Introduction to Text Mining

Hongning Wang
CS@UVa
What is “Text Mining”?  

• “Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text.” - wikipedia  

• “Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known.” - Hearst, 1999
Two different definitions of mining

• Goal-oriented (effectiveness driven)
  – Any process that generates useful results that are non-obvious is called “mining”.
  – Keywords: “useful” + “non-obvious”
  – Data isn’t necessarily massive

• Method-oriented (efficiency driven)
  – Any process that involves extracting information from massive data is called “mining”
  – Keywords: “massive” + “pattern”
  – Patterns aren’t necessarily useful
Knowledge discovery from text data

• IBM’s Watson wins at Jeopardy! - 2011
An overview of Watson

- On questions, at the start of question analysis
- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring
What is inside Watson?

• “Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia” – PC World

• “The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used.” – AI Magazine
What is inside Watson?

• DeepQA system
  – “Watson's main innovation was not in the creation of a new algorithm for this operation but rather its ability to **quickly** execute hundreds of proven language analysis algorithms simultaneously to find the correct answer.” – New York Times
  – [The DeepQA Research Team](#)
Text mining around us

• Sentiment analysis
Text mining around us

- Sentiment analysis
Text mining around us

- Document summarization
Text mining around us

- Document summarization
Text mining around us

• Movie recommendation

FOREIGN SUGGESTIONS (about 104)  See all >

Tell No One
Because you enjoyed:
Memento
Syriana
Children of Men

Add
Not Interested

Let the Right One In
Because you enjoyed:
Sawan Samurai
This Is Spinal Tap
The Big Lebowski

Add
Not Interested

I've Loved You So Long
Because you enjoyed:
The Queen
Syriana
Good Night, and Good Luck

Add
Not Interested

Downfall
Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai

Add
Not Interested

DRAMA SUGGESTIONS (about 82)  See all >

The Wrestler
Because you enjoyed:
Sin City
Reservoir Dogs
The Big Lebowski

Add
Not Interested

The Visitor
Because you enjoyed:
Gandhi
The Motorcycle Diaries
The Queen

Add
Not Interested

Brick
Because you enjoyed:
The Big Lebowski
Rushmore
Fight Club

Add
Not Interested

The Pianist
Because you enjoyed:
Amadeus
The Killing Fields
Empire of the Sun

Add
Not Interested
Text mining around us

• News recommendation
How to perform text mining?

- As computer scientists, we view it as
  - Text Mining = **Data Mining** + **Text Data**

  - **Natural language processing**
  - **Applied machine learning**
  - **Information retrieval**

  - **Emails**
  - **Blogs**
  - **Web pages**
  - **Software documentations**
  - **Tweets**
  - **News articles**
  - **Scientific literature**
Text mining v.s. NLP, IR, DM...

- How does it relate to data mining in general?
- How does it relate to computational linguistics?
- How does it relate to information retrieval?

<table>
<thead>
<tr>
<th></th>
<th>Finding Patterns</th>
<th>Finding “Nuggets”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Novel</td>
</tr>
<tr>
<td>Non-textual data</td>
<td>General data-mining</td>
<td>Exploratory analysis</td>
</tr>
<tr>
<td>Textual data</td>
<td>Comp Ling</td>
<td>Text Mining</td>
</tr>
</tbody>
</table>
Text mining in general

Access

Serve for IR applications

Sub-area of DM research

Mining

Discover knowledge

Filter information

Based on NLP/ML techniques

Organization

Add Structure/Annotations
Challenges in text mining

• Data collection is “free text”
  – Data is not well-organized
    • Semi-structured or unstructured
  – Natural language text contains ambiguities on many levels
    • Lexical, syntactic, semantic, and pragmatic
  – Learning techniques for processing text typically need annotated training examples
    • Expensive to acquire at scale

• What to mine?
Text mining problems we will solve

- Document categorization
  - Adding structures to the text corpus
Text mining problems we will solve

• Text clustering
  – Identifying structures in the text corpus
Text mining problems we will solve

• Topic modeling
  – Identifying structures in the text corpus

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analysis to compare known genes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a single parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Sizheng et al. from the University of California at San Diego. “But coming up with a number on an answer may be more than just a matter of counting,” notes Cecilia, who analyzed an ever-growing number of genes, “whether they are actually needed.”


From the lesson’s slides, it seems the computer analysis yields an estimate of the minimum modern and ancient genomes.
Text mining problems we will solve

• Social media and network analysis
  – Exploring additional structure in the text corpus
We will also briefly cover

• Natural language processing pipeline
  – Tokenization
    • “Studying text mining is fun!” -> “studying” + “text” + “mining” + “is” + “fun” + “!”
  – Part-of-speech tagging
    • “Studying text mining is fun!” -> Studying text mining is fun!
  – Dependency parsing
    • “Studying text mining is fun!” ->

CS@UVa  CS6501: Text Mining  22
We will also briefly cover

• Machine learning techniques
  – Supervised methods
    • Naïve Bayes, k Nearest Neighbors, Logistic Regression
  – Unsupervised methods
    • K-Means, hierarchical clustering
  – Semi-supervised methods
    • Expectation Maximization
Text mining in the era of Big Data

• Huge in size
  – Google processes 5.13B queries/day (2013)
  – Twitter receives 340M tweets/day (2012)
  – Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
  – eBay has 6.5 PB of user data + 50 TB/day

• 80% data is unstructured (IBM, 2010)

640K ought to be enough for anybody.
Scalability is crucial

- Large scale text processing techniques
  - MapReduce framework
State-of-the-art solutions

- Apache Spark ([spark.apache.org](http://spark.apache.org))
  - In-memory MapReduce
    - Specialized for machine learning algorithms
  - Speed
    - 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.
State-of-the-art solutions

• Apache Spark ([spark.apache.org](http://spark.apache.org))
  – In-memory MapReduce
    • Specialized for machine learning algorithms
  – Generality
    • Combine SQL, streaming, and complex analytics
State-of-the-art solutions

- GraphLab ([graphlab.com](http://graphlab.com))
  - Graph-based, high performance, distributed computation framework
State-of-the-art solutions

- **GraphLab** ([graphlab.com](http://graphlab.com))
  - Specialized for sparse data with local dependencies for iterative algorithms
Text mining in the era of Big Data

As knowledge consumer
Challenges:
1. Implicit feedback
2. Diverse and dynamic

Human: big data producer and consumer

As data producer
Challenges:
1. Unstructured data
2. Rich semantic

Knowledge service system

Knowledge Discovery

Human-generated data
Text data
Behavior data

Decision Support

Data Generation Modeling
Text books


What to read?

Text Mining

- Machine Learning
- Pattern Recognition
- Data Mining
- NLP
- Information Retrieval
- Web Applications, Bioinformatics...
- Library & Info Science

- Algorithms
- Statistics
- Optimization

- Text Mining
  - ICML, NIPS, UAI
  - KDD, ICDM, SDM
  - ACL, EMNLP, COLING
  - SIGIR, WWW, WSDM, CIKM

- Find more on course website for resource
Welcome to the class of “Text Mining”!