Data Process

Hendro Margono

Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human
- Regression
  - smooth by fitting the data into regression functions

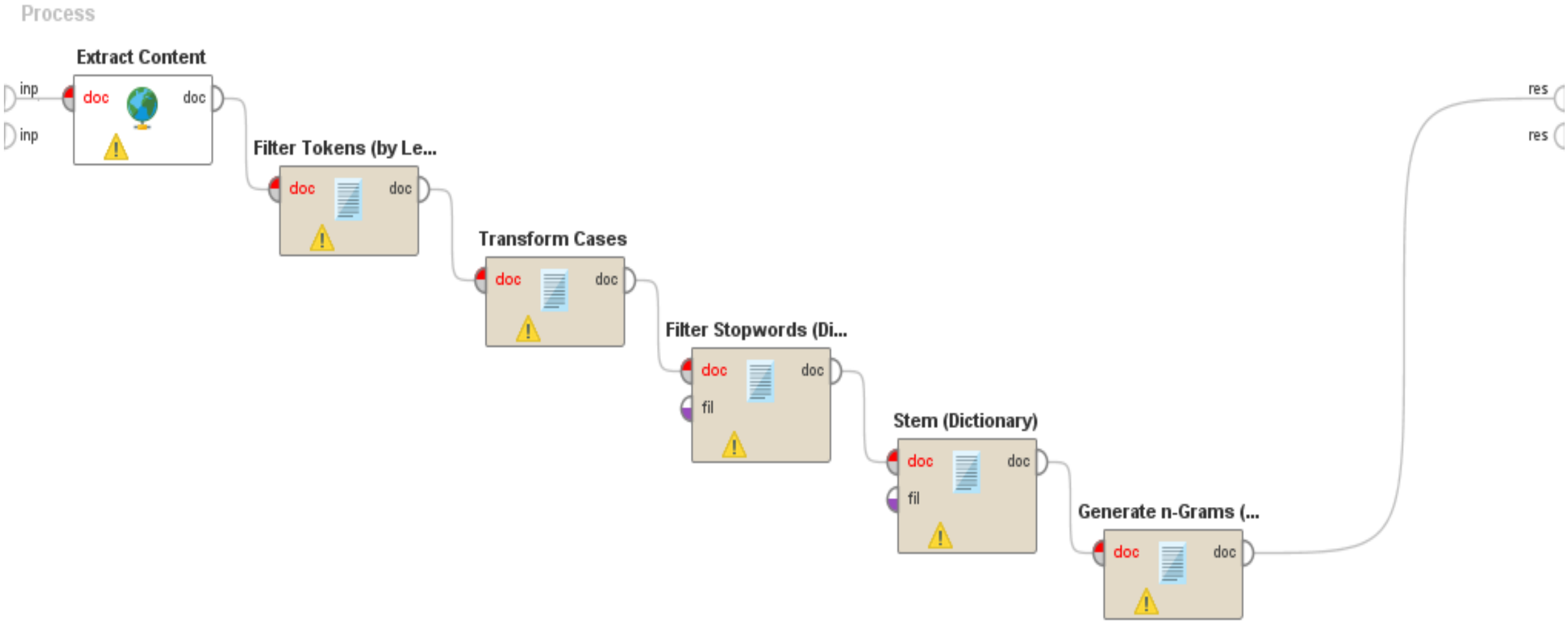
Data Integration

- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundant Data

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases. Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Cleaning Data



Tokenize

- **Tokenisasi** adalah proses untuk membagi teks yang dapat berupa kalimat, paragraf atau dokumen, menjadi token-token/bagian-bagian tertentu.
- Sebagai contoh, tokenisasi dari kalimat "Aku baru saja makan bakso pedas" menghasilkan enam token, yakni: "Aku", "baru", "saja", "makan", "bakso", "pedas". Biasanya, yang menjadi acuan pemisah antar token adalah spasi dan tanda baca·
- Biasanya, yang menjadi acuan pemisah antar token adalah spasi dan tanda baca. Tokenisasi seringkali dipakai dalam ilmu linguistik dan hasil tokenisasi berguna untuk analisis teks lebih lanjut

Transform Case

- fungsinya untuk mengkonversi karakter teks menjadi huruf besar (kapital) atau huruf kecil di awal kalimat, awal kata, atau seluruh karakter huruf.

Stop words

- Stop words adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna.
- Contoh stop words untuk bahasa Inggris diantaranya “of”, “the”. Sedangkan untuk bahasa Indonesia diantaranya “yang”, “di”, “ke”.
- menggunakan stop words untuk mengurangi jumlah kata yang harus diproses

Stemming

- adalah proses mengubah kata berimbuhan menjadi kata dasar. Aturan-aturan bahasa diterapkan untuk menanggalkan imbuhan-imbuhan itu.
- Contohnya:

  membetulkan -> betul

  berpegangan -> pegang
- Imbuhan pada Bahasa Indonesia cukup kompleks, terdiri dari:
  - Prefiks, imbuhan di depan kata: **ber**-tiga
  - Suffiks, imbuhan di akhir kata: makan-**an**
  - Konfiks, imbuhan di depan dan di akhir kata: **per**-ubah-**an**
  - Infiks, imbuhan di tengah kata: k**em**ilau.
  - Imbuhan dari bahasa asing: final-**isasi**, sosial-**isasi**
  - Aturan perubahan prefiks, seperti (me-) menjadi (meng-, mem-, men-, meny-)

N-grams

- N-gram merupakan salah satu proses yang secara luas digunakan dalam *text mining* (pengolahan teks) dan pengolahan bahasa.
- N-gram merupakan sekumpulan kata yang diberikan dalam sebuah paragraf dan ketika menghitung n-gram biasanya dilakukan dengan menggerakkan satu kata maju ke depan (Meskipun dalam prosesnya terdapat suatu proses dimana kata yang dimajukan sejumlah X kata).
- Sebagai contoh terdapat sebuah kaliman “The cow jumps over the moon”. Jika N=2 makan dikenal dengan *bigram*. Dimana ngram menjadi :
    - The cow
    - Cow jumps
    - Jumps over
    - Over the
    - The moon
- N-gram adalah potongan N-karakter yang diambil dari suatu string.