Text Mining: Overview, Applications and Issues

Group: 8
Course: CSE 537 Artificial Intelligence
Instructor: Professor Anita Wasilewska
Presenters: Subathra Vijaykumar 110744766
            Pooja Sonawane 110739621
Sources cited


[8] “Text mining tools techniques and applications”, Nathan Treloar, Aquaquest Inc.


[10] Common Text Mining Workflow by Ricky Ho


[12] Overview and Semantic Issues of Text Mining, Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, SIGMOD Record, September 2007 (Vol. 36, No. 3)
Overview

- Early history
- Applications
- Introduction to text mining
- Need for text mining
- Challenges in text mining
- Text mining process
- Areas of text mining
- Case study: Text Mining South Park
Brief Early History

- The challenge of exploiting the large proportion of enterprise information that originates in "unstructured” form was first recognized in IBM Journal article by H.P. Luhn[4]
- As BI emerged in the '80s and '90s as a software category, the emphasis was on numerical data stored in relational databases.
- Technological advances have enabled the field to advance during the past decade

Applications

• The technology is now broadly applied for a wide variety of government, research and business needs.
• Application categories include:
  • Security applications - monitoring online text sources
  • Biomedical applications - knowledge based search engine for biomedical texts.
  • Marketing applications - analytical customer relationship management, improve predictive analysis models, etc.
  • Sentiment analysis - analysis of data for predicting desired results
  • Software applications – used by major firms to automate analysis
  • Academic, online media, digital humanities, etc.

“Text Analytics”, Medallia, Retrieved 2015-02-23
What is Text Mining?

• It is also referred as text data mining and is roughly equivalent to text analytics.
• Process of extracting interesting and non-trivial knowledge from unstructured text.
• Text analysis also involves following:
  • information retrieval
  • lexical analysis to study word frequency distribution
  • pattern recognition
  • predictive analytics
  • visualization, etc.
• Overall goal is to turn text into data for analysis via application of Natural Language Processing and analytical methods.

“Unstructured data and 80 percent rule”, Seth Grimes; Breakthrough analysis, Retrieved 2015-02-23
Search vs Discover

Structured Data
- Data Retrieval

Unstructured Data (Text)
- Information Retrieval

Search (goal-oriented)

Discover (opportunistic)
- Data Mining
- Text Mining

“Text mining tools techniques and applications”, Nathan Treloar, Aquaquest Inc.
Data Retrieval

- It find records within a structured database.
- **Database type:** Structured
- **Search mode:** Goal-driven
- **Example Information need:** “Find a restaurant that serves vegetarian food”
Information Retrieval

- It finds a relevant information in unstructured information source.
- **Database type**: Unstructured
- **Search mode**: Goal-driven
- **Example Information need**: “Find a restaurant that serves vegetarian food”
Data Mining

- It discovers new knowledge through analysis of data.

- **Database type**: Structured

- **Search mode**: Opportunistic

- **Example Information need**: “Find the trend over number of visits to a vegetarian restaurant”
Text Mining

- It discovers new knowledge through analysis of text.

- **Database type:** Unstructured

- **Search mode:** Opportunistic

- **Example Information need:** “Find the types of food poisoning most often associated with junk food”
Need for Text Mining

- Vast amounts of new information and data are generated everyday through economic, academic and social activities, much with significant potential economic and societal value. Techniques such as text and data mining and analytics are required to exploit this potential.

- Approximately **90%** of the world’s data is held in unstructured formats (source: Oracle Corporation)

Image source: 2002, AvaQuest Inc.
Challenges in Text Mining

• Large datasets
• Noisy data
• Word Ambiguity and Context Sensitivity
  • Apple (the company) or apple (the fruit)
• Context Sensitivity
  • automobile = car = vehicle = Toyota
• Complex and subtle relationship between concepts in text
  • Eg: “AOL merges with Time-Warner” “Time-Warner is bought by AOL”
• Multilingual
Text Mining Process

- Extract Documents
- Text Transformation
- Feature Extraction
- Reduce Dimensions
- Apply standard Data Mining
- Interpretation/Evaluation
Extract Documents

- In this phase, we are extracting text document from various types of external sources into a text index (for subsequent search) as well as a text corpus (for text mining).
- Document source can be a public web site, an internal file system, etc.
How to?

- Perform a google search or crawl a predefined list of web sites, then download the web page from the list of URL, parse the DOM to extract text data from its sub-elements, and eventually creating one or multiple documents, store them into the text index as well as text Corpus.

- Invoke the Twitter API to search for tweets (or monitor a particular topic stream of tweets), store them into the text index and text Corpus.

- If the text is in a different language, we may also invoke some machine translation service (e.g. Google translate) to convert the language to English.
Text Preprocessing and Transformation

- **Tokenization**: Text documents contain a collection of statements. This step segments the whole text into words by removing blank spaces, commas etc.

- **Stop word Removal**: This step involves removing of HTML, XML tags from web pages. Then the process of removal of stop words such as ‘a’, ‘is’, ‘of’, etc is performed.

Common Text Mining Workflow by Ricky Ho
• Stemming: Stemming refers to the process of identifying the root of a certain word.
To extract information about some entities mentioned in the document, we need to conduct sentence segmentation, paragraph segmentation in order to provide some local context from which we can analyze the entity with respect to its relationship with other entities.

Attach Part-Of-Speech tagging, or Entity tagging (person, place, company) to each word.

Apply standard text processing such as lower case, removing punctuation, removing numbers, removing stop word, stemming.

Optionally, normalize the words to its synonyms using Wordnet or domain specific dictionary.

Common Text Mining Workflow by Ricky Ho
Feature Selection

● It is also known as variable selection.

● It is the process of selecting a subset of important features for use in model creation.

● This phase mainly performs removing features which are redundant or irrelevant.

● For text mining, the "bag-of-words model" is commonly used as the feature set.

● After this phase, the Corpus will turn into a large and sparse document term matrix.

Common Text Mining Workflow by Ricky Ho
Reduce Dimensions

Why?

- For efficiency reason, we want to reduce the memory footprint for storing the corpus.
- We want to transform the vector from the "term" space to a "topic" space, which allows document of similar topics to situate close by each other even they use different terms. (e.g. document using the word "pet" and "cat" are map to the same topic based on their co-occurrence)
How?

- SVD (Singular Value Decomposition) is a common matrix factorization technique to convert a "term" vector into a "concept" vector. SVD can be used to factor a large sparse matrix of M by N into the multiplication of three smaller dense matrix M*K, K*K, K*N.

- Another popular technique called topic modeling is also commonly used to transform the document into a smaller set of topic dimensions.
Text Mining Techniques

- Document Clustering
- Text Categorization
- Text Clustering
- Sentiment Analysis

A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal
Document Clustering

- It is the process of application of cluster analysis to textual data
- Two algorithms:
  - Hierarchical- Agglomerative and Decisive
  - K-means and its variant
- Applications:
  - Document Organization and browsing
  - Corpus Summarization

A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal
Text Categorization

- It aims at assigning documents to one or more classes or categories.
- Content-based classification: the weight given to particular subjects in a document determines the class to which the document is assigned.
- Applications:
  - Sentiment Analysis
  - Language identification of text

A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal
Interpretation/ Evaluation

- Analyzing the results
- Check accuracy
- Repeat the algorithm with refined data
Areas of Text Mining

- Information Extraction
- Information Retrieval
- Natural Language Processing
- Data Mining

Information Extraction

• It is the process of automatically extracting structured information from unstructured and/or semi structured text documents.
• Pattern matching is the output of the IE process.
• Functions performed by IE Systems are:
  • Term Analysis
  • Named Entity Recognition
  • Fact Extraction
Information Retrieval

- It is defined as the methods used for representation, storage and accessing of information items where the information handled is mostly in the form of textual documents, newspapers and books which are retrieved from databases according to the user request or queries.

- An IR system allows us to narrow down the set of documents that are relevant to a specific problem. The most well known IR systems are search engines such as Google.
Natural Language Processing

- The role of NLP in text mining is to provide the Information Extraction Phase with linguistic data that they need to perform their task.
- Often it is annotating documents with information like sentence boundaries, part-of-speech tagging, parsing which can then be read by the information extraction tools.
Data Mining

• Discussed in class

Case Study : Text Mining South Park

• South Park follows four fourth grade boys (Stan, Kyle, Cartman and Kenny) and an extensive ensemble cast of recurring characters.
• This analysis reviews their speech to determine which words and phrases are distinct for each character.

http://kaylinwalker.com/text-mining-south-park/
How it is done?

• The programming language R and packages tm, RWeka and stringr were used to scrape South Park episode transcripts from the internet.
• Preprocessing and transformation
• Calculate the log likelihood for each character pair, and rank them to create a list of most characteristic words/phrases for each character.
• The results were visualized using ggplot2, wordcloud and RColorBrewer.
Word cloud after initial steps
## Number of words spoke per character

<table>
<thead>
<tr>
<th>speaker</th>
<th>words</th>
<th>speaker</th>
<th>words</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>cartman</td>
<td>61110</td>
<td>jimmy</td>
<td>3738</td>
<td>narrator</td>
</tr>
<tr>
<td>stan</td>
<td>34762</td>
<td>gerald</td>
<td>3285</td>
<td>principal.victoria</td>
</tr>
<tr>
<td>kyle</td>
<td>31277</td>
<td>jimbo</td>
<td>3157</td>
<td>jesus</td>
</tr>
<tr>
<td>randy</td>
<td>14994</td>
<td>announcer</td>
<td>2900</td>
<td>mayor</td>
</tr>
<tr>
<td>butters</td>
<td>13690</td>
<td>wendy</td>
<td>2893</td>
<td>craig</td>
</tr>
<tr>
<td>mr. garrison</td>
<td>9436</td>
<td>sheila</td>
<td>2794</td>
<td>reporter</td>
</tr>
<tr>
<td>chef</td>
<td>5493</td>
<td>liane</td>
<td>2477</td>
<td>satan</td>
</tr>
<tr>
<td>mr..mackey</td>
<td>4829</td>
<td>stephen</td>
<td>2245</td>
<td>linda</td>
</tr>
<tr>
<td>sharon</td>
<td>4284</td>
<td>kenny</td>
<td>2112</td>
<td>all.others</td>
</tr>
</tbody>
</table>
Analysis of which character talks the most

Cartman Talks the Most

- Butters: 3.7%
- Randy: 4.1%
- Kyle: 8.4%
- Stan: 9.4%
- Cartman: 16.4%
- All others: 58.1%
Analysis of share of words spoken across seasons
Analysis of swear words spoken by characters

Kenny Swears at the Highest Rate

<table>
<thead>
<tr>
<th>Character</th>
<th>Profanities per 1000 Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenny</td>
<td>54.8</td>
</tr>
<tr>
<td>Cartman</td>
<td>20.8</td>
</tr>
<tr>
<td>Chef</td>
<td>16.4</td>
</tr>
<tr>
<td>Kyle</td>
<td>16.3</td>
</tr>
<tr>
<td>Mr. Garrison</td>
<td>14.6</td>
</tr>
<tr>
<td>Stan</td>
<td>13.3</td>
</tr>
<tr>
<td>Randy</td>
<td>8.9</td>
</tr>
<tr>
<td>Mr. Mackey</td>
<td>8.7</td>
</tr>
<tr>
<td>Jimmy</td>
<td>5.6</td>
</tr>
<tr>
<td>Butters</td>
<td>5</td>
</tr>
<tr>
<td>Sharon</td>
<td>4.8</td>
</tr>
</tbody>
</table>
Log-Likelihood

\[ 2 \sum O_i \times \ln \left( \frac{O_i}{E_i} \right) \]

Which can be computed from the contingency table below as

\[ (2 \times (a \times \log \left( \frac{a}{E_1} \right) + b \times \log \left( \frac{b}{E_2} \right)) ; E_1 = (a + c) \times \frac{a + b}{N} ; E_2 = (b + d) \times \frac{a + b}{N} \]
Basic Framework

<table>
<thead>
<tr>
<th>Group</th>
<th>Corpus.ONE</th>
<th>Corpus.TWO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Not Word</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>N=a+b+c+d</td>
</tr>
</tbody>
</table>

An Example with Log Likelihood 101.7

<table>
<thead>
<tr>
<th>Group</th>
<th>Cartmans.Text</th>
<th>Remaining.Text</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘hippies’</td>
<td>36</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>Not ‘hippies’</td>
<td>28170</td>
<td>144058</td>
<td>172228</td>
</tr>
<tr>
<td>Total</td>
<td>28206</td>
<td>144063</td>
<td>172269</td>
</tr>
</tbody>
</table>

Computed:

\[
E_1 = 28206 \times \left( \frac{41}{172269} \right) = 6.71
\]

\[
E_2 = 144063 \times \left( \frac{41}{172269} \right) = 34.28
\]

\[
LL = 2 \times \left( 36 \log \left( \frac{36}{6.71} \right) + 5 \times \log \left( \frac{5}{34.28} \right) \right) = 101.7
\]
Analysis of characteristic words per person

Most Characteristic Words/Phrases per Person

- Kyle: kyle, guys, butters, mom, kenny, clyde, seriously, token, sweet, balls, im gonna, scott, bitch, jews, guys im, cool, son, aw man, goddamnit, craig, fuck, suck balls, awesome
- Stan: dude, dad, cartman, yeah, come guys, chef, kyle, kenny, guys, dont care, grandpa, cartmans, killed, gonna get, wendy, gotta, poor, cool, son, bitch, aw man, goddamnit, craig, fuck, suck balls, awesome
- Kyle: cartman, dude, like, mr hankey, little brother, stan, hankey, brother, cartmans, kenny, yeah, stan dont, fat, grandpa, ill try, gotta get, wendy, gotta, poor, cool, son, bitch, aw man, goddamnit, craig, fuck, suck balls, awesome
- Kenny: woohoo, yeah, fuck, mr hankey, hey guys, uh huh, fucking, hey, fuckin, guys, guys im, huh uh, uh oh, stick, dude, giant douche, hell, bastards, cool, mom dad, stupid, bastard, chef, dont, hot, lets go, learned something, play
- Butters: oh jeez, eric, well, jeez, oh boy, mom dad, hot, hey stan, just dont, yeah, oh yeah, gerald, jersey, porn, internet, hey uh, boys, stan dont, stan im, obama, gonna go, wave, oh god, little bit
- Randy: im gonna go, nelson, hey stan, just dont, yeah, oh yeah, gerald, jersey, porn, internet, hey uh, boys, stan dont, stan im, obama, gonna go, wave, oh god, little bit
QUESTIONS?