Social Media and Opinion Mining
(社群媒體與意見探勘)

時間：2016/10/25 (二) (2:10-5:00pm)
地點：政治大學綜合院館270407，北棟407教室
主持人：陳恭主任

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2016-10-25
Outline

• Social Media
  – Social Media Marketing Analytics
    (社群媒體行銷分析)

• Opinion Mining
  – Text Mining and Analytics Technology
    (文字探勘分析技術)
Social Media Marketing Analytics
(社群媒體行銷分析)

Min-Yuh Day
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2016-07
Outline

• Consumer Psychology and Behavior on Social Media

• Social Media Marketing Analytics
  – Social Media Listening
  – Search Analytics
  – Content Analytics
  – Engagement Analytics

• Social Analytics Lifecycle
Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine

Emotions

Love

Joy

Surprise

Anger

Sadness

Fear

Example of Opinion: review segment on iPhone

“I bought an iPhone a few days ago.
It was such a nice phone.
The touch screen was really cool.
The voice quality was clear too.
However, my mother was mad with me as I did not tell her before I bought it.
She also thought the phone was too expensive, and wanted me to return it to the shop. ... ”

Example of Opinion: review segment on iPhone

“(1) I bought an _______ a few days ago.

(2) It was such a nice phone.

(3) The _______ screen was really cool.

(4) The _______ voice quality was clear too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

Social Media Marketing Analytics
Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Chuck Hemann and Ken Burbary, Que. 2013

Consumer Psychology and Behavior on Social Media
How consumers think, feel, and act

Analyzing Consumer Markets

• The aim of marketing is to meet and satisfy target customers’ needs and wants better than competitors.

• Marketers must have a thorough understanding of how consumers think, feel, and act and offer clear value to each and every target consumer.

Customer Perceived Value, Customer Satisfaction, and Loyalty

Social Media Marketing Analytics

- Social Media Listening
- Search Analytics
- Content Analytics
- Engagement Analytics

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
The Convergence of Paid, Owned & Earned Media

Paid Media
Traditional Ads

Owned Media
Corporate Ads

Earned Media
Organic

Promoted Brand Content

Brands that ask for shared

Sponsored Customer

Press Coverage


http://www.altimetergroup.com/2012/07/the-converged-media-imperative/
Converged Media
Top 11 Success Criteria

Social Listening / Analysis of Crowd

http://www.altimetergroup.com/2012/07/the-converged-media-imperative/
Figure 3 Content Tool Stack Hierarchy

Creation
Tools that aid in developing, building, and deploying consistent content.

Curation & Aggregation
Tools or processes that aid in the discovery, compiling, organizing, presenting, and publishing of existing content in a meaningful way that is on-brand and relevant to campaign goal.

Optimization
Tools designed for ongoing optimization of content marketing results over time.

Analytics
Independent of basic web analytics packages, content tools often contain their own specific analytics and dashboards. These can be wide ranging and are, of course, closely aligned with tool functionality.

Audience & Targeting
Tools to help identify who the target audience(s) is/are, where they are online, and the types of content that would attract them.

Distribution
Tools that help content publishers find audiences via, for example, suggested headlines or stories across publisher sites or social networks.

Workflow
Tools that aid in processes associated with content strategy including creating governance documentation (style, editing and brand guidelines), content audits, production, review, approval and publishing processes, etc.

Legal & Compliance
Tools designed for review/approval and compliance across all necessary stakeholders.

Source: Altimeter Group

Source: Rebecca Lieb, "Content marketing in 2015 -- research, not predictions", December 16, 2014
http://www.imediaconnection.com/content/37909.asp
Competitive Intelligence

- Gather competitive intelligence data

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Google Alexa Compete

• Which audience segments are competitors reaching that you are not?
• What keywords are successful for your competitors?
• What sources are driving traffic to your competitors’ websites?

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Competitive Intelligence

- Facebook competitive analysis
- Facebook content analysis
- YouTube competitive analysis
- YouTube channel analysis
- Twitter profile analysis

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Web Analytics (Clickstream)

• Content Analytics
• Mobile Analytics
Mobile Analytics

- Where is my mobile traffic coming from?
- What content are mobile users most interested in?
- How is my mobile app being used? What’s working? What isn’t?
- Which mobile platforms work best with my site?
- How does mobile user’s engagement with my site compare to traditional web users’ engagement?

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Identifying a Social Media Listening Tool

- Data Capture
- Spam Prevention
- Integration with Other Data Sources
- Cost
- Mobile Capability
- API Access
- Consistent User Interface
- Workflow Functionality
- Historical Data

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Search Analytics

• Free Tools for Collecting Insights Through Search Data
  – Google Trends
  – YouTube Trends
  – The Google AdWords Keyword Tool
  – Yahoo! Clues

• Paid Tools for Collecting Insights Through Search Data

• The BrightEdge SEO Platform

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Owned Social Metrics

- Facebook page
- Twitter account
- YouTube channel

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Own Social Media Metrics: Facebook

- Total likes
- Reach
  - Organic
  - Paid reach
  - Viral reach
- Engaged users
- People taking about this (PTAT)
- Likes, comments, and shares by post

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Own Social Media Metrics: Twitter

- Followers
- Retweets
- Replies
- Clicks and click-through rate (CTR)
- Impressions

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Own Social Media Metrics: YouTube

- Views
- Subscribers
- Likes/dislikes
- Comments
- Favorites
- Sharing

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Own Social Media Metrics: SlideShare

- Followers
- Views
- Comments
- Shares

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Own Social Media Metrics: Pinterest

• Followers
• Number of boards
• Number of pins
• Likes
• Repins
• Comments

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Own Social Media Metrics: Google+

- Number of people who have an account circled
- +1s
- Comments

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Earned Social Media Metrics

• Earned conversations
• In-network conversations

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Earned Social Media Metrics: Earned conversations

- Share of voice
- Share of conversation
- Sentiment
- Message resonance
- Overall conversation volume

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Demystifying Web Data

• Visits
• Unique page views
• Bounce rate
• Pages per visit
• Traffic sources
• Conversion

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Searching for the Right Metrics

Paid Searches

Organic Searches

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Paid Searches

• Impressions
• Clicks
• Click-through rate (CTR)
• Cost per click (CPC)
• Impression share
• Sales or revenue per click
• Average position

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Organic Searches

- Known and unknown keywords
- Known and unknown branded keywords
- Total visits
- Total conversions from known keywords
- Average search position

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Aligning Digital and Traditional Analytics

• Primary Research
  – Brand reputation
  – Message resonance
  – Executive reputation
  – Advertising performance

• Traditional Media Monitoring

• Traditional CRM Data

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Media Listening Evolution

- Location of conversations
- Sentiment
- Key message penetration
- Key influencers

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web
(blogs, social networks, forums/message boards, Video/phone sharing)

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web
(blogs, social networks, forums/message boards, Video/phone sharing)

Distill relevant signal from social noise

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover

2. Analyze

3. Segment

4. Strategy

5. Execution

Distill relevant signal from social noise

Data Segmentation
(Filter, Group, Tag, Assign)

- Strategic Planning
- Product Development
- Corps Communication
- Marketing & Advertising
- Customer Care
- Sales

Social Web
(blogs, social networks, forums/message boards, Video/phone sharing)

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web
(blogs, social networks, forums/message boards, Video/phone sharing)

Distill relevant signal from social noise

Data Segmentation
(Filter, Group, Tag, Assign)

Insights drive focused business strategies

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover

2. Analyze

Distill relevant signal from social noise

3. Segment

Data Segmentation
(Filter, Group, Tag, Assign)

4. Strategy

Insights drive focused business strategies

5. Execution

Innovation
Future Direction
Reputation Management
Campaigns
CRM
Customer Satisfaction Improvements

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web
(blogs, social networks, forums/message boards, Video/phone sharing)

Distill relevant signal from social noise

Data Segmentation
(Filter, Group, Tag, Assign)

Insights drive focused business strategies

Innovation
Future Direction
Reputation Management
Campaigns
CRM
Customer Satisfaction Improvements

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
How consumers think, feel, and act

Emotions

Love
Joy
Surprise
Anger
Sadness
Fear

Maslow’s Hierarchy of Needs

1. Physiological Needs
   (food, water, shelter)

2. Safety Needs
   (security, protection)

3. Social Needs
   (sense of belonging, love)

4. Esteem Needs
   (self-esteem, recognition, status)

5. Self-actualization Needs
   (self-development and realization)

Maslow’s hierarchy of human needs

(Maslow, 1943)

Maslow’s Hierarchy of Needs

**Physiological needs:**
food, water, warmth, rest

**Safety needs:**
security, safety

**Belongingness and love needs:**
intimate relationships, friends

**Esteem needs:**
prestige and feeling of accomplishment

**Self-actualization:**
achieving one’s full potential, including creative activities

Source: http://sixstoriesup.com/social-psyche-what-makes-us-go-social/
Social Media Hierarchy of Needs

Maslow’s Hierarchy of Needs:
- **Physiological**
  - air, sleep, food, hunger, thirst, warmth
- **Safety & Security**
  - Shelter, protection, Safety & Stability
- **Love & Belonging**
  - Love, Affection, family, & relationships
- **Esteem**
  - self-esteem, Status, Reputation
- **Self Actualization**
  - Self Actualization
- **Personal Fulfillment**
  - building a solid image. Engaging in conversations. Voicing your expertise.
- **Optimization & Monetization**

Social Media:
- **Community Building**
- **Personal Branding**
- **Structure**
- **Existence (Presence)**

Social Media Hierarchy of Needs - by John Antonios

Source: http://2.bp.blogspot.com/_Rta1VZltiMk/TPavcanFtfI/AAAAAAAAACo/OBGnRL5arSU/s1600/social-media-heirarchy-of-needs1.jpg
Social Media Hierarchy of Needs

Source: http://www.pinterest.com/pin/18647785930903585/
The Social Feedback Cycle
Consumer Behavior on Social Media

Marketer-Generated

User-Generated

Awareness  Consideration  Purchase  Use  Form Opinion  Talk

Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement
The New Customer Influence Path

Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement
Attensity: Track social sentiment across brands and competitors

http://www.attensity.com/

http://www.youtube.com/watch?v=4goxmBEg2lw#!
## Sentiment Analysis vs. Subjectivity Analysis

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Subjectivity Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Subjective</td>
</tr>
<tr>
<td>Negative</td>
<td>Objective</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
</tr>
</tbody>
</table>

- **Sentiment Analysis**
  - Positive
  - Neutral
  - Negative

- **Subjectivity Analysis**
  - Subjective
  - Objective
### Example of SentiWordNet

<table>
<thead>
<tr>
<th>POS</th>
<th>ID</th>
<th>PosScore</th>
<th>NegScore</th>
<th>SynsetTerms</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>00217728</td>
<td>0.75</td>
<td>0</td>
<td>beautiful#1</td>
<td>delighting the senses or exciting intellectual or emotional admiration; &quot;a beautiful child&quot;; &quot;beautiful country&quot;; &quot;a beautiful painting&quot;; &quot;a beautiful theory&quot;; &quot;a beautiful party&quot;</td>
</tr>
<tr>
<td>a</td>
<td>00227507</td>
<td>0.75</td>
<td>0</td>
<td>best#1</td>
<td>(superlative of `good') having the most positive qualities; &quot;the best film of the year&quot;; &quot;the best solution&quot;; &quot;the best time for planting&quot;; &quot;wore his best suit&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00042614</td>
<td>0</td>
<td>0.625</td>
<td>unhappily#2</td>
<td>sadly#1 in an unfortunate way; &quot;sadly he died before he could see his grandchild&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00093270</td>
<td>0</td>
<td>0.875</td>
<td>woefully#1</td>
<td>sadly#3 lamentably#1 deplorably#1 in an unfortunate or deplorable manner; &quot;he was sadly neglected&quot;; &quot;it was woefully inadequate&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00404501</td>
<td>0</td>
<td>0.25</td>
<td>sadly#2</td>
<td>with sadness; in a sad manner; &quot;`She died last night,' he said sadly&quot;</td>
</tr>
</tbody>
</table>
Summary

• Consumer Psychology and Behavior on Social Media
• Social Media Marketing Analytics
  – Social Media Listening
  – Search Analytics
  – Content Analytics
  – Engagement Analytics
• Social Analytics Lifecycle
References

• Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
• Dave Evans, Susan Bratton, and Jake McKee, Social Media Marketing: The Next Generation of Business Engagement, , Sybex, 2010
• Liana Evans, Social Media Marketing: Strategies for Engaging in Facebook, Twitter & Other Social Media, Que, 2010.
• Hiroshi Ishikawa, Social Big Data Mining Hardcover, CRC Press, 2015
• Data Science for Business: What you need to know about data mining and data-analytic thinking, Foster Provost and Tom Fawcett, O'Reilly, 2013
Text Mining and Analytics Technology
(文字探勘分析技術)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2016-07
Outline

• Text Mining
  – Differentiate between text mining, Web mining and data mining

• Natural Language Processing (NLP)

• Text Mining Tools and Applications
Text Mining and Analytics Technology
Text Mining Techniques
Natural Language Processing (NLP)
Text Mining

Web Mining and Social Networking

Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
Search Engines:
Information Retrieval in Practice

http://www.amazon.com/Search-Engines-Information-Retrieval-Practice/dp/0136072240
Christopher D. Manning and Hinrich Schütze (1999), Foundations of Statistical Natural Language Processing, The MIT Press

Natural Language Processing with Python
– Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

The NLTK book is currently being updated for Python 3 and NLTK 3. This is work in progress; chapters that still need to be updated are indicated. The first edition of the book, published by O’Reilly, is available at http://nltk.org/book_1ed. A second edition of the book is anticipated in early 2016.

0. Preface
1. Language Processing and Python
2. Accessing Text Corpora and Lexical Resources
3. Processing Raw Text
4. Writing Structured Programs
5. Categorizing and Tagging Words (minor fixes still required)
6. Learning to Classify Text
7. Extracting Information from Text
8. Analyzing Sentence Structure
9. Building Feature Based Grammars
10. Analyzing the Meaning of Sentences (minor fixes still required)
11. Managing Linguistic Data (minor fixes still required)
12. Afterword: Facing the Language Challenge

Bibliography
Term Index

This book is made available under the terms of the Creative Commons Attribution Noncommercial No-Derivative-Works 3.0 US License. Please post any questions about the materials to the nltk-users mailing list. Please report any errors on the issue tracker.

http://www.nltk.org/book/
Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing

http://www.amazon.com/NLTK-Essentials-Nitin-Hardeniya/dp/1784396907
Text Mining
(text data mining)

the process of deriving high-quality information from text

http://en.wikipedia.org/wiki/Text_mining
Typical Text Mining Tasks

- Text categorization
- Text clustering
- Concept/entity extraction
- Production of granular taxonomies
- Sentiment analysis
- Document summarization
- Entity relation modeling
  - i.e., learning relations between named entities.

http://en.wikipedia.org/wiki/Text_mining
Web Mining

• Web mining
  – discover useful information or knowledge from the Web hyperlink structure, page content, and usage data.

• Three types of web mining tasks
  – Web structure mining
  – Web content mining
  – Web usage mining

Text Mining Concepts

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining versus Text Mining

- Both seek for novel and useful patterns
- Both are semi-automated processes
- Difference is the nature of the data:
  - Structured versus unstructured data
  - **Structured data**: in databases
  - **Unstructured data**: Word documents, PDF files, text excerpts, XML files, and so on
- Text mining – first, impose structure to the data, then mine the structured data

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Concepts

• Benefits of text mining are obvious especially in text-rich data environments
  – e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.

• Electronic communization records (e.g., Email)
  – Spam filtering
  – Email prioritization and categorization
  – Automatic response generation

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application Area

• Information extraction
• Topic tracking
• Summarization
• Categorization
• Clustering
• Concept linking
• Question answering

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Terminology

- Unstructured or semistructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Terminology

- Term dictionary
- Word frequency
- Part-of-speech tagging (POS)
- Morphology
- Term-by-document matrix (TDM)
  - Occurrence matrix
- Singular Value Decomposition (SVD)
  - Latent Semantic Indexing (LSI)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• Structuring a collection of text
  – **Old approach**: bag-of-words
  – **New approach**: natural language processing

• NLP is ...
  – a very important concept in text mining
  – a subfield of artificial intelligence and computational linguistics
  – the studies of "understanding" the natural human language

• **Syntax** versus **semantics** based text mining

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• What is “Understanding”?  
  – Human understands, what about computers? 
  – Natural language is vague, context driven 
  – True understanding requires extensive knowledge of a topic

  – Can/will computers ever understand natural language the same/accurate way we do?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• Challenges in NLP
  – Part-of-speech tagging
  – Text segmentation
  – Word sense disambiguation
  – Syntax ambiguity
  – Imperfect or irregular input
  – Speech acts

• Dream of AI community
  – to have algorithms that are capable of automatically reading and obtaining knowledge from text

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• WordNet
  – A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  – A major resource for NLP
  – Need automation to be completed

• Sentiment Analysis
  – A technique used to detect favorable and unfavorable opinions toward specific products and services
  – CRM application

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
NLP Task Categories

- Information retrieval (IR)
- Information extraction (IE)
- Named-entity recognition (NER)
- Question answering (QA)
- Automatic summarization
- Natural language generation and understanding (NLU)
- Machine translation (ML)
- Foreign language reading and writing
- Speech recognition
- Text proofing
- Optical character recognition (OCR)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications

• Marketing applications
  – Enables better CRM

• Security applications
  – ECHELON, OASIS
  – Deception detection (...)

• Medicine and biology
  – Literature-based gene identification (...)

• Academic applications
  – Research stream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications

• Application Case: Mining for Lies
• Deception detection
  – A difficult problem
  – If detection is limited to only text, then the problem is even more difficult
• The study
  – analyzed text based testimonies of person of interests at military bases
  – used only text-based features (cues)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications

• Application Case: Mining for Lies

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications

• Application Case: Mining for Lies

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Cues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity</td>
<td>Verb count, noun-phrase count, ...</td>
</tr>
<tr>
<td>Complexity</td>
<td>Avg. no of clauses, sentence length, …</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>Modifiers, modal verbs, ...</td>
</tr>
<tr>
<td>Nonimmediacy</td>
<td>Passive voice, objectification, ...</td>
</tr>
<tr>
<td>Expressivity</td>
<td>Emotiveness</td>
</tr>
<tr>
<td>Diversity</td>
<td>Lexical diversity, redundancy, ...</td>
</tr>
<tr>
<td>Informality</td>
<td>Typographical error ratio</td>
</tr>
<tr>
<td>Specificity</td>
<td>Spatiotemporal, perceptual information …</td>
</tr>
<tr>
<td>Affect</td>
<td>Positive affect, negative affect, etc.</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications

• Application Case: Mining for Lies
  – 371 usable statements are generated
  – 31 features are used
  – Different feature selection methods used
  – 10-fold cross validation is used
  – Results (overall % accuracy)
    • Logistic regression 67.28
    • Decision trees 71.60
    • Neural networks 73.46

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications
(gene/protein interaction identification)

..expression of Bcl-2 is correlated with insufficient white blood cell death and activation of p53.

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Process

Context diagram for the text mining process

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Process

Task 1: Establish the Corpus: Collect & Organize the Domain Specific Unstructured Data

The inputs to the process include a variety of relevant unstructured (and semi-structured) data sources such as text, XML, HTML, etc.

Task 2: Create the Term-Document Matrix: Introduce Structure to the Corpus

The output of the Task 1 is a collection of documents in some digitized format for computer processing.

Task 3: Extract Knowledge: Discover Novel Patterns from the T-D Matrix

The output of the Task 2 is a flat file called term-document matrix where the cells are populated with the term frequencies.

The output of Task 3 is a number of problem specific classification, association, clustering models and visualizations.

The three-step text mining process

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Process

• **Step 1:** Establish the corpus
  – Collect all relevant unstructured data
    (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings...)
  – Digitize, standardize the collection
    (e.g., all in ASCII text files)
  – Place the collection in a common place
    (e.g., in a flat file, or in a directory as separate files)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
## Text Mining Process

**Step 2: Create the Term–by–Document Matrix**

| Terms                  | investment risk | project management | software engineering | development | SAP | ...
|------------------------|-----------------|---------------------|-----------------------|-------------|-----|-------
| Documents              |                 |                     |                       |             |     |       |
| Document 1             | 1               |                     | 1                     |             |     |       |
| Document 2             |                 | 1                   |                       |             |     |       |
| Document 3             |                 |                     | 3                     | 1           |     |       |
| Document 4             |                 |                     | 1                     |             |     |       |
| Document 5             |                 |                     | 2                     | 1           |     |       |
| Document 6             | 1               |                     | 1                     |             |     |       |
| ...                    |                 |                     |                       |             |     |       |

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Process

• **Step 2:** Create the Term–by–Document Matrix (TDM), cont.
  
  – Should all terms be included?
    • Stop words, include words
    • Synonyms, homonyms
    • Stemming
  
  – What is the best representation of the indices (values in cells)?
    • Row counts; binary frequencies; log frequencies;
    • Inverse document frequency

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Process

• Step 2: Create the Term–by–Document Matrix (TDM), cont.
  – TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    • Manual - a domain expert goes through it
    • Eliminate terms with very few occurrences in very few documents (?)
    • Transform the matrix using singular value decomposition (SVD)
    • SVD is similar to principle component analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Process

• **Step 3:** Extract patterns/knowledge
  – Classification (text categorization)
  – Clustering (natural groupings of text)
    • Improve search recall
    • Improve search precision
    • Scatter/gather
    • Query-specific clustering
  – Association
  – Trend Analysis (...)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application
(research trend identification in literature)

- Mining the published IS literature
  - MIS Quarterly (MISQ)
  - Journal of MIS (JMIS)
  - Information Systems Research (ISR)
  - Covers 12-year period (1994-2005)
  - 901 papers are included in the study
  - Only the paper abstracts are used
  - 9 clusters are generated for further analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
<table>
<thead>
<tr>
<th>Journal</th>
<th>Year</th>
<th>Author(s)</th>
<th>Title</th>
<th>Vol/No</th>
<th>Pages</th>
<th>Keywords</th>
<th>Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>MISQ</td>
<td>2005</td>
<td>A. Malhotra, S. Gosain and O. A. El Sawy</td>
<td>Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation</td>
<td>29/1</td>
<td>145-187</td>
<td>knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches</td>
<td>The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing</td>
</tr>
<tr>
<td>ISR</td>
<td>1999</td>
<td>D. Robey and M. C. Boudreau</td>
<td>Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications</td>
<td>2-Oct</td>
<td>167-185</td>
<td>organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems</td>
<td>Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory</td>
</tr>
<tr>
<td>JMIS</td>
<td>2001</td>
<td>R. Aron and E. K. Clemons</td>
<td>Achieving the optimal balance between investment in quality and investment in self-promotion for information products</td>
<td>18/2</td>
<td>65-88</td>
<td>information products internet advertising product positioning signaling signaling games</td>
<td>When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
### Text Mining Application

(research trend identification in literature)

<table>
<thead>
<tr>
<th>YEAR</th>
<th>No of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>0</td>
</tr>
<tr>
<td>1995</td>
<td>5</td>
</tr>
<tr>
<td>1996</td>
<td>10</td>
</tr>
<tr>
<td>1997</td>
<td>15</td>
</tr>
<tr>
<td>1998</td>
<td>20</td>
</tr>
<tr>
<td>1999</td>
<td>25</td>
</tr>
<tr>
<td>2000</td>
<td>30</td>
</tr>
<tr>
<td>2001</td>
<td>35</td>
</tr>
<tr>
<td>2002</td>
<td>0</td>
</tr>
<tr>
<td>2003</td>
<td>5</td>
</tr>
<tr>
<td>2004</td>
<td>10</td>
</tr>
<tr>
<td>2005</td>
<td>15</td>
</tr>
</tbody>
</table>

**Source:** Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application
(research trend identification in literature)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Tools

• Commercial Software Tools
  – SPSS PASW Text Miner
  – SAS Enterprise Miner
  – Statistica Data Miner
  – ClearForest, ...

• Free Software Tools
  – RapidMiner
  – GATE
  – Spy-EM, ...

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
SAS Text Analytics

https://www.youtube.com/watch?v=l1rYdrRCZJ4
Web Mining Overview

• Web is the largest repository of data
• Data is in HTML, XML, text format
• Challenges (of processing Web data)
  – The Web is too big for effective data mining
  – The Web is too complex
  – The Web is too dynamic
  – The Web is not specific to a domain
  – The Web has everything

• Opportunities and challenges are great!

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Mining

- Web mining (or Web data mining) is the process of discovering intrinsic relationships from Web data (textual, linkage, or usage).

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Content/Structure Mining

• Mining of the textual content on the Web
• Data collection via Web crawlers

• Web pages include hyperlinks
  – Authoritative pages
  – Hubs
  – hyperlink-induced topic search (HITS) alg

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining

• Extraction of information from data generated through Web page visits and transactions...
  – data stored in server access logs, referrer logs, agent logs, and client-side cookies
  – user characteristics and usage profiles
  – metadata, such as page attributes, content attributes, and usage data

• Clickstream data

• Clickstream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining

• Web usage mining applications
  – Determine the lifetime value of clients
  – Design cross-marketing strategies across products.
  – Evaluate promotional campaigns
  – Target electronic ads and coupons at user groups based on user access patterns
  – Predict user behavior based on previously learned rules and users' profiles
  – Present dynamic information to users based on their interests and profiles…

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining
(clickstream analysis)

User / Customer

Website

Pre-Process Data
- Collecting
- Merging
- Cleaning
- Structuring
  - Identify users
  - Identify sessions
  - Identify page views
  - Identify visits

Extract Knowledge
- Usage patterns
- User profiles
- Page profiles
- Visit profiles
- Customer value

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, ...
- Website Optimization Ecosystem

Customer Interaction on the Web → Analysis of Interactions → Knowledge about the Holistic View of the Customer

- Web Analytics
- Voice of Customer
- Customer Experience Management

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
歐巴馬是美國的一位總統

文章的文字檔
擷取未知詞過程
包含未知詞的斷詞標記結果
未知詞列表

歐巴馬(Nb) 是(SHI) 美國(Nc) 的(DE) 一(Neu) 位(Nf) 總統(Na)
抗氣候變遷 白宮籲採緊急行動

（中央社華盛頓6日綜合外電報導）白宮今天公布全球暖化對全美及美國經濟關鍵產業造成何種衝擊的新報告，呼籲採取緊急行動對抗氣候變遷。

這份為期4年的調查報告，極端氣候事件將對住家、基礎設施及產業帶來嚴重威脅。

美國總統歐巴馬2008年當選總統時曾在競選造勢時誓言，要讓美國成為對抗氣候變遷與相關「安全威脅」的領頭羊。

但歐巴馬在任上一直未能說服美國國會採取重大行動。

在本週對這項議題採取的新作為中，歐巴馬今天將與數名氣象學家接受電視訪問，討論美國全國氣候評估第3版調查結果。

美國數百名來自政府與民間的頂尖氣候科學家及技術專家，共同投入這項研究，檢視氣候變遷對當今帶來的衝擊並預測將對下個世紀帶來何種影響。

研究人員警告，加州可能發生旱災、奧克拉荷馬州發生草原大火，東岸則可能遭遇海平面上升，尤其佛羅里達，而這些事件多為人類造成。

海平面上升也將吞噬密西西比等低窪地區。

至於超過8000萬人居住且擁有全美部分成長最快都會區的東南部與加勒比海區，「海平面上升加上其他與氣候變遷有關的衝擊，以及地層下陷等既有問題，將對經濟和生態帶來重大影響」。

https://tw.news.yahoo.com/%E6%8A%97%E6%B0%A3%E5%80%99%E8%AE%8A%E9%81%B7-%E7%99%BD%E5%AE%AE%E7%B1%B2%E6%8E%A1%E7%B7%8A%E6%80%A5%E8%A1%8C%E5%8B%95-145804493.html
線上展示使用簡化詞類進行斷詞標記，僅供參考並且系統不再進行更新。線上服務斷詞和授權mirror site僅提供精簡詞類，結果也與舊版的展示系統不同。

自2014/01/06起，本斷詞系統已經處理過929136篇文章。
CKIP 中研院中文斷詞系統
http://ckipsvr.iis.sinica.edu.tw/
The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Jython), Ruby, Perl, Javascript, and F# or other .NET languages.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

All these software distributions are open source, licensed under the GNU General Public License (v2 or later). Note that this is the full GPL, which allows many free uses, but does not allow its incorporation into any type of distributed proprietary software, even in part or in translation. Commercial licensing is also available; please contact us if you are interested.

Stanford CoreNLP
An integrated suite of natural language processing tools for English and (mainland) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.

Stanford Parser
Implementations of probabilistic natural language parsers in Java: highly optimized PCFG and dependency parsers, a lexicalized PCFG parser, and a deep learning reranker. See also: Online parser demo, the Stanford Dependencies page, and Parser FAQ.

Stanford POS Tagger
A maximum-entropy (CMM) part-of-speech (POS) tagger for English,
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.

Part-of-Speech:
Stanford University is located in California. It is a great university.

**Named Entity Recognition:**

1. Stanford University is located in California.
2. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Collapsed dependencies:

1. Stanford University is located in California.

2. It is a great university.

Collapsed CC-processed dependencies:

1. Stanford University is located in California.

2. It is a great university.

Visualisation provided using the brat visualisation/annotation software.

Copyright © 2011, Stanford University, All Rights Reserved.
Stanford University is located in California. It is a great university.

Stanford CoreNLP XML Output

<table>
<thead>
<tr>
<th>Sentence #1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tokens</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

Parse tree
(ROOT (S (NP (NNP Stanford) (NNP University))) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California)))))) (. .))
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California. It is a great university.

<table>
<thead>
<tr>
<th>Id</th>
<th>Word</th>
<th>Lemma</th>
<th>Char begin</th>
<th>Char end</th>
<th>POS</th>
<th>NER</th>
<th>Normalized NER</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stanford</td>
<td>Stanford</td>
<td>0</td>
<td>8</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>2</td>
<td>University</td>
<td>University</td>
<td>9</td>
<td>19</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>be</td>
<td>20</td>
<td>22</td>
<td>VBZ</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>4</td>
<td>located</td>
<td>located</td>
<td>23</td>
<td>30</td>
<td>JJ</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>5</td>
<td>in</td>
<td>in</td>
<td>31</td>
<td>33</td>
<td>IN</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>6</td>
<td>California</td>
<td>California</td>
<td>34</td>
<td>44</td>
<td>NNP</td>
<td>LOCATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>44</td>
<td>45</td>
<td>.</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
</tbody>
</table>

Parse tree

(ROOT (S (NP (NNP Stanford) (NNP University))) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (.) (. .)))
Stanford University is located in California. It is a great university.
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California. It is a great university.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Head</th>
<th>Text</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 (gov)</td>
<td>Stanford University</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>It</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>a great university</td>
<td></td>
</tr>
</tbody>
</table>
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million.

That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares.

Related: Gates reclaims title of world's richest billionaire
Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.
It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation.
The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates Foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

Copyright © 2011, Stanford University, All Rights Reserved.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft, bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates Foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process
Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.conll.4class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates, who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
LOCATION
ORGANIZATION
PERSON
MISC
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON
Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE
Bill Gates no longer Microsoft's biggest shareholder. By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. New York (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT), Fortune 500, bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Summary

• Text Mining
  – Differentiate between text mining, Web mining and data mining

• Natural Language Processing (NLP)

• Text Mining Tools and Applications
References

• Nitin Hardeniya, NLTK Essentials, 2015, Packt Publishing
• Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, 2010, Wiley
• Guandong Xu, Yanchun Zhang, Lin Li, Web Mining and Social Networking: Techniques and Applications, 2011, Springer
• Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, O'Reilly Media
• Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2009, Springer
• Christopher D. Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, 1999, The MIT Press