



# KE5205 TEXT MINING 2018

## INTRODUCTION TO TEXT ANALYTICS

**Leong Mun Kew**  
**Institute of Systems Science**  
**National University of Singapore**

email: [munkew@nus.edu.sg](mailto:munkew@nus.edu.sg)

© 2015 National University of Singapore. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.



# Objectives of this module

**At the end of this module, you can:**

- **Describe the difference between data mining and text mining**
- **List the 5 basic use cases for text mining and provide examples relevant to real business usage**



# Outline for these modules

- **Setting the stage**
- **What is text mining?**
- **What can text mining do?**
  - The 5 Basic Use Cases of text mining
- **Workshop Assessment & Discussion**



# WHAT IS TEXT MINING

## What is Data Mining?

---

---

---

---

---

---



# What is Data Mining?

*^  
the process of*

- From Wikipedia:
  - The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
  - The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining)



# What is Data Mining?

*^  
the outcome of*

- From a business perspective:
  - Data mining is the transformation of structured **data** into **answers** to **business questions**
    - If you don't have a business context...
      - Then data mining is an academic exercise
    - If you don't have a business question...
      - Then data mining is a waste of time
    - If you don't have data...
      - Then data mining is really easy, but really useless

# What is Data Mining?

*^  
the outcome of*

- From a business perspective:
  - Data mining is the transformation of structured **data** into **answers** to **business questions**
    - If you don't have a business context...
      - Then data mining is an academic exercise
    - If you don't have a business question...
      - Then data mining is a waste of time
    - If you don't have data...
      - Then data mining is really easy, but really useless

**Data  
Mining  
Process!**



# What is Data Mining?

^  
the outcome of

- From a business perspective:
  - Data mining is the transformation of structured **data** into **answers** to **business questions**
    - If you don't have a business context...
      - Then data mining is an academic exercise
    - If you don't have a business question...
      - Then data mining is a waste of time
    - If you don't have data...
      - Then data mining is really easy, but really useless



the right  
answer

## What is the Outcome of Text Mining?

---

---

---

---

---

---



# What is the Outcome of Text Mining?

- Obvious answer?
  - Text mining is the task of transforming **unstructured text data** into answers to business questions



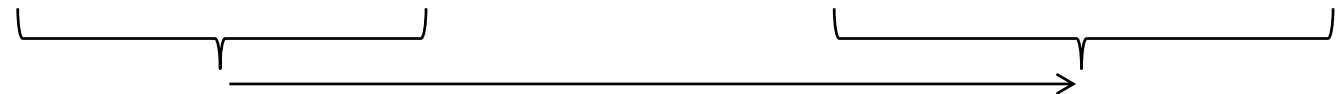
# So, what is Text Mining?

- **What do you learn in “text mining”?**
  - Text mining is the task of transforming unstructured text data into structured numerical data so that automatic algorithms can be applied to large document databases
  - Converting text to numbers requires the use of techniques for handling text at the individual word/character level to semi-structured documents to unstructured documents to document databases

# Unstructured to structured

Cust ID	Date:time	Model	Comments
00010	20121203:2201	8560	Doesn't work – may have dropped. Out of warranty. Sent to svc.
00023	20121203:1034	8850	Cannot roam. System is enabled. Reset settings on phone. Done.
00025	20121203:1640	2338	No sound. Rebooted many times. Sent to svc. 3 months old.
01003	20121203:1030	6000-1	Bought 2 weeks back. Gift. No receipt. Wants to upgrade. Sent to svc.
20456	20121203:1025	6000-1	Out of space. 4GB uSD. Set default save to uSD for songs. Done.

Cust ID	Date:time	Model	Svc	Closed	Issue	... etc
00010	20121203:2201	8560	1	1	99	
00023	20121203:1034	8850	0	1	45	
00025	20121203:1640	2338	1	1	12	
01003	20121203:1030	6000-1	1	1	99	
20456	20121203:1025	6000-1	0	1	28	



Text Mining converts  
unstructured text fields into one  
or more columns of easily  
processed numeric data



# Why is text mining so tough?

- Feature extraction is necessary (and not easy!)
  - Need background knowledge and resources
- Documents represented by very many features
  - Short fat databases
  - Features that are significant may not be intuitive
- Patterns supported by small number of documents can be significant
- Very large numbers of patterns
  - Which patterns are significant in what context and domain?
  - Training data with outcomes to prune patterns
  - Interactive exploration also useful



# WHAT CAN TEXT MINING DO?

## THE 5 BASIC USE CASES OF TEXT MINING



# Why text mining?

- **Data mining works**

- Most information in the world is not in structured data form
- The information in text needs to be unlocked
- Text is being **created in digital format** and available
  - Formal documents: word processing
  - Semi-structured text documents: patents, websites, ...
  - Informal text: email, social media, sms, tweets, ...
- Analyzing text, by itself or in conjunction with data, provides **better outcomes** for business decisions



## Text Analytics Software, What Is It and Why is It Worth \$1.8 Billion?

InternetTimeMachine



Subscribe

127 videos ▾



Like



Share



35,744

Uploaded by [InternetTimeMachine](#) on Sep 24, 2010

<http://www.TheInternetTimeMachine.com> looks at text analytics software and why IBM bought Netezza for \$1.7 billion recently. Text mining or text digging has been around for years, so why is it so valuable now? Check out this video on data mining to see..

46 likes, 1 dislike

From: <https://www.youtube.com/watch?v=soFQT5RAdMk>



# What can text mining do?

## 5 basic Use Cases:

1. **Extract “meaning” from unstructured text**
2. **Automatically put text into categories**
3. **Improve accuracy in predictive modeling or unsupervised learning**
4. **Identify specific or similar/relevant documents**
5. **Extract specific information from the text**



# 1. Extract “meaning” from unstructured text

- **Extract answers from large corpus of small documents or small corpus or large documents that is not doable by human eye**
- **Sentiment analysis**
  - What are my customers saying about me?
  - What are the areas of concern to a target group?
  - Analyzing open-ended responses to survey questions
- **Trending themes in a stream of text**
  - Insurance claims trends, warranty claims analysis
- **Summarizing text**
  - Gisting – main theme of text documents/websites
  - Automatic keyword extraction

# Overview -- Analyzing Twitter data with IBM BigSheets

IBMetinfo



Subscribe

28 videos

Curt Hall	Curtsiphone	Sat Sep 11 17:08:03 +0000 2010
Martin Richard	Marlen1929	Fri Sep 10 19:55:06 +0000 2010
????? Bieber	RachSmiles4JB	Wed Sep 15 22:34:43 +0000 2010
Curt Hall	Curtsiphone	Sat Sep 11 17:08:04 +0000 2010
KickPost	KickPost	Fri Sep 10 19:55:06 +0000 2010
Leonidas Koustimpis	leonbis2000	Wed Sep 15 22:34:43 +0000 2010
Curt Hall	Curtsiphone	Sat Sep 11 17:08:03 +0000 2010
Black.Mamba	MsLadyJoycelynn	Fri Sep 10 19:55:06 +0000 2010
Jennifer ?	jenn4sgb13	Wed Sep 15 22:34:43 +0000 2010
Tweets España	espana_es	Sat Sep 11 17:08:03 +0000 2010
Hairulnizar	rullysmully	Fri Sep 10 19:55:06 +0000 2010
J.As	R_Angel_9	Wed Sep 15 22:34:43 +0000 2010
iPhone?????? ??	iphone_akashi	Sat Sep 11 17:08:03 +0000 2010



24 Hour Twitte... 36



Like



Share



2,193



Uploaded by [IBMetinfo](#) on Oct 31, 2010

This demonstration shows how IBM BigSheets can be used to find buyer sentiment in Twitter data. This is a shortened version of the demo. You can see the full step-by-step version at <http://www.youtube.com/watch?v=Jqq66INIQUU>

8 likes, 0 dislikes

From: <http://www.youtube.com/watch?v=PSq7hZ0shLs>



# Things to Note

**This does the hard work**

**New Sheet: Macro**

Sheet Name:

**LW - Sentiment Analysis**

This is a Languageware UDF for sentiment analysis.

Fill in parameters:

content\*

type\*  
  
com.ibm.Watchlist  
com.ibm.NegativeIndicator  
com.ibm.PositiveIndicator  
com.ibm.QuestionIndicator  
com.ibm.TwitterID  
com.ibm.URL

Parameters

☐ ☐

A	B	created_at
id	name	
628	????	0 19:55:06 +0000 2010
029	waldheins	15 22:34:43 +0000 2010
126	Curt Hall	11 17:08:03 +0000 2010
352	Alan Phillips	0 19:55:05 +0000 2010
165	Bryan Hammond	15 22:34:43 +0000 2010
355	Curt Hall	11 17:08:03 +0000 2010
765	Martin Richard	0 19:55:06 +0000 2010
394	?????? Bieber	15 22:34:43 +0000 2010
624	Curt Hall	11 17:08:04 +0000 2010
155	KickPost	0 19:55:06 +0000 2010
709	Leonidas Koustimpis	15 22:34:45 +0000 2010
855	Curt Hall	11 17:08:04 +0000 2010
244	Black.Mamba	0 19:55:06 +0000 2010
290	Jennifer ?	15 22:34:46 +0000 2010
679	Tweets Espa a	11 17:08:04 +0000 2010
341	Hairulnizar	0 19:55:06 +0000 2010
785	J.As	Wed Sep 15 22:34:46 +0000 2010

R Angel 9

# Things to Note

Result Data:		Ready		
	type	name	screen_name	
1	com.ibm.en.PositiveIndicator	Special	Marlen1929	@m
2	com.ibm.en.PositiveIndicator	fantastic	Curtsiphone	Mak
3	com.ibm.en.PositiveIndicator	glorious	sharding	@Ke
4	com.ibm.en.PositiveIndicator	Amazing	thaibisz	Top
5	com.ibm.en.PositiveIndicator	like	MagsJB	My a
6	com.ibm.en.PositiveIndicator	Cool	Cellphonez	Ipho
7	com.ibm.en.PositiveIndicator	best	THE_Efram	@HI
8	com.ibm.en.PositiveIndicator	Quick	Leesa19043	@ilo
9	com.ibm.en.PositiveIndicator	First	paladigarisbiz	(HO
10	com.ibm.en.PositiveIndicator	wow	sjbuchanan007	@Ar
11	com.ibm.en.PositiveIndicator	Fast	Leesa19043	@ilo
12	com.ibm.en.PositiveIndicator	great	grattonboy	I'm
13	com.ibm.en.NegativeIndicator	doubt	gadgetinn	App
14	com.ibm.en.NegativeIndicator	hurt	mslurvmylife	Girl
15	com.ibm.en.PositiveIndicator	like	POSTMODERNISM_	Rec
16	com.ibm.en.PositiveIndicator	like	gadgetinn	App

**Decides  
sentiment  
based on  
lists of  
built-in  
keywords**



**What was the Business Context?**

---

**What was the Business Question/Need?**

---

**What was the data that was used?**

---

**What was the answer that was obtained?**

---

**What advantage did text mining provide in this case?**

---

**Exercise**





## 2. Automatically put text into categories

- **Classification – assigning one or more predefined categories to a text document, for subsequent processing**
- **Automatic actions based on category**
  - Email routing, spam filtering
  - News filtering
- **Identifying anomalies based on text descriptions**
  - Fraud detection, normally flag for human intervention





## Predicting Article Subject Matter

- Project Goal
  - To automatically classify articles as either related to financial earning or not
- Project Plan
  - Using 5,000 expertly classified articles from Reuters, index the text and build predictive models that will classify new articles

### Clear Business Objective

Shoppers continue the drought since termpomo, althou Comersaria Smith barpmo will be more 135,221 bag 5.81 mln against delivered earlier Comersaria Smith cocoa is still avail total Bahia crop a 6.2 mln there are mid-Bernen, espe the cocoa would difficulties in obtain quality over rece on consignment, created per annu nearly shipment 1,738 to 1,788 de also light and all and at 31 and 45 1,888 disc per ton sold at 4,184, 4,3 New York May, Ju 8, 458 disc and ol dms and 2.27 time the U.S., Comerble more registered a dms for Aug and U.S., Argentina, U bated with Harl dms and at 1.25 to times New York 5 Smith said, Total against the 1888. Final figures for 0 the Brazilian Coc February 27, Mar

Standard Oil Co as to manage the re companies. BP is called BP/Stand under the oversig

Texas Commerce t filed an application create the largest network would led dms in deposits. Bt

BankAmerica Corp is not under pressure to act quickly on its proposed equity offering and would do well to delay it because of the stock's recent poor performance, banking analysts said. Some analysts said they have recommended BankAmerica delay its up to one-billion-dollar equity offering, which has yet to be approved by the Securities and Exchange Commission. BankAmerica's stock fell this week, along with other banking issues, on the news that Brazil has suspended interest payments on a large portion of its foreign debt. The stock traded around 12, down 1/4, this afternoon, after falling to 11 1/2 earlier this week on the news. Banking analysts said that with the immediate threat of the First Interstate Bancorp CEO takeover bid gone, BankAmerica is under no pressure to sell the securities into a market that will be nervous on bank stocks in the near term. BankAmerica filed the offer on January 26. It was seen as one of the major factors leading the First Interstate withdrawing its takeover bid on February 6. A BankAmerica spokesman said SEC approval is taking longer than expected and market conditions must now be re-evaluated. "The circumstances at the time will determine what we do," said Arthur Miller, BankAmerica's Vice President for Financial Communications, when asked if BankAmerica would proceed with the offer immediately after it receives SEC approval. "I'd put it off as long as they conceivably could," said Lawrence Cohen, analyst with Herrell Lynch, Pierce, Fennner and Smith. Cohen said the longer BankAmerica waits, the longer they have to show the market an improved financial outlook. Although BankAmerica has yet to specify the types of equities it would offer, most analysts believed a convertible preferred stock would encompass at least part of it. Such an offering at a depressed stock price would mean a lower conversion price and more dilution to BankAmerica's stock holders, noted Daniel Williams, analyst with Salp Group. Several analysts said that while they believe the Brazilian debt problem will continue to hang over the banking industry through the quarter, the initial shock reaction is likely to ease over the coming weeks. Nevertheless, BankAmerica, which holds about 2.78 billion dms in Brazilian loans, stands to lose 15-20 mln dms if the interest rate is reduced on the debt, and as much as 200 mln dms if Brazil pays no interest for a year, said Joseph Aronson, analyst with Rye, Wilson and Co. He noted, however, that any potential losses would not show up in the current quarter. With other major banks standing to lose even more than BankAmerica if Brazil fails to service its debt, the analysts said they expect the debt will be restructured, similar to way Mexico's debt was, minimizing losses to the creditor banks. Reuter

Data has "ground truth" established by experts



**What was the Business Context?**

---

---

**What was the Business Question/Need?**

---

---

**What was the data that was used?**

---

---

**What was the answer that was obtained?**

---

---

**What advantage did text mining provide in this case?**

---

---

**Exercise**



### 3. Improve predictive accuracy in predictive modeling or unsupervised learning

- **Use text mining to improve data mining results (“Lift”)**
- **Changing text to numbers to work with data mining**
  - Build a data matrix based on word/phrase counts
  - Compute various indices based on those matrices
  - Merge indices, counts with structured data for mining
- **Predicting insurance fraud from claims processing notes**
- **Using dictionaries to control vocabulary, reduce variance**

# Text Mining Series: Predicting Fraudulent Claims

StatSoft



Subscribe

88 videos ▾



1,165

Uploaded by [StatSoft](#) on Nov 15, 2011

In this case study, fraud detection models are built using the structured variables and provide a good predictive model, finding fraudulent claims. Then with the aid of STATISTICA Text Miner, the notes for each claim were indexed and the results were added to the predictive variable pool. Predictive models built with the added text mining results gave a 10% improvement in finding fraudulent claims.

5 likes, 0 dislikes

From: <http://www.youtube.com/watch?v=OlQpm8qTog4>



## Predicting Fraudulent Claims

- Can predictability of fraudulent claims be improved by adding Text Mining results?
- Variables for analysis include
  - Delay in report
  - Policy age
  - Unusual injury
  - Minor injury
  - Text notes

Data: Fraud.sta (11v by 2732c)

	1 Fraud	2 Delay _in_ Report	3 Date _of_ Loss	4 Vehicle _Year	5 Age _In_ Days	6 Vehicle _Age	7 Unusual _Injury	8 Minor Injury	9 Text Notes	10 Cause	11 stratified sample
1	Yes	5	1/17/07	2010	1510	1	1	0	Car had right-of-way 'B' cited but 'A' w/ train		
2	Yes	4	3/1/05	2009	2196	2	1	0	Car had right-of-way Accident per inner train		
3	No	0	7/23/09	2006	591.1	5	0	0	Left turn other than Attempting to pas. exclude		
4	No	0	4/10/09	1998	695.1	13	0	0	Left turn other than Attempting to pas. exclude		
5	Yes	0	4/26/06	2009	1775	2	1	0	Car had right-of-way Attempting to pas train		
6	Yes	1	12/3/04	1995	2285	16	1	1	Car had right-of-way Attempting to pas train		
7	Yes	2	2/23/06	2008	1838	3	0	1	Hit by other car - nc Attempting to pas train		

Data: fraud with TM results.sta\* (39v by 2732c)

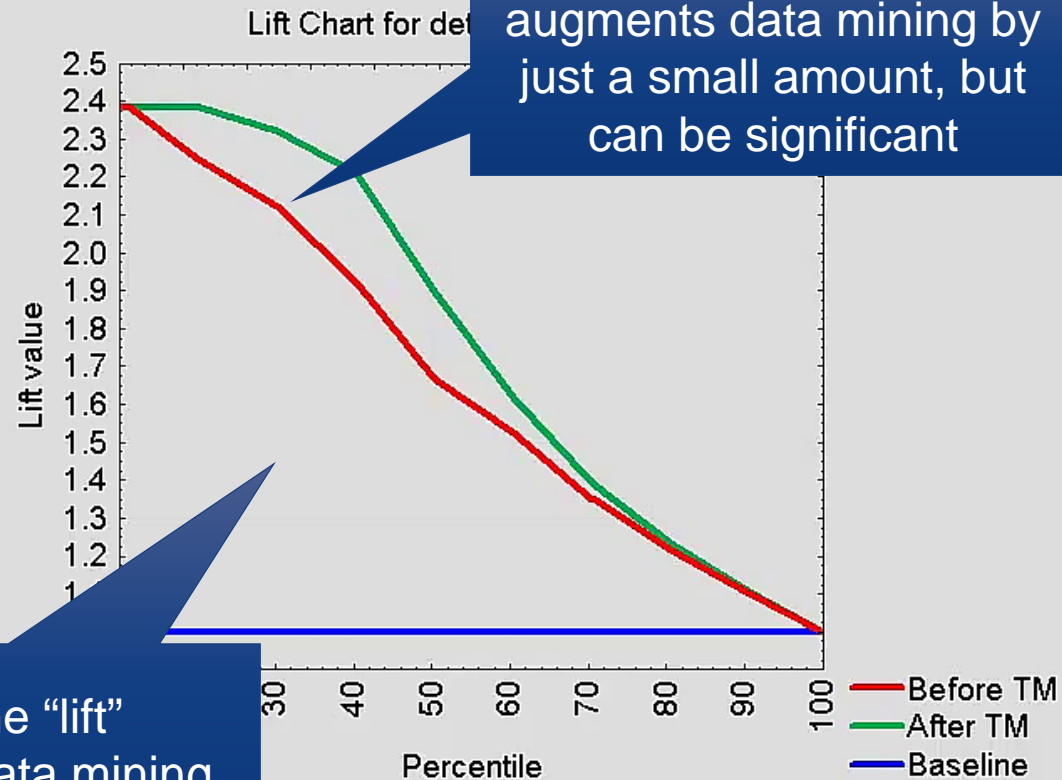
Text Miner results: may be used as Input spreadsheet for subsequent analyses

	1 Fraud	2 Delay _in_ Report	3 Age _In_ Days	4 Unusual _Injury	5 Minor Injury	6 stratified sample	7 appli	8 better	9 call	10 car	11 client	12 descript	13 hit	14 injury	15 intersect	16 lawyer	17 legal	18 move	19 neck	20 park	21 rear	22 right	23 right-of- way	24 stop	25 way
1	Yes	5	1509.6	1	0	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
2	Yes	4	2196.03	1	0	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
3	No	0	591.108	0	0	exclude																			
4	No	0	695.129	0	0	exclude																			
5	Yes	0	1775.05	1	0	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
6	Yes	1	2284.79	1	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
7	Yes	2	1837.84	0	1	train	4.357	4.357		1.465		4.30187	0.85												
8	Yes	3	880.781	0	0	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
9	Yes	0	376.028	1	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
10	Yes	4	339.661	1	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
11	No	0	684.083	0	0	exclude							0.85				0.902				0.86			0.9	
12	No	0	408.183	0	0	test			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
13	Yes	0	1922	1	0	test			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
14	Yes	1	2560.34	0	0	test			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
15	Yes	7	2043.96	1	0	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
16	Yes	0	1463.17	1	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
17	Yes	2	1818.81	1	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
18	No	0	998.468	0	0	exclude			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
19	No	6	573.433	0	1	train							0.85				0.902			0.86				0.9	
20	No	2	571.677	0	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
21	Yes	0	2486.18	1	1	test			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		
22	Yes	2	1494.84	1	1	train			1.596	1.465	1.596			1.596		1.5938		1.596					1.1363		

Text converted into numbers

## Predicting Fraudulent Claims – Project Steps

1. Build predictive models using the structured data
2. Index the text notes for accident claims
3. Build predictive models using the structured data and text mining results
4. Compare model performance





**What was the Business Context?**

---

---

**What was the Business Question/Need?**

---

---

**What was the data that was used?**

---

---

**What was the answer that was obtained?**

---

---

**What advantage did text mining provide in this case?**

---

---

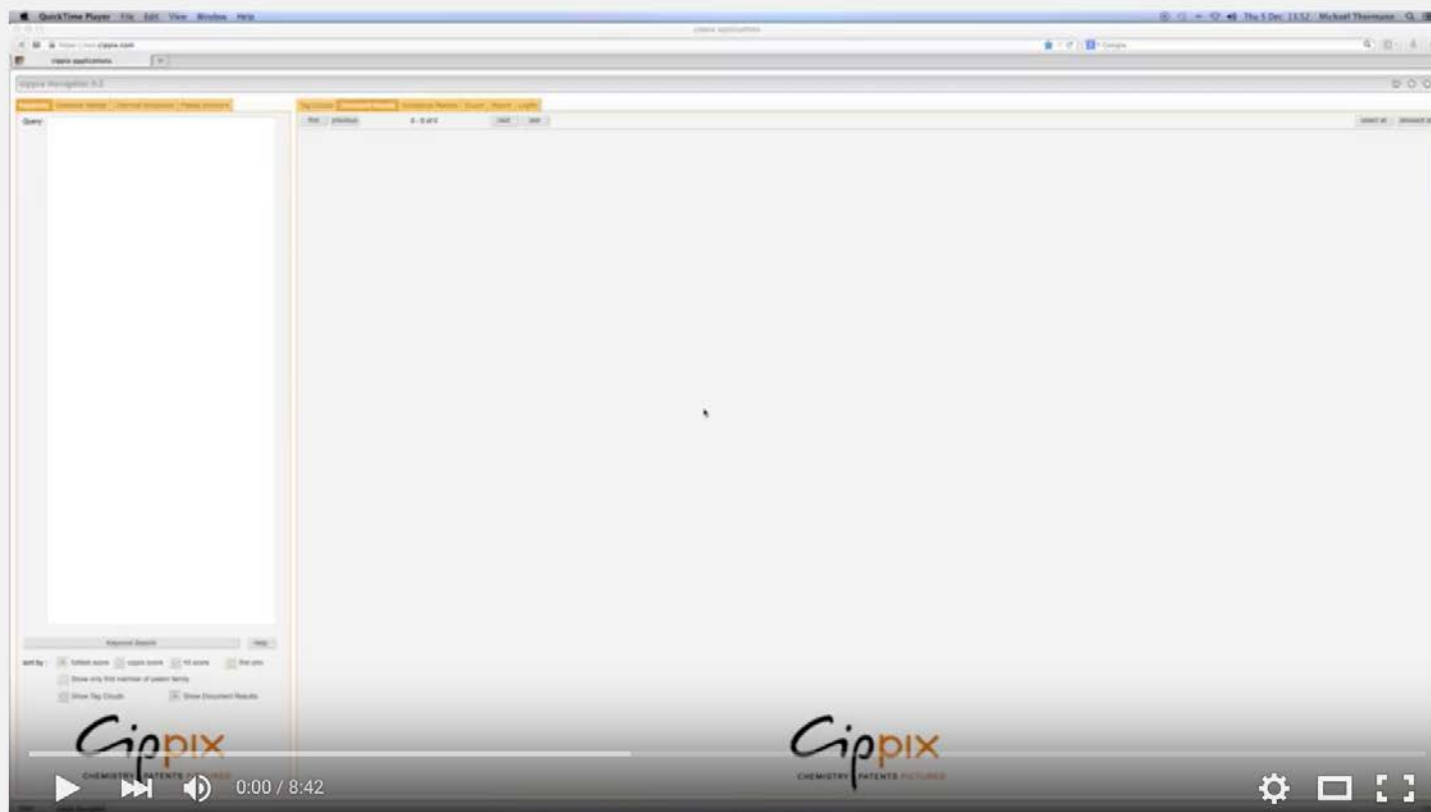
**Exercise**





## 4. Identify specific or similar/relevant documents

- **Document searching – given a specific documents, identify other documents in the corpus which are similar and relevant**
- **Create a pool of similar/linked documents for analysis**
  - Patent search, primary research
  - Forensic investigations into text
- **Web search**



## Cippix Tutorial: How to search for similar documents



Mestrelab Research

Subscribe 142

32 views

+ Add to Share ... More

0 0

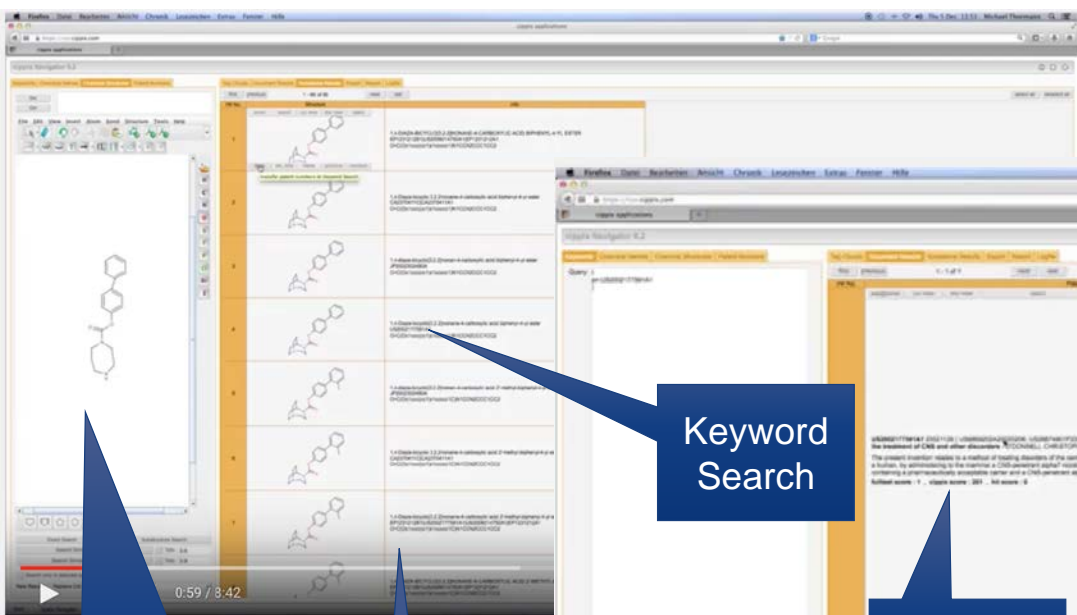
<https://www.youtube.com/watch?v=evLDjHQzMRU>

# Findings to Note

Chemical Substructure query

Matching Documents

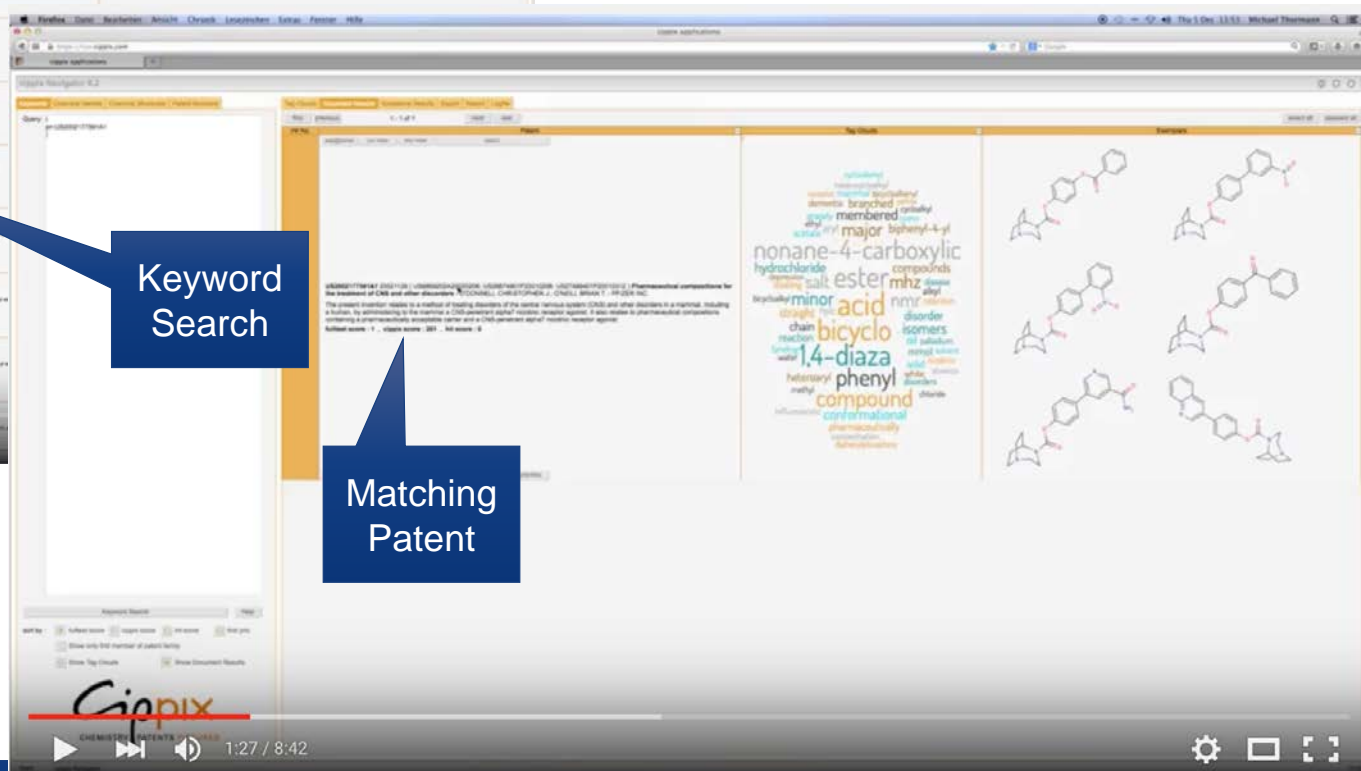
# Findings to Note



Keyword Search

Matching Patent

Matching Documents



# Findings to Note

Chemical  
Substructure  
query

Keyword  
Search

Matching  
Patent

Matching  
Documents

Word  
Analysis

# Findings to Note

Chemical  
Substructure  
query

Keyword  
Search

Matching  
Patent

Continue  
Searching

Matching  
Documents

Word  
Analysis





**What was the Business Context?**

---

---

**What was the Business Question/Need?**

---

---

**What was the data that was used?**

---

---

**What was the answer that was obtained?**

---

---

**What advantage did text mining provide in this case?**

---

---

**Exercise**



## 5. Extract specific information from the text

- **There are many answers in text documents. The problem is given a question, how to get the answer, not just the document. The task is called “question answering” (QA)**
- **At a more basic level, identify and extract “named entities” from documents and corpora**
- **Automatic QA**
  - Compare interest rates at banks for best deal
  - Automated help desk and FAQs
- **Name Entity Extraction (NER)**
  - Dates, money sums, organizations, stock symbols, etc.



# How IBM's Watson supercomputer wins at Jeopardy!, with IBM's Dave



Subscribe

447 videos ▾



Like



Share



113,383



Uploaded by [engadget](#) on Jan 13, 2011

How IBM's Watson supercomputer wins at Jeopardy, with IBM's Dave Gondek.

343 likes, 5 dislikes

<http://www.engadget.com/2011/01/13/ibms-watson-supercomputer-destroys-all-hum...>

From: [http://www.youtube.com/watch?v=d\\_yXV22O6n4](http://www.youtube.com/watch?v=d_yXV22O6n4)

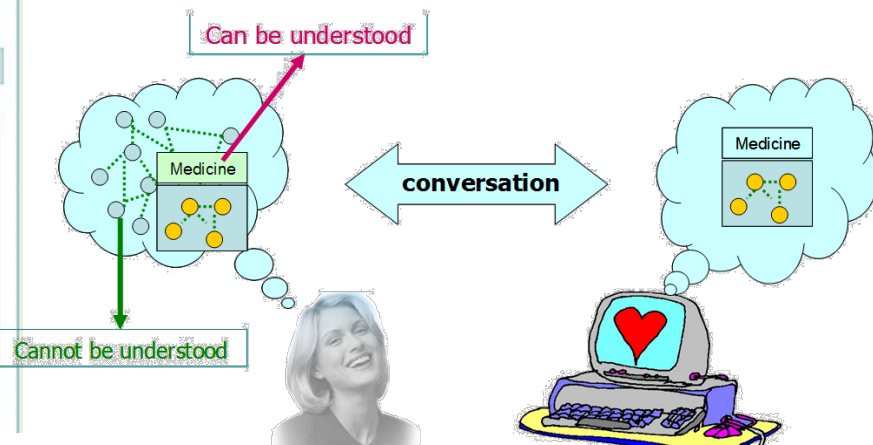
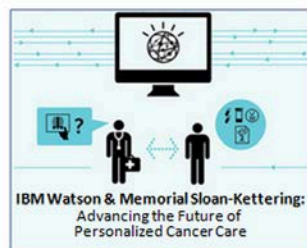
# High Tech Advancing Future of Personalized Cancer Care

01/19/2013

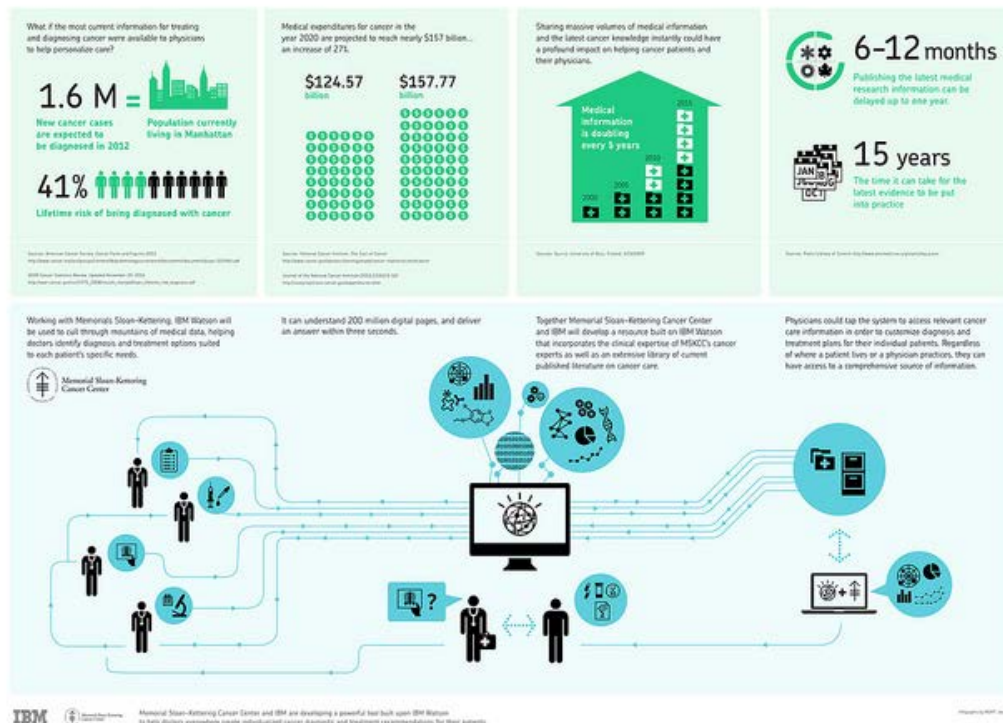
**Memorial Sloan-Kettering Cancer Center, IBM to Collaborate in Applying Watson Technology to Help Oncologists**

**IBM Watson combined with MSKCC's clinical knowledge will help physicians access and integrate latest science and knowledge**

**New York City – 22 Mar 2012:** Memorial Sloan-Kettering Cancer Center and IBM have agreed to collaborate on the development of a powerful tool built upon IBM Watson in order to provide medical professionals with improved access to current and comprehensive cancer data and practices. The resulting decision support tool will help doctors everywhere create individualized cancer diagnostic and treatment recommendations for their patients based on current evidence.



## Memorial Sloan Kettering & IBM Watson: Advancing the Future of Personalized Cancer Care





## IBM Watson Demo Oncology Diagnosis and Treatment 2 min.



kuresurem

Subscribe 72

1,315

Add to Share More

0 0

**Published on 14 Aug 2013**

The IBM Watson Cancer Diagnosis and Treatment Adviser demo was created in close collaboration with Memorial Sloan Kettering, one of the world's preeminent cancer treatment and research institutions. The demo scenario follows the interactions of a hypothetical oncologist and patient as they move through consultations, tests, treatment options, patient preferences and pre-authorization. It showcases IBM Watson's

<https://www.youtube.com/watch?v=uwbGgvEY244>



**What was the Business Context?**

---

---

**What was the Business Question/Need?**

---

---

**What was the data that was used?**

---

---

**What was the answer that was obtained?**

---

---

**What advantage did text mining provide in this case?**

---

---

**Exercise**

**What is the Business of my organisation?**

---

**What is the Business Need?**

---

**Do we have data that can be used? Who understands it?**

---

**What answer do we want? What action will result from that?**

---

---



# Reference & Resources

- **Chris Manning & Hinrich Schutze**, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- **NLP resources**: <http://nlp.stanford.edu/links/statnlp.html>
- **Christopher Potts (Stanford University)**, **Sentiment Symposium Tutorial**,  
<http://sentiment.christopherpotts.net/index.html>
- **John Elder, Gary Miner, Bob Nisbet**. *Practical Text Mining and Statistical Analysis for non-Structured Text Data Applications*, Academic Press, 2012
- **Roger Bilisoly**. *Practical Text Mining with PERL*, John Wiley & Sons, 2008