TEXT MINING

Text Mining

- Text mining mengacu pada proses ekstasi informasi dari dokumen-dokumen teks tak terstruktur (unstructur)
- Text mining dapat didefinisikan sebagai penemuan informasi baru dan tidak diketahui sebelumnya oleh computer, yang secara otomatis mengekstrak informasi dari sumber-sumber teks tak terstruktur yang berbeda.

Informati
on
Extractio
n

Text Mining VS Data Mining

- Perbedaan Mendasar terletak pada sumber data yang digunakan
- Pada data mining, pola-pola diekstrak dari basis data yang terstruktur

- sedangkan di text mining, pola-pola diekstrak dari data tekstual (natural language).
-

Tahapan Text Mining

- Text Prepocessing
- Text Trasformation
- Feature Selection
- Pattern Discovery

Text Prepocessing

- Mempersiapkan teks menjadi data yang akan mengalami pengolahan lebih lanjut.
- Operasi yang dilakukan ditahap ini adalah
  - Part of Speech (Pos) / Tagging
  - Parse tree
  - Pembersihan teks

Text Trasformation

- Pembentukan atribut mengacu pada proses untuk mendapatkan representasi dokumen yang diharapkan
- Pengubahan kata-kata kebentuk dasarnya
- Stemming
- Stop words

Feature Selection

- Memilih kata-kata yang memiliki arti penting
- mengurangi dimensi, pemilihan hanya dilakukan terhadap kata-kata yang relevan yang benar-benar merepresentasikan isi dari suatu dokumen.

Pattern Discovery

- Pattern discoverymerupakan tahap penting untuk menemukan pola atau pengetahuan (knowledge) dari keseluruhan teks.
- Tindakan yang lazim dilakukan pada tahap ini adalah operasi text mining, dan biasanya menggunakan teknik-teknik data mining.

Case Folding

- mengubah semua huruf dalam dokumen menjadi huruf kecil.
- Karakter selain huruf dihilangkan dan dianggap delimiter.

Tokenizing

- Tahap tokenizing/ parsing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.

Filtering

- Filtering adalah tahap mengambil kata-kata penting dari hasil token.
- Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words.
- Contoh stopwords adalah “yang”, “dan”, “di”, “dari”, dan seterusnya

Stemming

- Tahap stemming adalah tahap mencari root kata dari tiap kata hasil filtering.
- Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama.
- Sebagai contoh, kata bersama, kebersamaan, menyamai, akan distem ke root word nya yaitu “sama”.

Word, Token and Tokenization

- Words are separated by a special character: a blank space
- Each word is called a token
- The process of discretizing words within a document is called tokenization
- For our purpose here, each sentence can be considered a separate document, although what is considered an individual document may depend upon the context
- For now, a document here is simply a sequential collection of tokens

15

Matrix of Terms

- We can impose some form of structure on this raw data by creating a matrix, where:
  - the columns consist of all the tokens found in the two documents

- the cells of the matrix are the counts of the number of times a token appears

- Each token is now an attribute in standard data mining parlance and each document is an example

16

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

17

TF–IDF

- We could have also chosen to use the TF–IDF scores for each term to create the document vector
- N is the number of documents that we are trying to mine
- N_k is the number of documents that contain the keyword, k

Stopwords

- In the two sample text documents was the occurrence of common words such as “a,” “this,” “and,” and other similar terms
- Clearly in larger documents we would expect a larger number of such terms that do not really convey specific meaning
- Most grammatical necessities such as articles, conjunctions, prepositions, and pronouns may need to be filtered before we perform additional analysis
  - Such terms are called stopwords and usually include most articles, conjunctions, pronouns, and prepositions
  - Stopword filtering is usually the second step that follows immediately after tokenization

19

- Notice that our document vector has a significantly reduced