Introduction to Text Analytics
Data for Analytics

<table>
<thead>
<tr>
<th>Product</th>
<th>Name</th>
<th>Data Type</th>
<th>Nullable?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PRODUCT_ID</td>
<td>VARCHAR</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td>CATEGORY</td>
<td>VARCHAR</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td>LIST_PRICE</td>
<td>DECIMAL</td>
<td>NO</td>
</tr>
</tbody>
</table>

The challenge is:

- Approximately 75-90% of data is unstructured (while IT is built for structured data)
- Unstructured data is growing at nearly 10x the rate of structured data
- Less than 5% of unstructured data is proactively managed

Source: Natasha DeKroon and Brian Karp
Why Need Text Analytics?

“Eight-five percent of data is unstructured, and you need text analysis and text abstraction along with a relational database to arrive at an integrated view,” says Jerry Hill, vice president of manufacturing, for Teradata.

This figure has been boosted by Social Media
What Text Analytics can do?

Text Analytics Functions

- Social network analysis
- Mobility Analysis
- Information Retrieval
- Information Extraction
- Web behavior analysis
- Opinion Mining

Text Analytics Domains

- Legal
- Government
- Consumer business
- Biotech
- Financial
- Manufacturing
- Security
- Software
- Mobile
How- Text Mining Techniques

Text Mining Techniques

Classification
- SVM, decision trees, Rule-based, Neural network, Bayesian classifiers, Regression classifiers

Clustering
- Hierarchical, K-means, Distance-based, Word-based

Topic Models
- Probabilistic latent semantic analysis, Latent Dirichlet allocation, Correlated topic models

Graph Models
- Bayesian networks, HMM, Markov random fields, CRFs

Other Methods
- Chinese Restaurant Process, Pitman-yor models, NLP techniques, Linguistic models
Stages in Applying Text Analytics

**Extraction**
- Crawling
- Data processing
- Format/Representation

**Modeling**
- Generic data model
- Semantic data model
- Document modeling
- Predictive modeling

**Analytics**
- Descriptive analytics
- Predictive analytics
- Prescriptive analytics
- Decisive analytics

**Applications**
- Consumer business applications
- Financial applications
- Law applications
- Government applications
  - etc.,
Example Text Analytics Application

Users’ Feedback Channels:
Forums, Blogs, Political Sites, Facebook, Debate sites, etc.,

Sociopolitical Opinion Mining

Opinion Targets

Extraction + Classification + Normalization Methods

Sentiments

Polarized Sentiment

Suggestions

Stances

Issues

Entities

Opinion Mining of Sociopolitical Comments from Social Media, Swapna Gottipati, Thesis, 2014
Example Comment from Social Media

“My father was telling me in the past they worry abt 3 meals. The government that could give them 3 meals wins. Right now I worry abt my retirement and housing. The gov that can lower hdb cost and discuss abt job and retirement wins my vote. And I won't take 'no retirement' as an option”.

What do we see?
1. Issues – Housing, Retirement, Job
2. Entities (People/Organizations)- Government
3. Suggestive opinions – Lower HDB cost
4. Valuable Comment
## Results – Extracting Issues

<table>
<thead>
<tr>
<th>Issue</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economy</td>
<td><em>ft</em>, government, good, jobs, job, money, time, pay, working, bad, oil, country, workers, employers, simple</td>
</tr>
<tr>
<td>Immigration</td>
<td>foreign, workers, jobs, citizens, foreigners, chinese, talent, economy, immigrants, understand, world, local, foreigner</td>
</tr>
<tr>
<td>National</td>
<td>foreigners, salary, country, govt, people, ns, nsmen, vote, election, pay,</td>
</tr>
<tr>
<td>Service</td>
<td>policies, send, lower, family, private, sporeans, service</td>
</tr>
<tr>
<td>Congestion</td>
<td>time, job, change, hours, line, work, problem, place, talented, trains, people, working, foreigners, coming, run, bad</td>
</tr>
<tr>
<td>Housing</td>
<td>hdb, flats, live, foreigners, people, local, housing, property, long, high, poor, time, work, clear, afford, fw, stop, population</td>
</tr>
<tr>
<td>Education</td>
<td>good, students, school, schools, education, programme, poly, universities, work, government, academic, normal, university, overseas</td>
</tr>
</tbody>
</table>

Results - Comment Linking

Singapore Sociopolitical Issues

Response to PM Lee’s National Day speech.

1. REACH has received close to 1,700 comments on the National Day Rally (NDR) 2010 Feedback Exercise. This has surpassed the number of inputs REACH received last year on NDR by more than 50%. Immigration, Housing and Education top the list of topics with the highest number of inputs.

US Sociopolitical Issues

Response to President Obama’s State union address.

Ranked issues by the influence on public
Results - Extracting Entities and Suggestions

Input – Example comments from Social Media:

“The government should lift diplomatic immunity of the ambassador.”
“Govt must inform the romanian government of what happened.”
“SG government needs to cooperate closely with romania.”
“Hope the government help the victims by at least paying the legal fees.”
“I believe that government will help the victims for legal expenses”

Output:

<table>
<thead>
<tr>
<th>Entity</th>
<th>Suggestive opinion (Sentiment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>government</td>
<td>lift diplomatic immunity of the ambassador and get him</td>
</tr>
<tr>
<td>government</td>
<td>inform the romanian government of what happened</td>
</tr>
<tr>
<td>government</td>
<td>cooperate closely with romania</td>
</tr>
<tr>
<td>government</td>
<td>help victims by at least paying the legal fees</td>
</tr>
</tbody>
</table>

Embedded Analytics in Business Process in Healthcare

- Prediction from clinical notes using NLP
- Recommendation by Question-Answering System