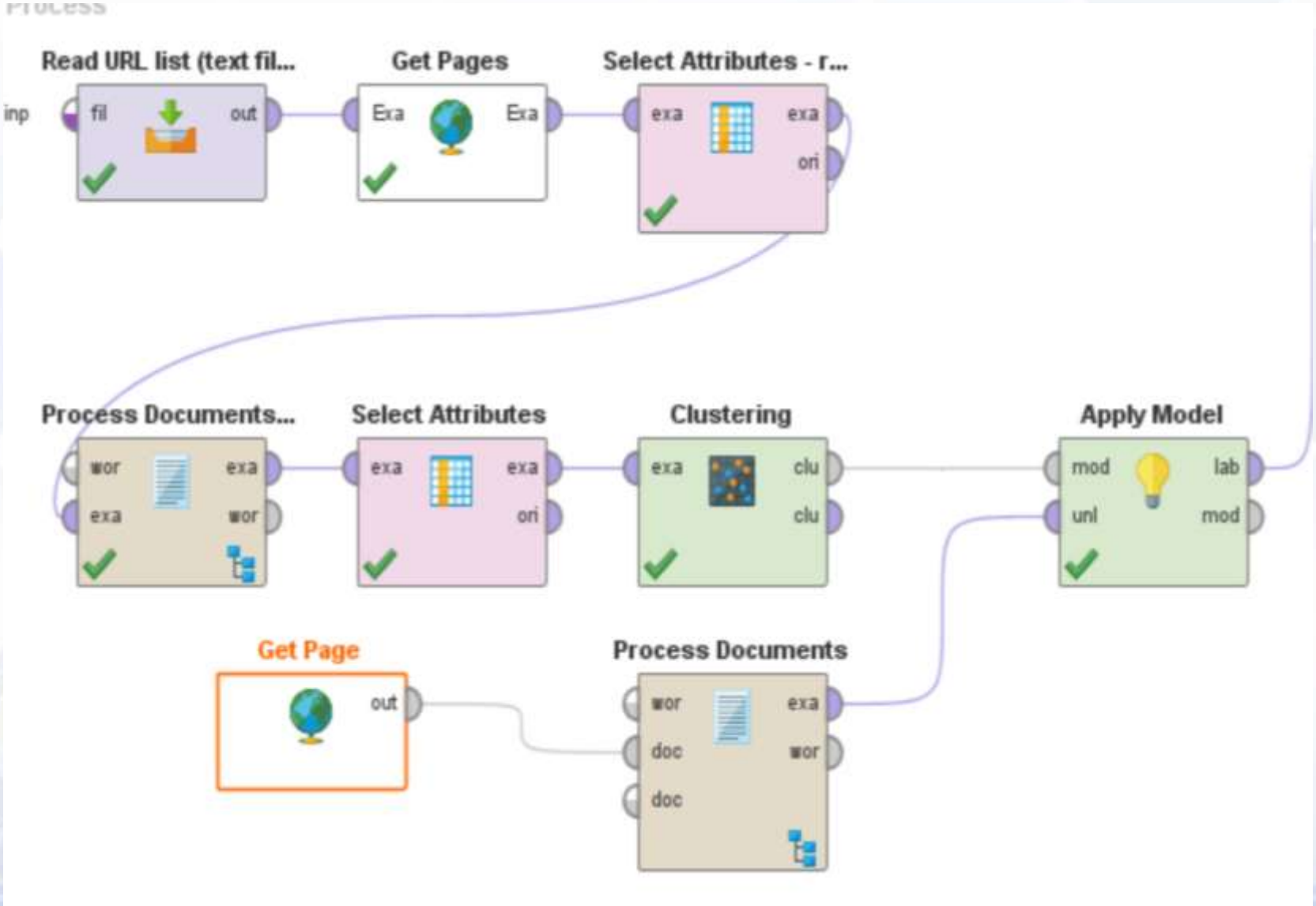
I.33

Testing Model (Read Document)



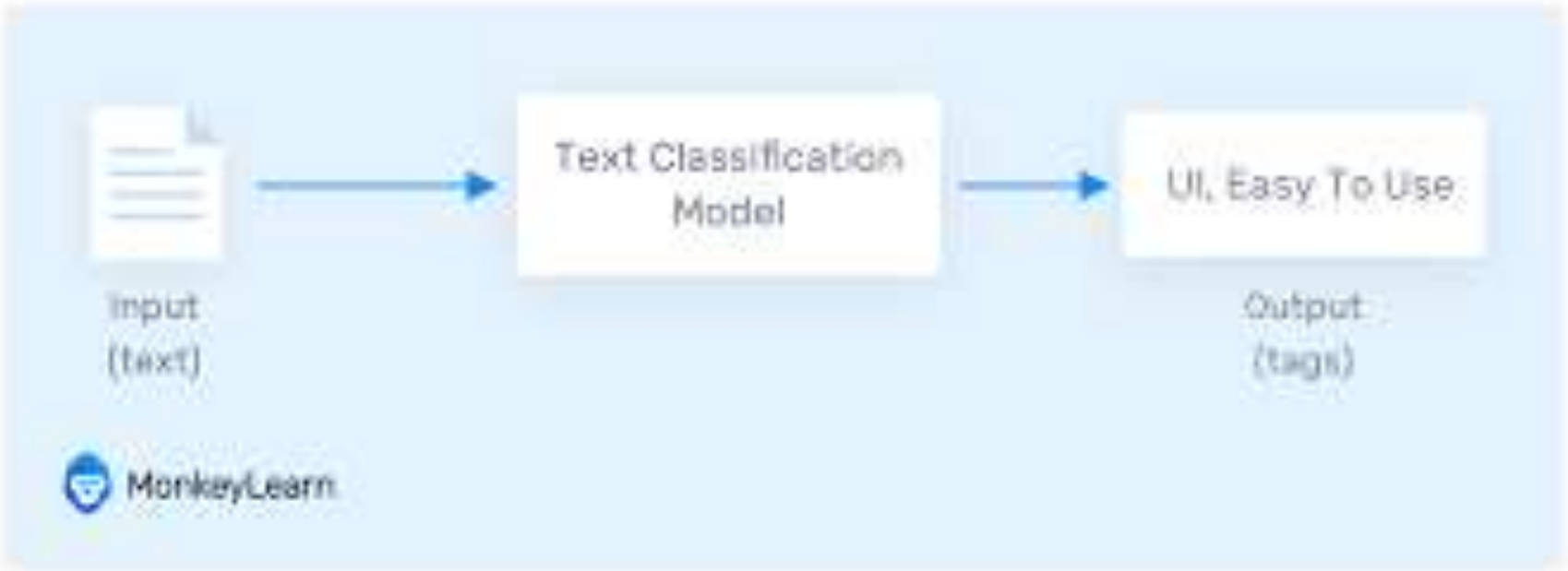
I.34

Testing Model (Get Page)



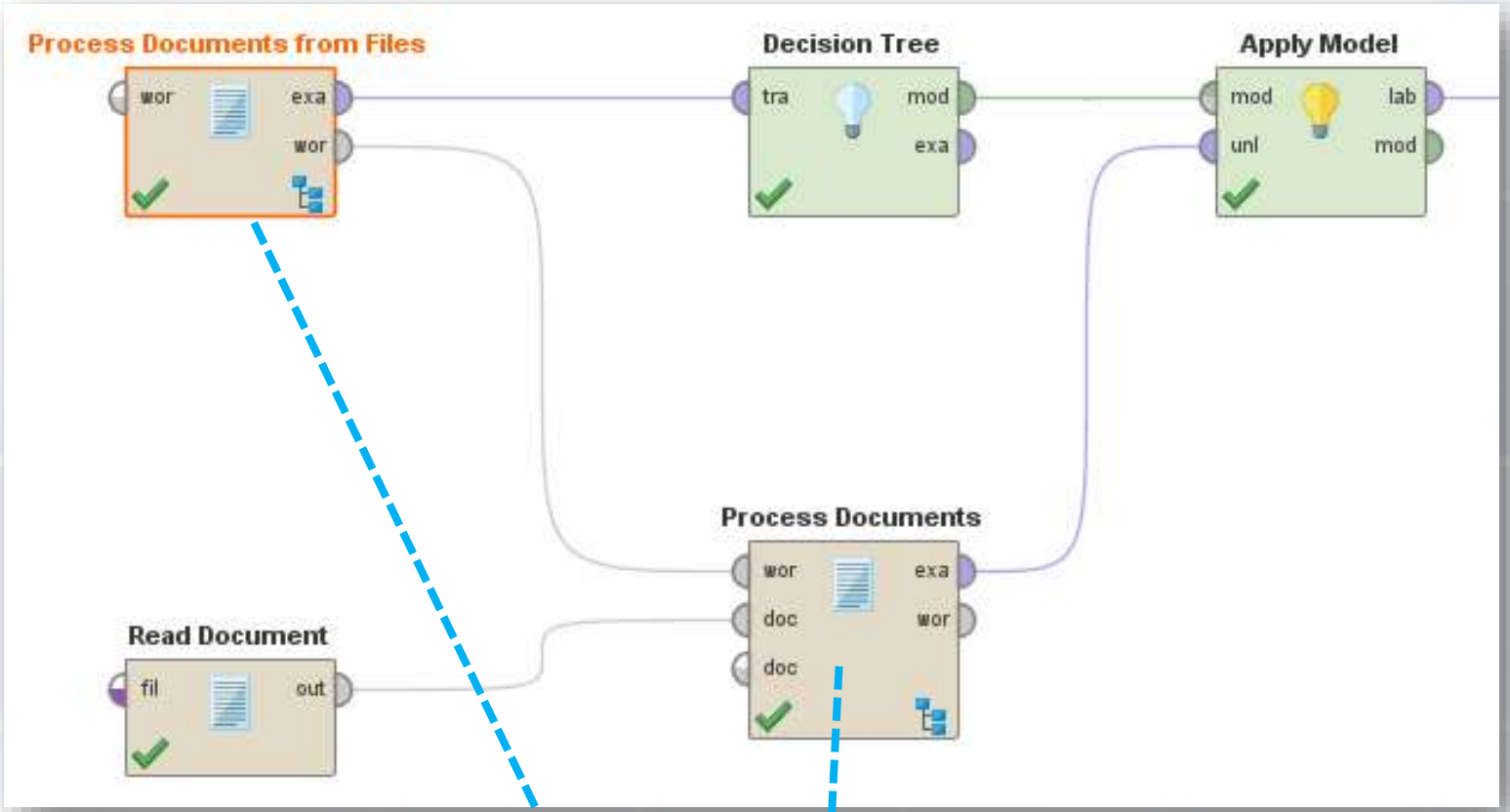
I.35

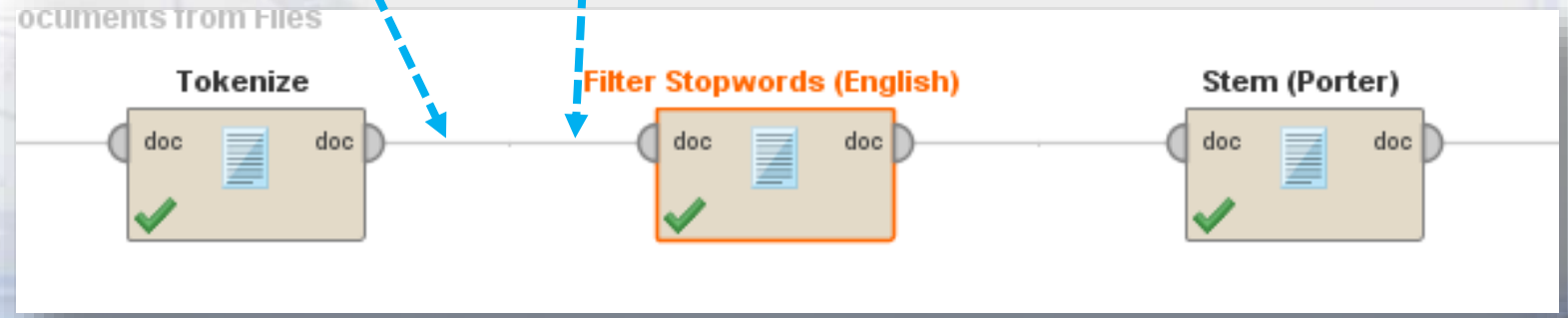
TEXT CLASSIFICATION

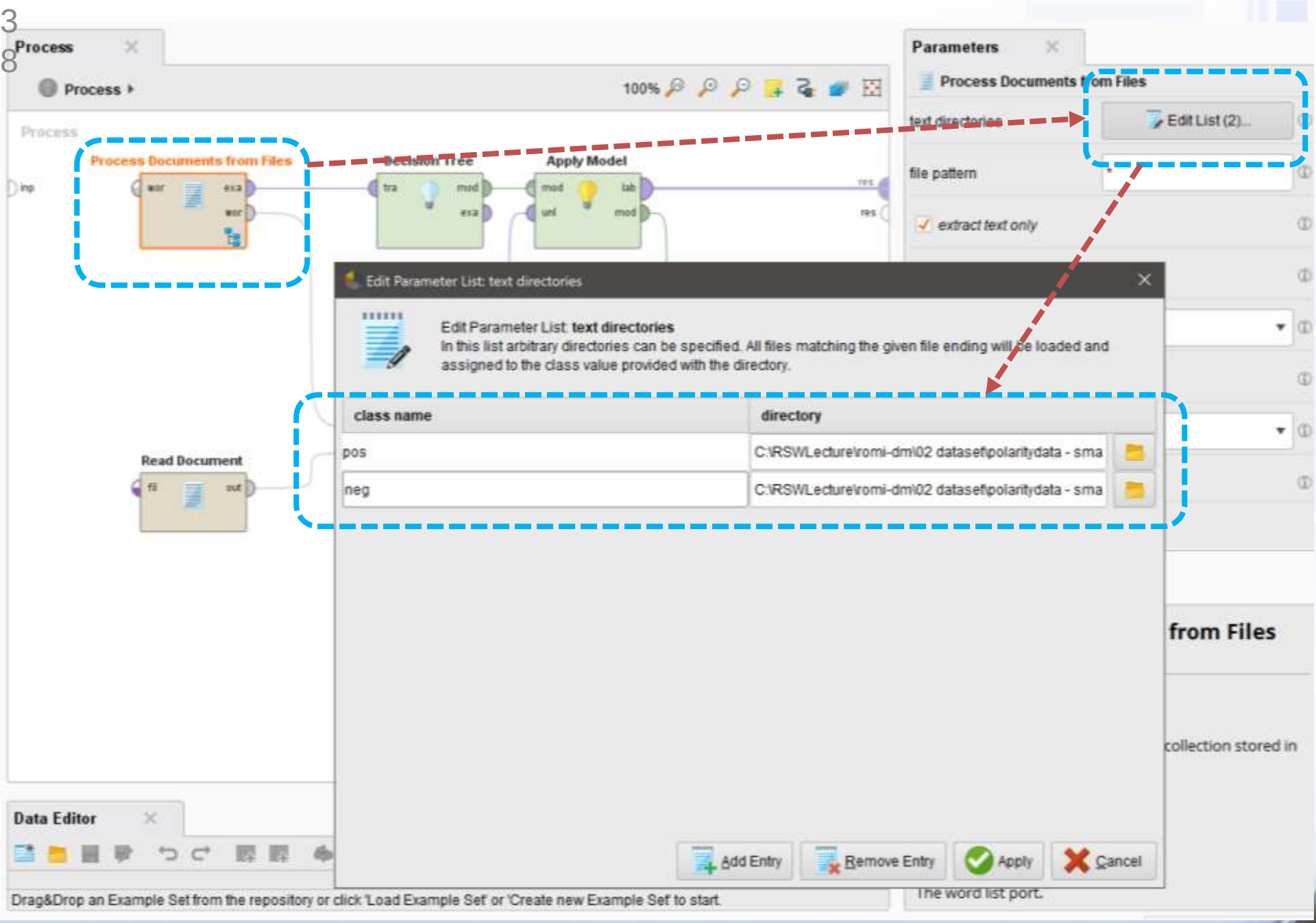


Latihan

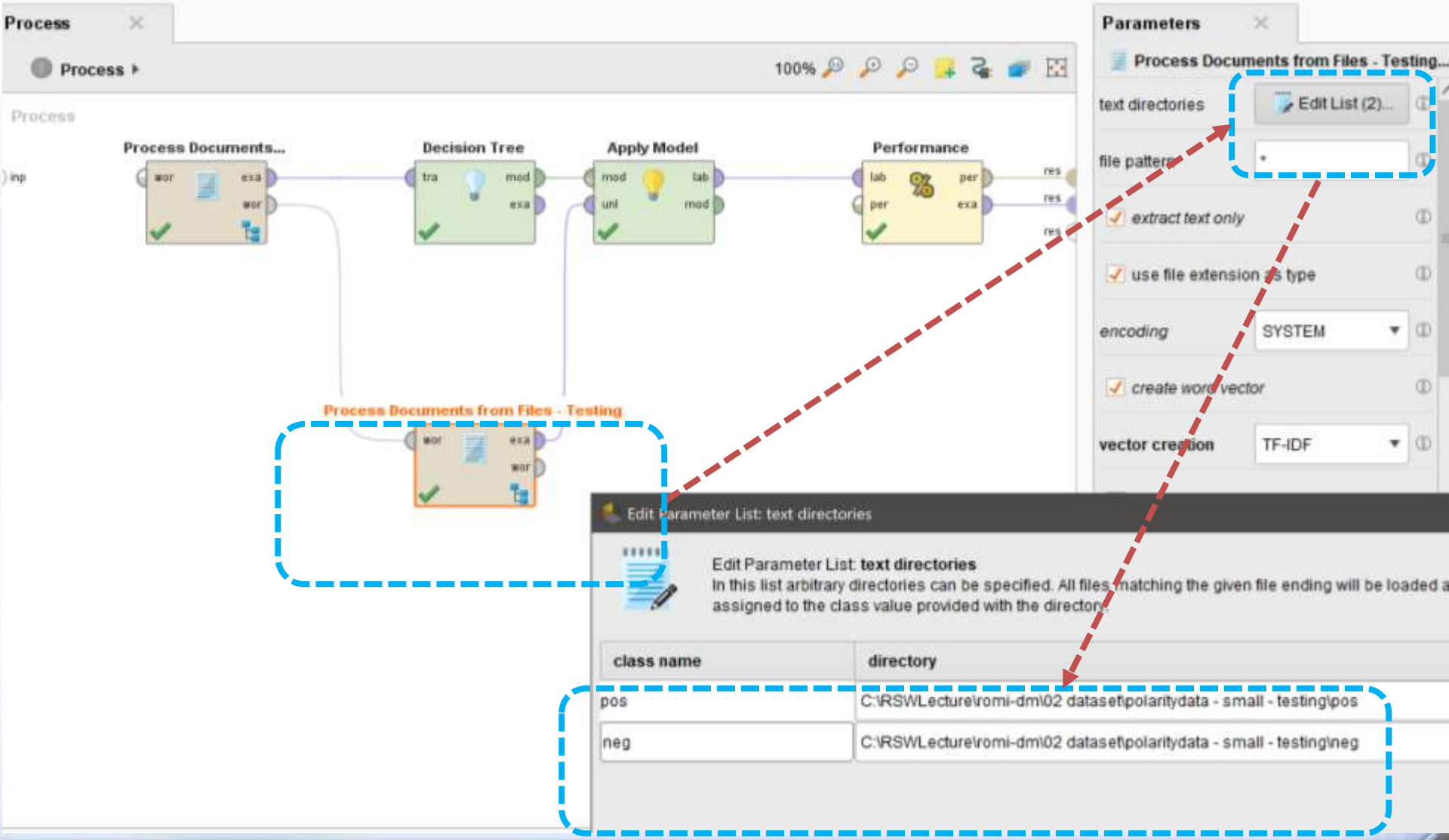
- Dengan berbagai konsep dan teknik yang anda kuasai, lakukan text classification pada dataset polarity data - small
- Gunakan algoritma Decision Tree untuk membentuk model
- Ambil 1 artikel di dalam folder polaritydata - small – testing , misalnya dalam folder pos, uji apakah artikel tersebut diprediksi termasuk sentiment negative atau positive


Process Documents from Files
wor
exa
wor
Decision Tree
tra
mod
exa
Apply Model
mod
lab
unl
mod
Process Documents
wor
exa
doc
wor
doc
Read Document
fil
out


ocuments from Files
Tokenize
doc
doc
Filter Stopwords (English)
doc
doc
Stem (Porter)
doc
doc


Process
Process
100%
Process Documents from Files
Decision Tree
Apply Model
Read Document
Parameters
Process Documents from Files
text directories
Edit List (2)...
file pattern
extract text only
Edit Parameter List: text directories
Edit Parameter List: text directories
In this list arbitrary directories can be specified. All files matching the given file ending will be loaded and assigned to the class value provided with the directory.
class name
directory
pos
C:\RSW\Lecture\romi-dm\02 dataset\polaritydata - sma
neg
C:\RSW\Lecture\romi-dm\02 dataset\polaritydata - sma
Add Entry
Remove Entry
Apply
Cancel
from Files
collection stored in
Data Editor
Drag&Drop an Example Set from the repository or click 'Load Example Set' or 'Create new Example Set' to start.
The word list port.

Ukur Akurasi dari polaritydata-small-testing



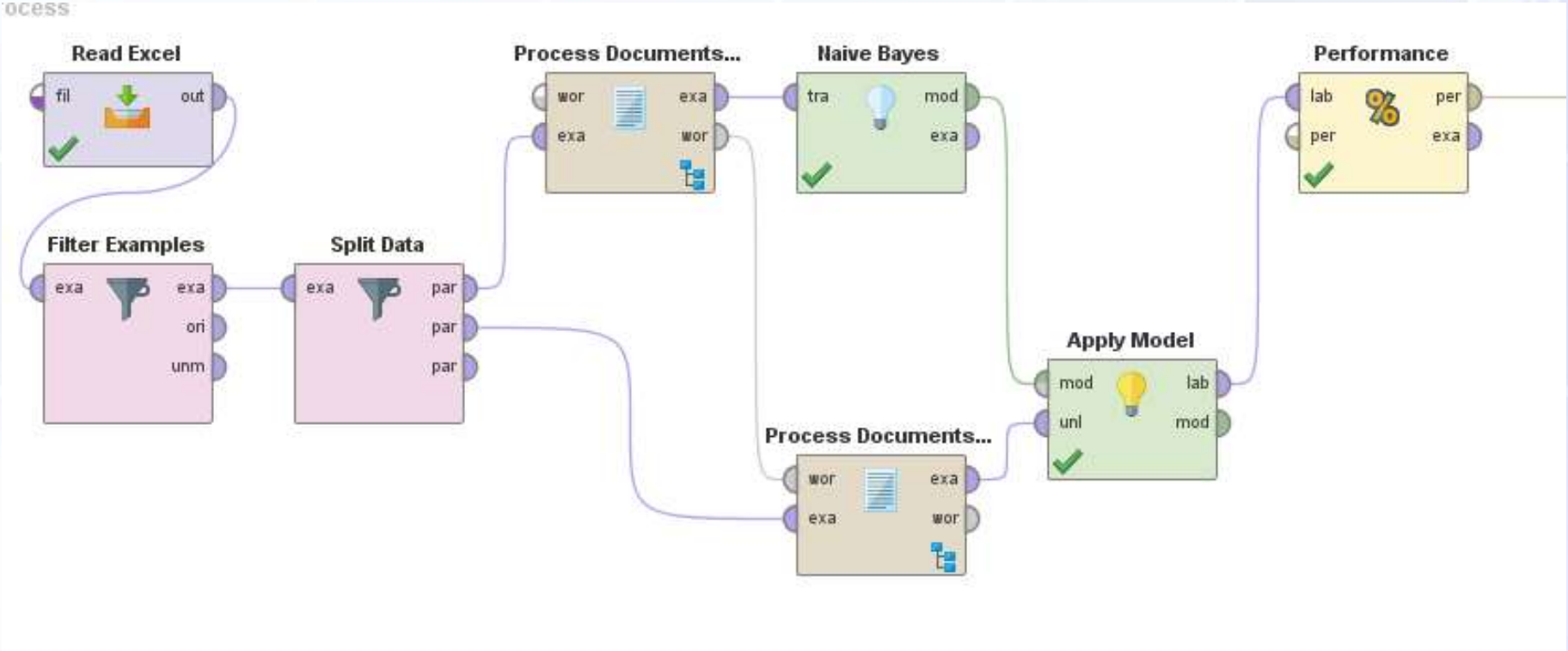
Latihan

- Dengan berbagai konsep dan teknik yang anda kuasai, lakukan text classification pada dataset polarity data
- Terapkan beberapa metode feature selection, baik filter maupun wrapper
- Lakukan komparasi terhadap berbagai algoritma klasifikasi, dan pilih yang terbaik

Latihan

- Lakukan eksperimen mengikuti buku Vijay Kotu (Predictive Analytics and Data Mining) Chapter 9 (Text Mining), Case Study 2: Predicting the Gender of Blog Authors, p 287-301
- Datasets: blog-gender-dataset.xslx
- Split Data: 50% data training dan 50% data testing
- Gunakan algoritma Naïve Bayes
- Apply model yang dihasilkan untuk data testing
- Ukur performance nya

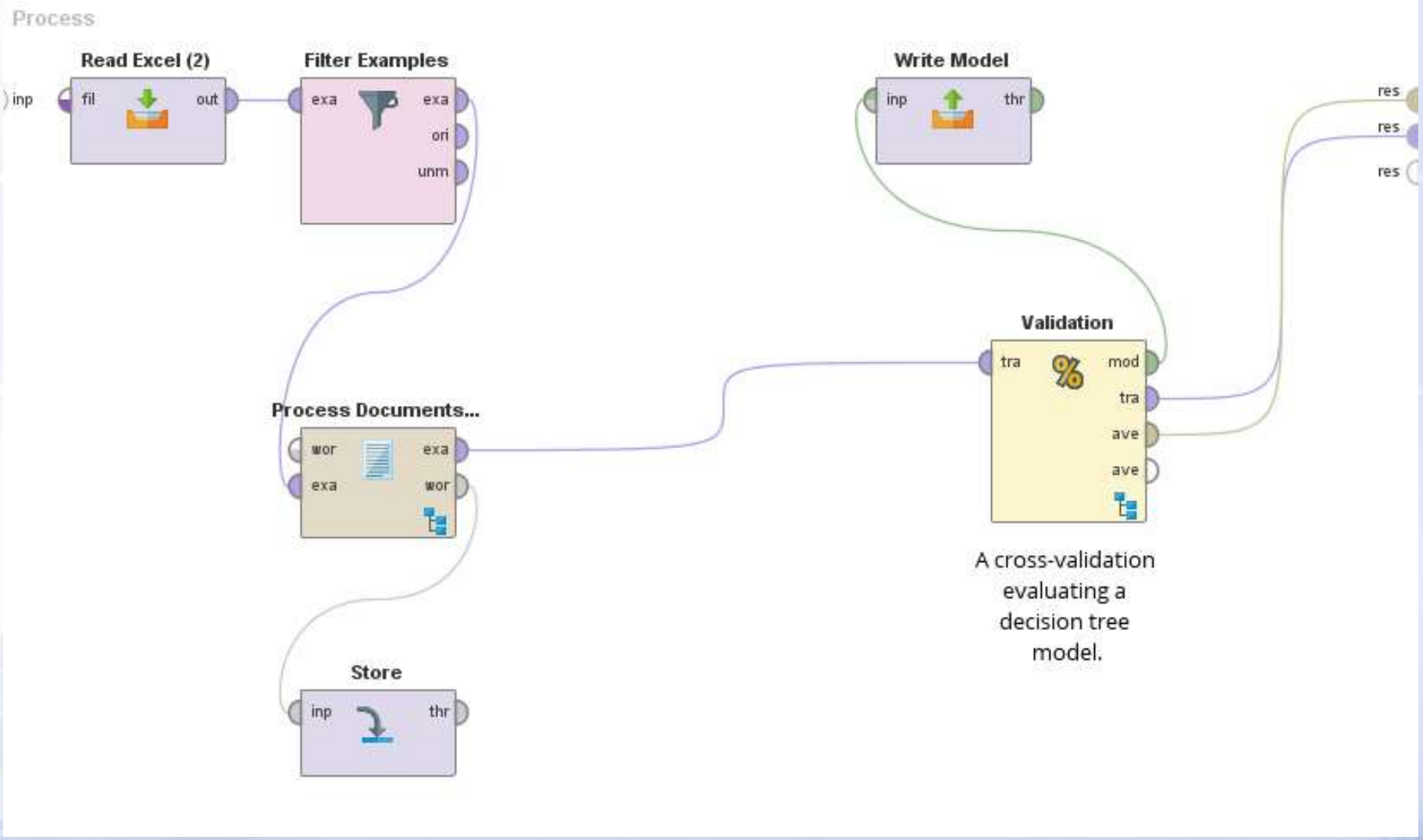
I.42


Read Excel
fil
out
Filter Examples
exa
exa
ori
unm
Split Data
exa
par
par
par
Process Documents...
wor
exa
exa
wor
Naive Bayes
tra
mod
exa
Process Documents...
wor
exa
exa
wor
Apply Model
mod
unl
lab
mod
Performance
lab
per
per
exa

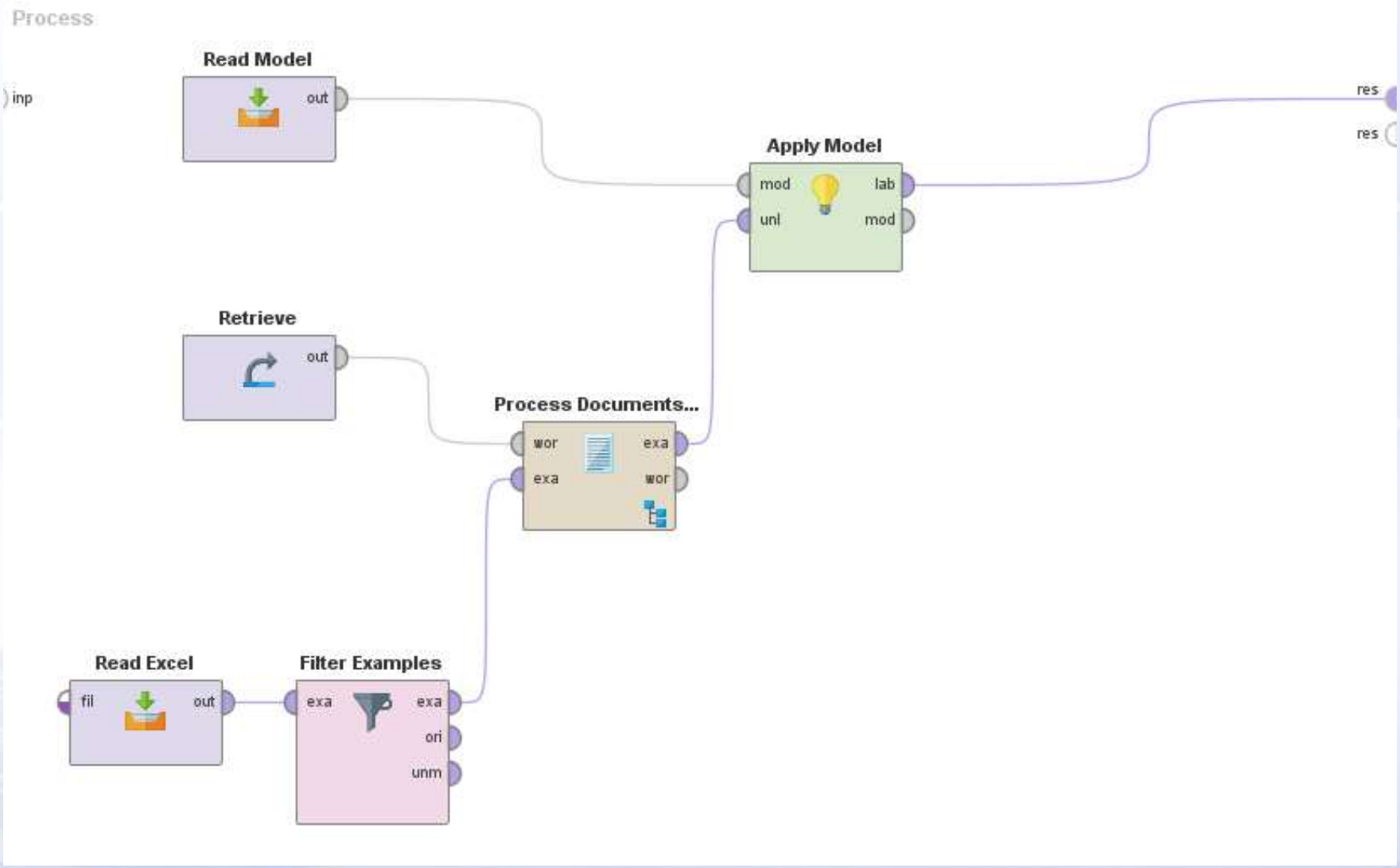
Latihan

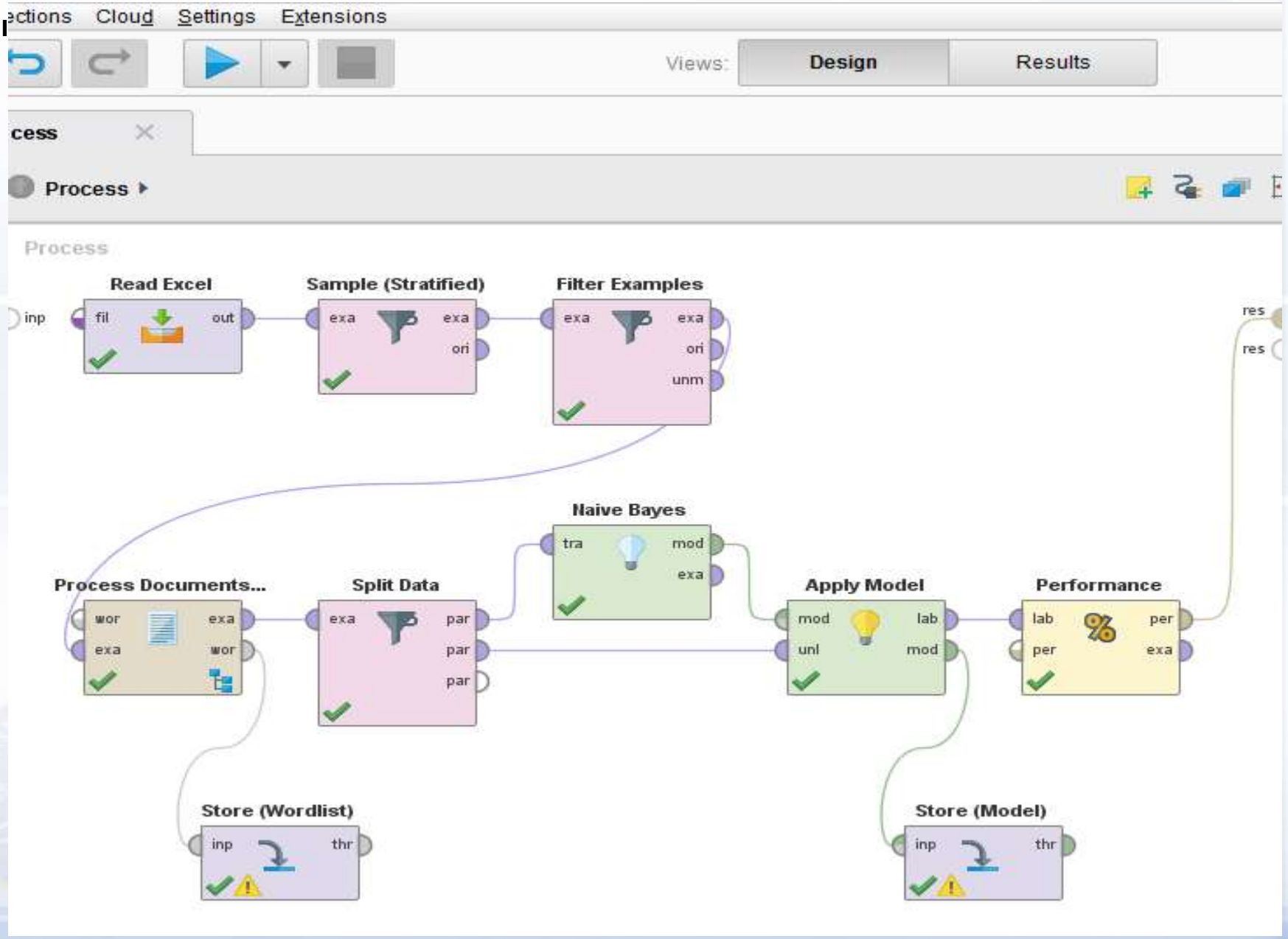
- Lakukan eksperimen mengikuti buku Vijay Kotu (Predictive Analytics and Data Mining) Chapter 9 (Text Mining), Case Study 2: Predicting the Gender of Blog Authors, p 287-301
- Datasets:
  - blog-gender-dataset.xslx
  - blog-gender-dataset-testing.xlsx
- Gunakan 10-fold X validation dan operator write model (read model), store (retrieve)

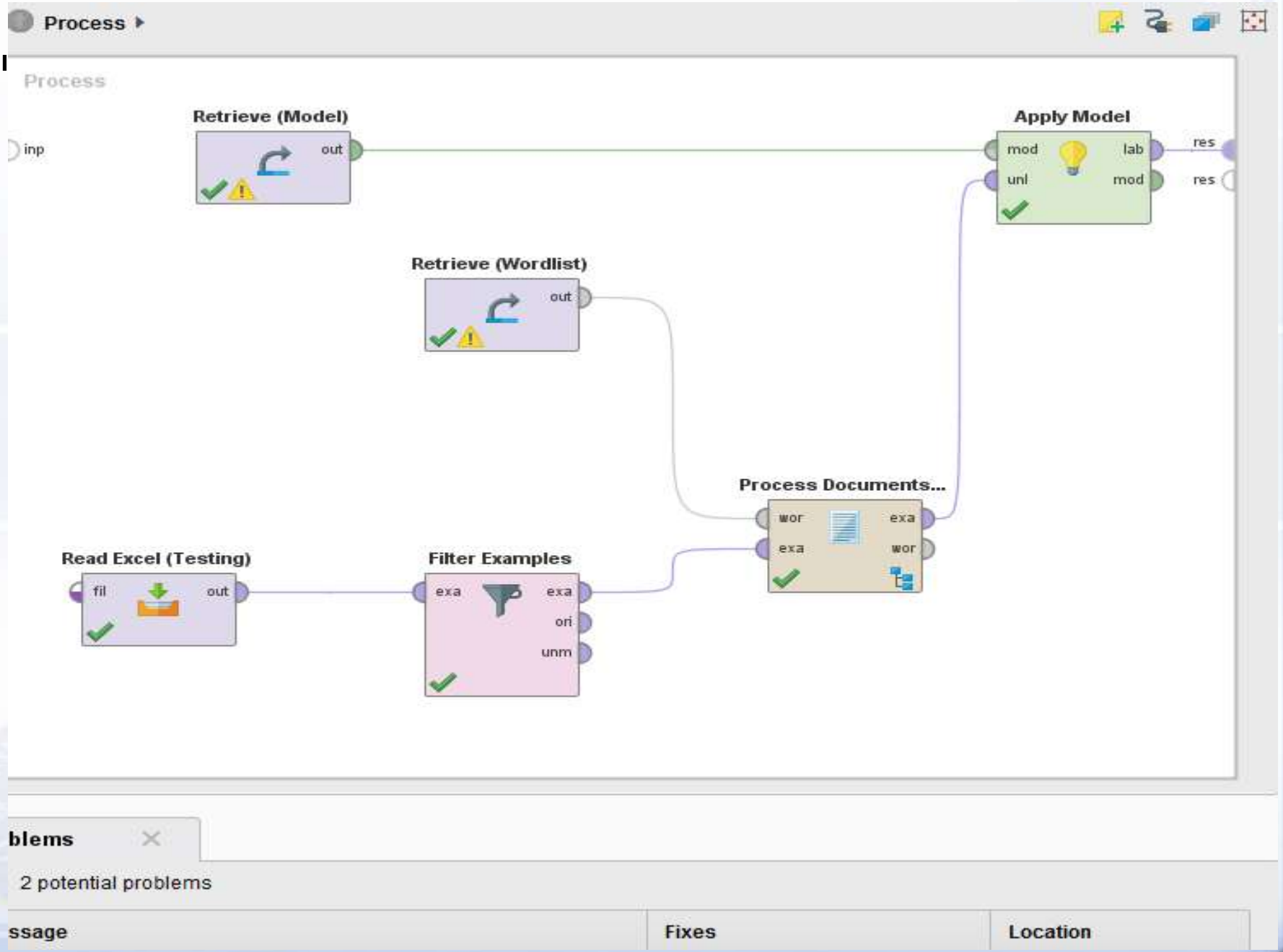
I.44


Process
Read Excel (2)
inp
fil
out
Filter Examples
exa
exa
ori
unm
Write Model
inp
thr
res
res
res
Validation
tra
mod
tra
ave
ave
Process Documents...
wor
exa
exa
wor
Store
inp
thr
A cross-validation evaluating a decision tree model.

I.45


Process
Read Model
inp
out
Apply Model
mod
lab
unl
mod
res
res
Retrieve
out
Process Documents...
wor
exa
exa
wor
Read Excel
fil
out
Filter Examples
exa
exa
ori
unm


ections Cloud Settings Extensions
Views: Design Results
cess
Process
Process
Read Excel
Sample (Stratified)
Filter Examples
Naive Bayes
Process Documents...
Split Data
Apply Model
Performance
Store (Wordlist)
Store (Model)


Process
Process
Retrieve (Model)
inp
out
Apply Model
mod
unl
lab
mod
res
res
Retrieve (Wordlist)
out
Process Documents...
wor
exa
exa
wor
Read Excel (Testing)
fil
out
Filter Examples
exa
exa
ori
unm
blems
2 potential problems
ssage
Fixes
Location

Post-Test

1. Jelaskan perbedaan antara data, informasi dan pengetahuan!
2. Jelaskan apa yang anda ketahui tentang data mining!
3. Sebutkan peran utama data mining!
4. Sebutkan pemanfaatan dari data mining di berbagai bidang!
5. Pengetahuan atau pola apa yang bisa kita dapatkan dari data di bawah?

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMAN 7	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

30.2 DATA MINING LAW

Tom Khabaza, Nine Laws of Data Mining, 2010
(*http://khabaza.codimension.net/index_files/9laws.htm*)

Data Mining Law

1. Tujuan bisnis adalah asal mula setiap solusi penambangan data
2. Pengetahuan bisnis sangat penting dalam setiap langkah proses penambangan data
3. Persiapan data lebih dari separuh proses penambangan data
4. Tidak ada makan siang gratis untuk penambang data
5. Selalu ada pola
6. Penambangan data memperkuat persepsi dalam domain bisnis
7. Prediksi meningkatkan informasi secara lokal melalui generalisasi
8. Nilai hasil data mining tidak ditentukan oleh keakuratan atau stabilitas model prediktif
9. Semua pola dapat berubah

Tom Khabaza, Nine Laws of Data Mining, 2010
(*http://khabaza.codimension.net/index_files/9laws.htm*)

1. Business Goals Law

Tujuan bisnis adalah asal mula setiap Solusi data mining

- Ini mendefinisikan bidang penambangan data: penambangan data berkaitan dengan pemecahan masalah bisnis dan mencapai tujuan bisnis
- Penambangan data pada dasarnya bukanlah sebuah teknologi; ini adalah sebuah proses, yang memiliki satu atau lebih tujuan bisnis sebagai intinya
- Tanpa tujuan bisnis, tidak ada data mining
- Pepatah: “Data Mining adalah Proses Bisnis”

2. Business Knowledge Law

Pengetahuan bisnis sangat penting dalam setiap langkah proses penambangan data

- Pembacaan CRISP-DM yang naif akan melihat pengetahuan bisnis digunakan pada awal proses dalam menentukan tujuan, dan pada akhir proses dalam memandu penerapan hasil.
- Hal ini berarti kehilangan properti utama dari proses penambangan data, bahwa pengetahuan bisnis memiliki peran sentral dalam setiap langkah

2 Business Knowledge Law

1. Pemahaman bisnis harus didasarkan pada pengetahuan bisnis, begitu pula pemetaan tujuan bisnis ke tujuan data mining
2. Pemahaman data menggunakan pengetahuan bisnis untuk memahami data mana yang terkait dengan masalah bisnis, dan bagaimana kaitannya
3. Persiapan data berarti menggunakan pengetahuan bisnis untuk membentuk data sehingga pertanyaan bisnis yang diperlukan dapat ditanyakan dan dijawab
4. Pemodelan berarti menggunakan algoritma penambangan data untuk membuat model prediktif dan menafsirkan model dan perilakunya dalam istilah bisnis – yaitu, memahami relevansi bisnisnya.
5. Evaluasi berarti memahami dampak bisnis dari penggunaan model
6. Deployment berarti menjadikan hasil data mining berfungsi dalam suatu proses bisnis

3 Data Preparation Law

Persiapan data lebih dari separuh proses penambangan data

- Maksim penambangan data: sebagian besar upaya dalam proyek penambangan data dihabiskan untuk akuisisi dan persiapan data, dan perkiraan informal bervariasi dari 50 hingga 80 persen
- Tujuan penyiapan data adalah:
- Untuk memasukkan data ke dalam bentuk di mana pertanyaan data mining dapat diajukan
- Untuk memudahkan teknik analisis (seperti algoritma data mining) dalam menjawabnya

4 No Free Lunch Theory

There is No Free Lunch for the Data Miner (NFL-DM)

The right model for a given application can only be discovered by experiment

- Axiom of machine learning: if we knew enough about a problem space, we could choose or design an algorithm to find optimal solutions in that problem space with maximal efficiency
- Arguments for the superiority of one algorithm over others in data mining rest on the idea that data mining problem spaces have one particular set of properties, or that these properties can be discovered by analysis and built into the algorithm
- However, these views arise from the erroneous idea that, in data mining, the data miner formulates the problem and the algorithm finds the solution
- In fact, the data miner both formulates the problem and finds the solution – the algorithm is merely a tool which the data miner uses to assist with certain steps in this process

4 No Free Lunch Theory

There is No Free Lunch for the Data Miner (NFL-DM)

Model yang tepat untuk aplikasi tertentu hanya dapat ditemukan melalui eksperimen

- Aksioma pembelajaran mesin: jika kita memiliki cukup pengetahuan tentang suatu ruang masalah, kita dapat memilih atau merancang suatu algoritma untuk menemukan solusi optimal dalam ruang masalah tersebut dengan efisiensi maksimal
- Argumen yang mendukung keunggulan satu algoritma dibandingkan yang lain dalam data mining bertumpu pada gagasan bahwa ruang masalah data mining memiliki satu set properti tertentu, atau bahwa properti ini dapat ditemukan melalui analisis dan dibangun ke dalam algoritma

4 No Free Lunch Theory

There is No Free Lunch for the Data Miner (NFL-DM)

Model yang tepat untuk aplikasi tertentu hanya dapat ditemukan melalui eksperimen

- Namun, pandangan ini muncul dari gagasan yang salah bahwa, dalam penambangan data, penambang data merumuskan masalahnya dan algoritma menemukan solusinya.
- Faktanya, data miner merumuskan masalah dan menemukan solusi – algoritma hanyalah sebuah alat yang digunakan data miner untuk membantu langkah-langkah tertentu dalam proses ini.

4 No Free Lunch Theory

- Jika ruang permasalahan dipahami dengan baik, proses penambangan data tidak diperlukan
  - Penambangan data adalah proses mencari koneksi yang belum diketahui
- Untuk aplikasi tertentu, tidak hanya ada satu ruang masalah
  - Model yang berbeda dapat digunakan untuk menyelesaikan bagian masalah yang berbeda
  - Cara penguraian masalah sering kali merupakan hasil penambangan data dan tidak diketahui sebelum proses dimulai
- Penambang data memanipulasi, atau “membentuk”, ruang masalah dengan persiapan data, sehingga dasar untuk mengevaluasi model terus berubah
- Tidak ada ukuran teknis mengenai nilai untuk model prediktif
- Tujuan bisnis itu sendiri mengalami revisi dan pengembangan selama proses penambangan data
    - sehingga tujuan data mining yang sesuai dapat berubah sepenuhnya

5 Watkins' Law

There are always patterns

- This law was first stated by David Watkins
- There is always something interesting to be found in a business-relevant dataset, so that even if the expected patterns were not found, something else useful would be found
- A data mining project would not be undertaken unless business experts expected that patterns would be present, and it should not be surprising that the experts are usually right

6 Insight Law

Data Mining memperkuat persepsi dalam domain bisnis

- Bagaimana data mining menghasilkan wawasan? Undang-undang ini mendekati inti dari penambangan data – mengapa ini harus merupakan proses bisnis dan bukan proses teknis
  - Masalah bisnis diselesaikan oleh manusia, bukan oleh algoritma
- Penambang data dan pakar bisnis “melihat” solusi suatu masalah, yaitu pola dalam domain yang memungkinkan tujuan bisnis tercapai
  - Jadi data Mining adalah, atau membantu sebagai bagian dari, proses persepsi
  - Algoritme penambangan data mengungkapkan pola yang biasanya tidak terlihat oleh persepsi manusia
- Proses penambangan data mengintegrasikan algoritma ini dengan proses persepsi manusia normal, yang bersifat aktif
- Dalam proses penambangan data, manusia pemecah masalah menafsirkan hasil algoritma penambangan data dan mengintegrasikannya ke dalam pemahaman bisnis mereka.

7 Prediction Law

Prediksi meningkatkan informasi secara lokal melalui generalisasi

- "Model prediktif" dan "analisis prediktif" berarti "memprediksi hasil yang paling mungkin"
- Jenis model penambangan data lainnya, seperti pengelompokan dan asosiasi, juga dicirikan sebagai "prediktif"; ini adalah pengertian istilah yang lebih longgar:
    - Model pengelompokan dapat digambarkan sebagai "memprediksi" kelompok yang akan dimasuki seseorang
    - Model asosiasi dapat digambarkan sebagai "memprediksi" satu atau lebih atribut berdasarkan atribut yang diketahui
- Apa yang dimaksud dengan "prediksi" dalam pengertian ini? Apa persamaan algoritma klasifikasi, regresi, pengelompokan dan asosiasi serta model yang dihasilkannya?
    - Jawabannya terletak pada "skoring", yaitu penerapan model prediktif pada contoh baru
    - Informasi yang tersedia tentang contoh yang dimaksud telah ditingkatkan, secara lokal, berdasarkan pola yang ditemukan oleh algoritma dan diwujudkan dalam model, yaitu berdasarkan generalisasi atau induksi.

8 Value Law

Nilai hasil data mining tidak ditentukan oleh keakuratan atau stabilitas model prediktif

- Akurasi dan stabilitas adalah ukuran yang berguna untuk mengetahui seberapa baik model prediktif membuat prediksinya
  - Akurasi berarti seberapa sering prediksi tersebut benar
  - Stabilitas berarti seberapa besar perubahan prediksi jika data yang digunakan untuk membuat model adalah sampel yang berbeda dari populasi yang sama
- Nilai model prediktif muncul dalam dua cara:
  - Prediksi model ini mendorong tindakan yang lebih baik (lebih efektif).
  - Model tersebut memberikan wawasan (pengetahuan baru) yang mengarah pada peningkatan strategi

9 Law of Change

Semua pola dapat berubah

- Pola yang ditemukan oleh data mining tidak bertahan selamanya
- Dalam aplikasi penambangan data pemasaran dan CRM, dipahami dengan baik bahwa pola perilaku pelanggan dapat berubah seiring waktu
  - Mode berubah, pasar dan persaingan berubah, dan perekonomian berubah secara keseluruhan; karena semua alasan ini, model prediktif menjadi ketinggalan jaman dan harus diperbarui secara berkala atau model tersebut tidak dapat lagi memprediksi secara akurat
  - Hal yang sama juga berlaku dalam aplikasi penambangan data yang terkait dengan risiko dan penipuan. Pola penipuan berubah seiring dengan perubahan lingkungan dan karena penjahat mengubah perilaku mereka agar tetap terdepan dalam upaya pencegahan kejahatan

Review dan Latihan

☺ END ☺