Mineração de Textos

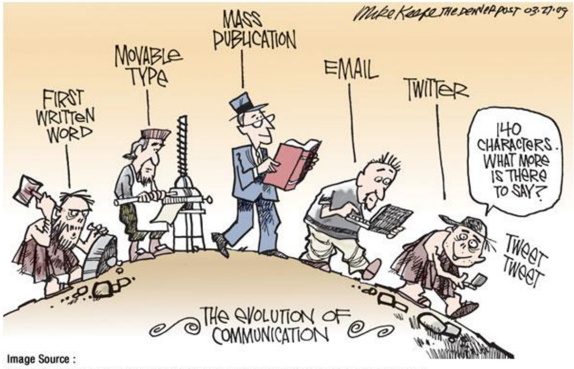

Mike Keefe THE DENVER POST 03·27·09
MASS PUBLICATION
MOVABLE TYPE
FIRST WRITTEN WORD
EMAIL
TWITTER
140 CHARACTERS. WHAT MORE IS THERE TO SAY?
TWEET TWEET
THE EVOLUTION OF COMMUNICATION

Image Source :
http://zenashapter.com/blog/wp-content/uploads/2014/05/20090326_085025_OP27keefedpo.jpg

A (brief) History of Text Mining

Prof. Rinaldo Lima
rinaldo.ufrpe@gmail.com

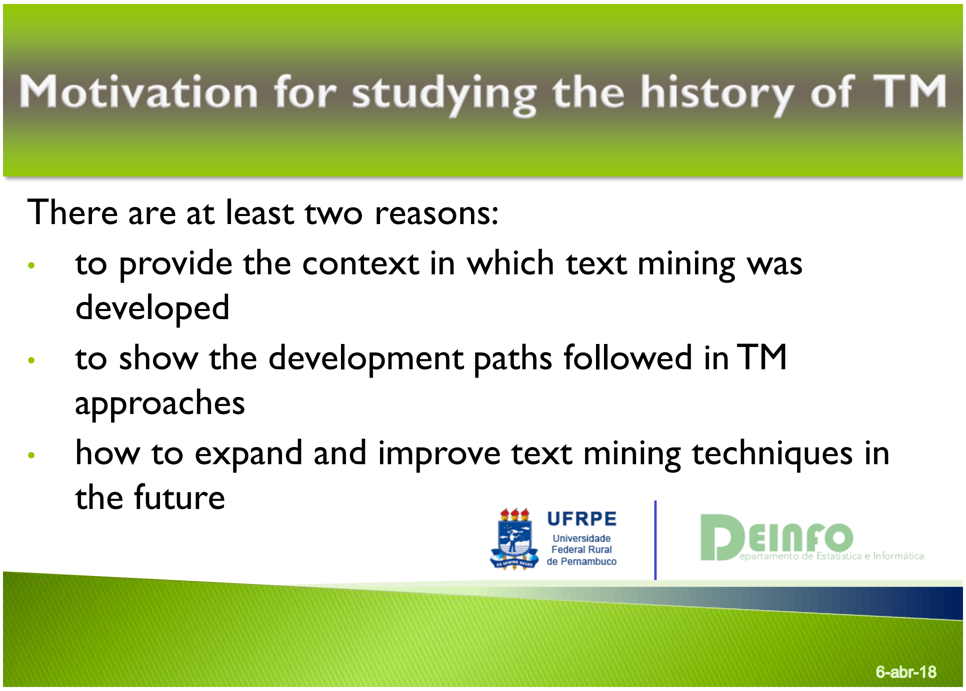
DEINFO Departamento de Estatística e Informática

6-abr-18

Contents

- Motivation for studying the history of TM
- The roots of TM
- Information Extraction (IE) and Modern Text Mining
- Major Innovations in TM since 2000
- Emerging Applications in TM

2

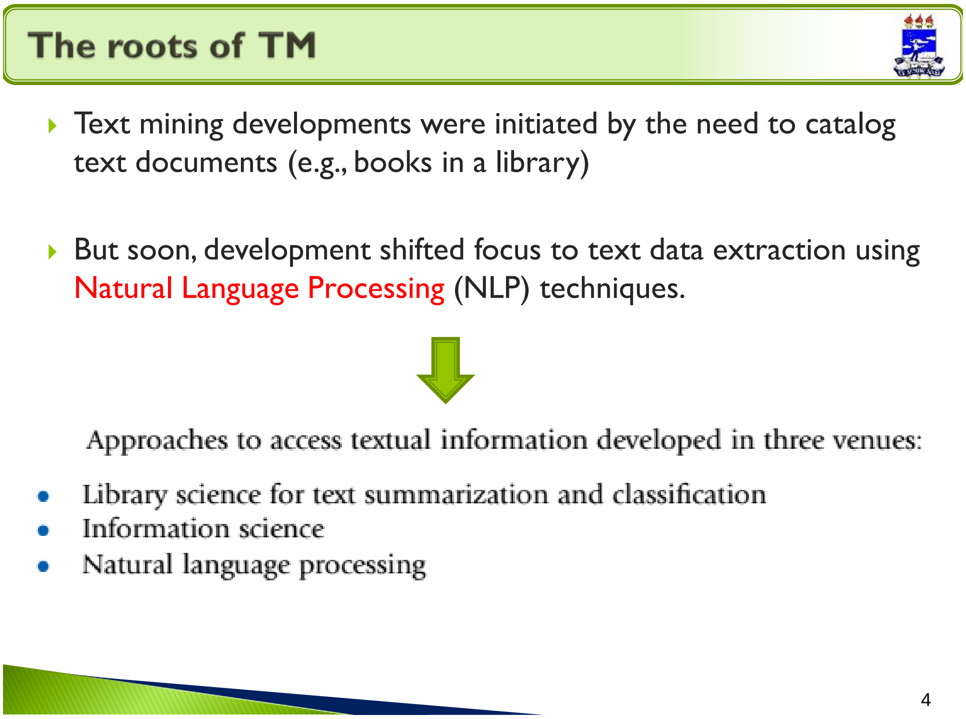


Motivation for studying the history of TM

There are at least two reasons:

- to provide the context in which text mining was developed
- to show the development paths followed in TM approaches
- how to expand and improve text mining techniques in the future

6-abr-18



The roots of TM

- Text mining developments were initiated by the need to catalog text documents (e.g., books in a library)
- But soon, development shifted focus to text data extraction using Natural Language Processing (NLP) techniques.

Approaches to access textual information developed in three venues:

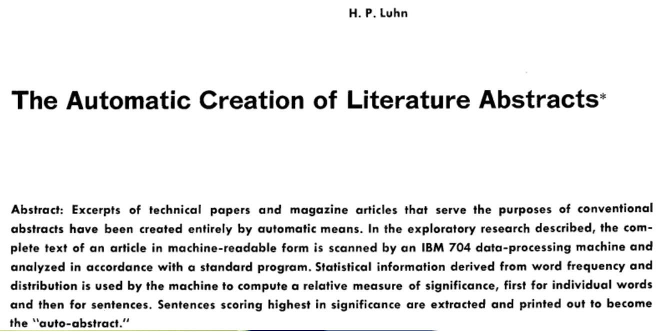
- Library science for text summarization and classification
- Information science
- Natural language processing

4

Book Summarization and Classification

One of the earliest examples of text summarization and classification was the library catalog.

Another step in the development of text processing was the summarization of text to generate abstracts



H. P. Luhn

The Automatic Creation of Literature Abstracts*

Abstract: Excerpts of technical papers and magazine articles that serve the purposes of conventional abstracts have been created entirely by automatic means. In the exploratory research described, the complete text of an article in machine-readable form is scanned by an IBM 704 data-processing machine and analyzed in accordance with a standard program. Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the "auto-abstract."

Seminal Work on Automatic Text Sumarization In 1954

Luhn's Summarization Method

1. It performed a word frequency analysis on an early IBM 701 computer (the first commercial computer, built with vacuum tubes)

2. a relative measure of significance was derived for the words
   - The number of relatively significant words was counted for each sentence and combined with the linear distances between the words to produce a metric of sentence significance

3. The most significant sentences, according to several criteria, were then extracted to compose an abstract for the document

Doyle 1961

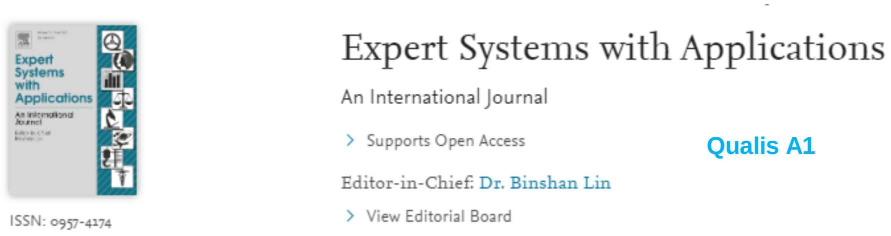
Based on the Luhn's work, Doyle proposed a new way to classify information in a library in the form of word frequencies and associations.

This system became a highly systematic and automated method for rapid browsing of information in digital libraries

Bibliometrics

- Applications of information theory to printed text developed along several lines in the 50's
- The science of bibliometrics arose to provide a numerical means to study and measure texts and information.
- One of the most common bibliometric applications was the formation of the citation index
  - It analyzes the references to one text document contained in other text documents
- The average number of citations per year can be used as a metric of importance for an article (journal)

Bibliometrics Exemplo de Example



Expert Systems with Applications

An International Journal

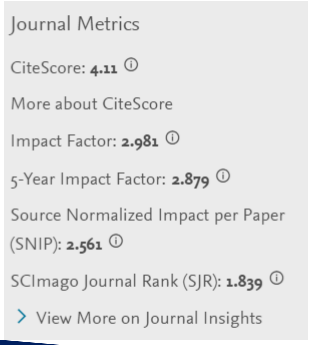
› Supports Open Access

Editor-in-Chief: Dr. Binshan Lin

› View Editorial Board

ISSN: 0957-4174

Qualis A1



Journal Metrics

CiteScore: **4.11** ⓘ

More about CiteScore

Impact Factor: **2.981** ⓘ

5-Year Impact Factor: **2.879** ⓘ

Source Normalized Impact per Paper (SNIP): **2.561** ⓘ

SCImago Journal Rank (SJR): **1.839** ⓘ

› View More on Journal Insights

Natural Language Processing

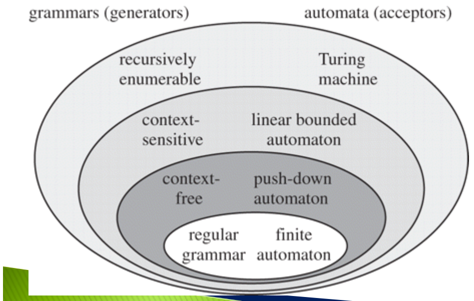
A hybrid discipline developed from the elements of linguistics and information science

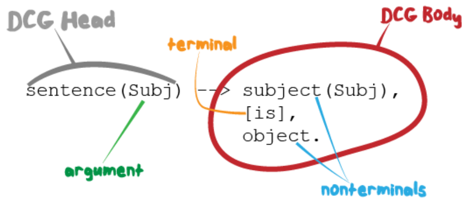
It attempts to understand how natural human language is learned and how it can be modeled

NLP began as an attempt to translate language on a computer

Noah Chomsky (1959) : the basis for PLN

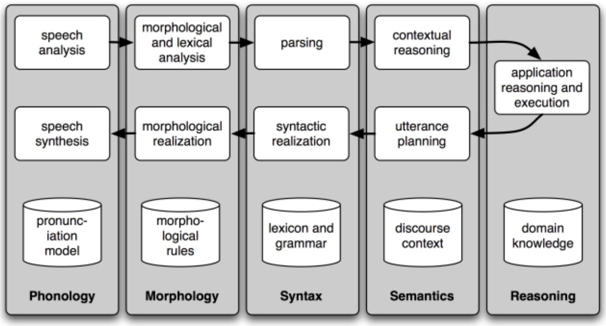
- The birth of computational linguistics in natural language processing (NLP)
- Chomsky championed the idea of "generative grammar": rule-based descriptions of syntactic structures.
- Between 1980 and 1990, Chomsky's generative linguistics approach ruled as the dominant philosophy in NLP.

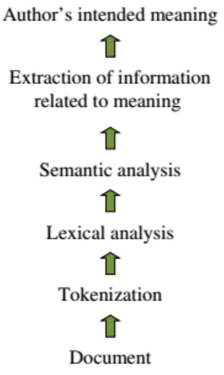




Modern NLP

- The next phase of NLP (by the year 2000) was primarily interested in understanding the meaning and the context of the information, rather than focusing just on the words themselves




Author's intended meaning
Extraction of information related to meaning
Semantic analysis
Lexical analysis
Tokenization
Document

Stages of analysis in natural language processing. *Source: Based on Dale et al., 2000.*

Information Extraction and Modern NLP

- **Information extraction (IE)** consists of an ordered series of steps designed to extract terms, terms attributes, relations, and events (Sanger and Feldman, 2007)
- IE was promoted back as far as 1987 in the Message Understanding Conferences (MUCs)
- The MUCs were initiated by the Naval Ocean Systems Center (NOSC), with assistance from the Defense Advanced Research Projects Agency DARPA).

The influence of the MUCs in TM community:

1. Defining the processes of named entity recognition (often referred to in text mining as proper name identification).
2. Formalizing the test metrics recall, precision and F-measure.
3. The importance of robustness in ML models.
   - The generalizability of a model is a measure of its successful application to data sets other than the one used for training and testing
4. The importance of making distinctions of coreference among noun phrases in the MUCs, i.e., the process of matching pairs of NLP expressions that refer to the same entity in the real world
5. The importance of word sense disambiguation led to the development of stochastic content-free grammars in probabilistic text mining models (Collins, 1997).

The Impact of Domain Knowledge on Text Mining

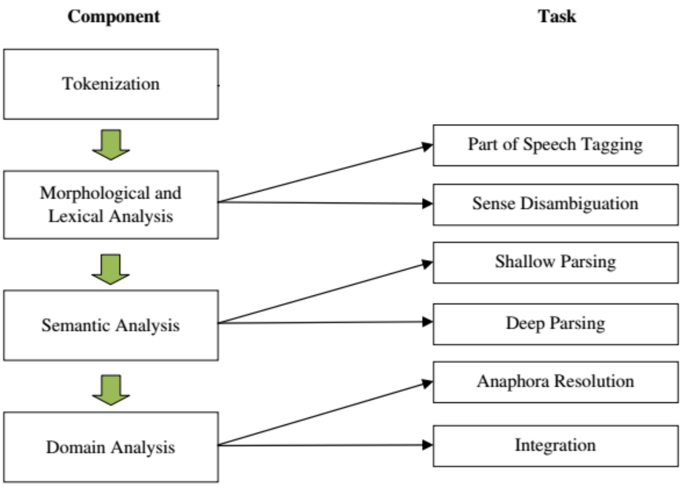
Recent research on NLP investigates the notion of including domain (or background) knowledge in processing was hotly contested.

An ontology can be viewed as the set of all concepts of interest and the relationships between them in a given domain knowledge base, ex. ontologies and domain lexicon.

- One example is the **gene ontology (GO**) knowledge base assembled by Princeton University beginning in 1998.
- The GO project developed three structured controlled vocabularies that describe various gene products across all of their functions and processes independent of species

Major Innovations in TM since 2000

Modern Information Extraction Engines



Bag-of-Words versus High-Dimensional Vector Spaces

- An important advance in the early 2000s was the extension of the **bag-of-words concept (BOW)** to a higher-dimensional space of "**features**" defined by nonlinear functions.
- Cortes and Vapnik (1995) proposed the kernel-based learning method which has been applied to various information extraction tasks.
- A kernel uses a nonlinear function to "map" text terms (words or phrases) to a higher-dimensional "feature space" (Renders, 2004)
- For BOW based-features, basic kernel functions such as
  - linear kernel,
  - polynomial kernel, and
  - Gaussian kernel are often used

IBM's Watson: an "intelligent" text mining machine?

- In 2007, IBM developed the world's most advanced **question answering machine** that was able to understand a question posed in human language (natural language) and respond with a precise, factual answer.
- Watson is a form of a text mining machine, dedicated to answering spoken or written questions. (QA System)
- This machine can
  - process the language of the questioner,
  - understand the meaning of the questions, and
  - produce an answer in terms of words and concepts stored in its memory based on a huge knowledge base gathered from web pages

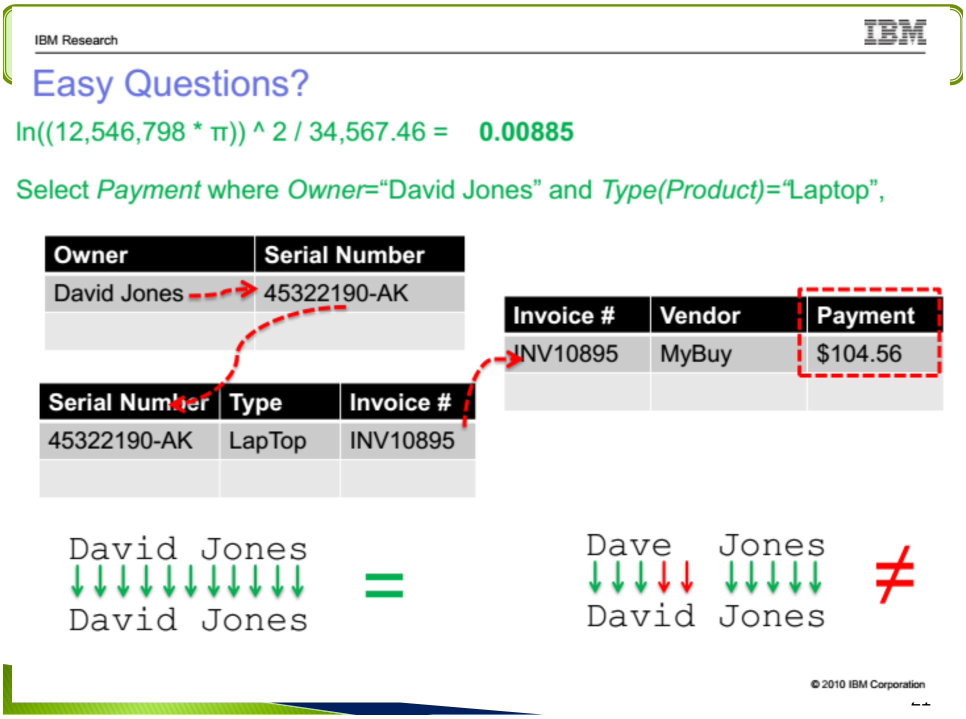
IBM Watson – A Question Answering Machine

IBM Watson – A Question Answering Machine

The Jeopardy! Challenge posed a unique and compelling artificial intelligence question…

"Can a computer system be designed to compete against the best humans at a task thought to require high levels of human intelligence, and if so, what kind of technology, algorithms, and engineering is required?"

Ferrucci, et al, *Building Watson*, AI Magazine, Fall 2010



IBM Research

Easy Questions?

ln((12,546,798 * π)) ^ 2 / 34,567.46 = **0.00885**

Select *Payment* where *Owner*="David Jones" and *Type(Product)*="Laptop",

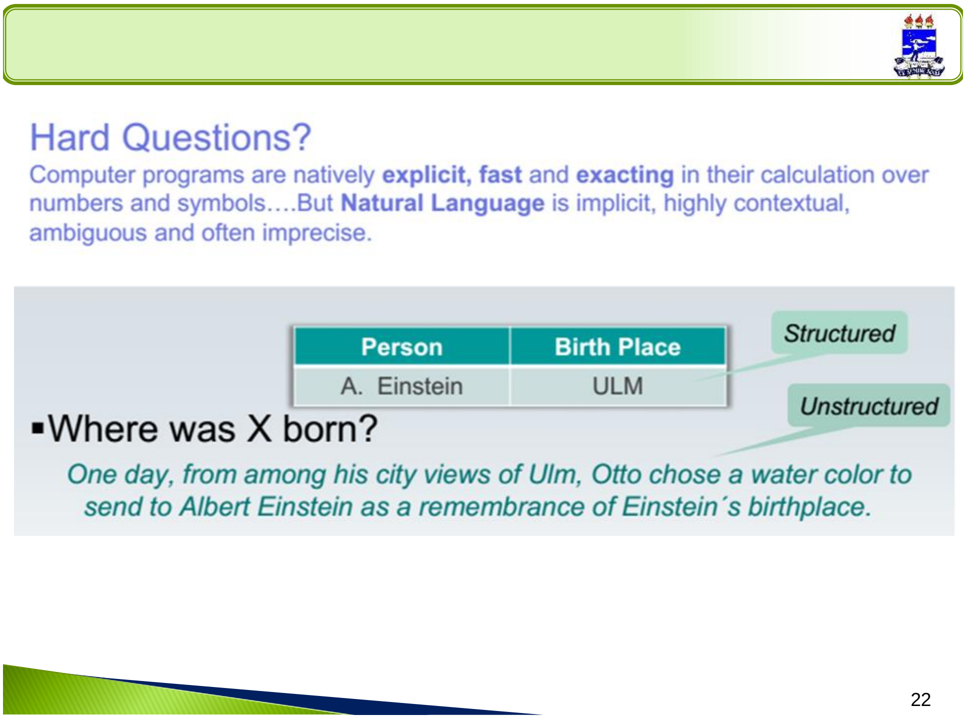
Owner	Serial Number
David Jones	45322190-AK

Serial Number	Type	Invoice #
45322190-AK	LapTop	INV10895

Invoice #	Vendor	Payment
INV10895	MyBuy	$104.56

```
David Jones
↓↓↓↓↓↓↓↓↓↓↓        =
David Jones
```

```
Dave  Jones
↓↓↓↓↓ ↓↓↓↓↓        ≠
David Jones
```



Hard Questions?

Computer programs are natively **explicit, fast** and **exacting** in their calculation over numbers and symbols….But **Natural Language** is implicit, highly contextual, ambiguous and often imprecise.

Person	Birth Place
A. Einstein	ULM

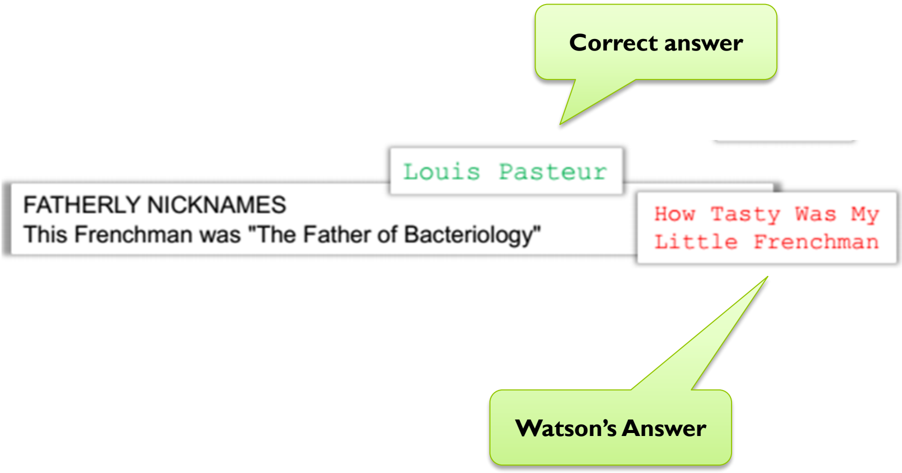
Structured

Unstructured

- Where was X born?

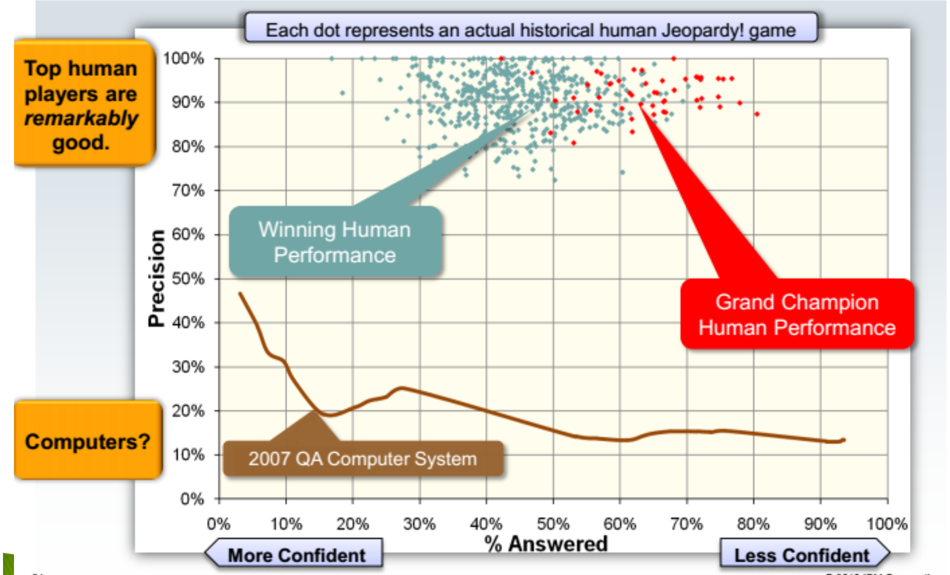
One day, from among his city views of Ulm, Otto chose a water color to send to Albert Einstein as a remembrance of Einstein´s birthplace.

Watson early Answers


Correct answer
Louis Pasteur
FATHERLY NICKNAMES
This Frenchman was "The Father of Bacteriology"
How Tasty Was My Little Frenchman
Watson's Answer

IBM Research

The Best Human Performance: *Our Analysis Reveals the* ***Winner's Cloud***

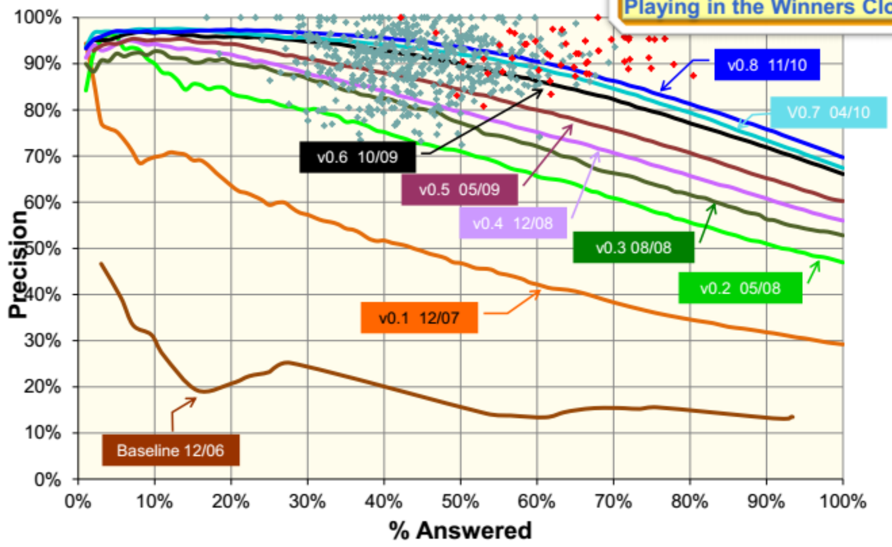


IBM Research

IBM

DeepQA: Incremental Progress in Answering Precision on the Jeopardy Challenge: 6/2007-11/2010

IBM *Watson* Playing in the Winners Cloud

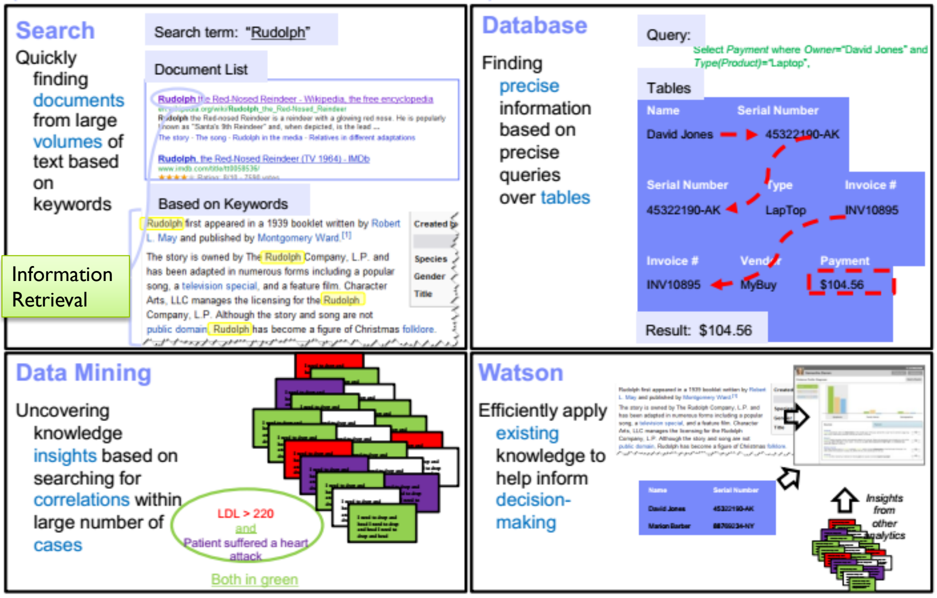


IBM Research

IBM

Search vs. Database vs. Data Mining vs. Watson

(Watson uses unstructured *and* structured sources)



Search

Quickly finding documents from large volumes of text based on keywords

Information Retrieval

Search term: "Rudolph"

Document List

Rudolph the Red-Nosed Reindeer - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Rudolph_the_Red-Nosed_Reindeer
Rudolph the Red-nosed Reindeer is a reindeer with a glowing red nose. He is popularly known as "Santa's 9th Reindeer" and, when depicted, is the lead ...
The story - The song - Rudolph in the media - Relatives in different adaptations

Rudolph, the Red-Nosed Reindeer (TV 1964) - IMDb
www.imdb.com/title/tt0058536/
★★★★ Rating: 8/10 - 7598 votes

Based on Keywords

Rudolph first appeared in a 1939 booklet written by Robert L. May and published by Montgomery Ward.[1]

The story is owned by The Rudolph Company, L.P. and has been adapted in numerous forms including a popular song, a television special, and a feature film. Character Arts, LLC manages the licensing for the Rudolph Company, L.P. Although the story and song are not public domain, Rudolph has become a figure of Christmas folklore.

Created by
Species
Gender
Title

Database

Finding precise information based on precise queries over tables

Query: *Select Payment where Owner="David Jones" and Type(Product)="Laptop",*

Tables

Name	Serial Number
David Jones	45322190-AK

Serial Number	Type	Invoice #
45322190-AK	LapTop	INV10895

Invoice #	Vendor	Payment
INV10895	MyBuy	$104.56

Result: $104.56

Data Mining

Uncovering knowledge insights based on searching for correlations within large number of cases

LDL > 220 and Patient suffered a heart attack

Both in green

Watson

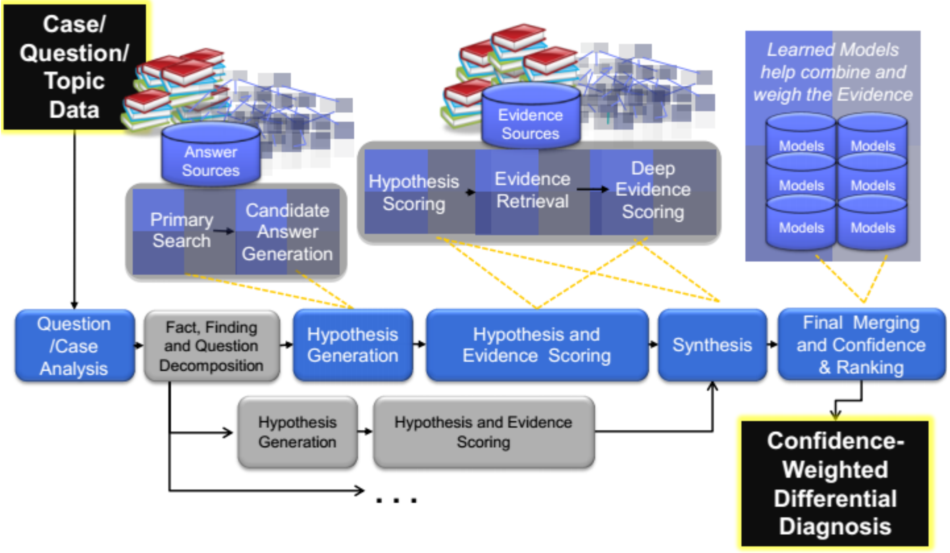
Efficiently apply existing knowledge to help inform decision-making

Name	Serial Number
David Jones	45322190-AK
Marion Barber	88769234-NY

Insights from other analytics

IBM Research

Generalized DeepQA Reasoning Paradigm



Case/ Question/ Topic Data

Answer Sources

Primary Search

Candidate Answer Generation

Evidence Sources

Hypothesis Scoring

Evidence Retrieval

Deep Evidence Scoring

Learned Models help combine and weigh the Evidence

Models

Models

Models

Models

Models

Models

Question /Case Analysis

Fact, Finding and Question Decomposition

Hypothesis Generation

Hypothesis and Evidence Scoring

Synthesis

Final Merging and Confidence & Ranking

Hypothesis Generation

Hypothesis and Evidence Scoring

. . .

Confidence-Weighted Differential Diagnosis

© 2010 IBM Corporation

IBM Research

Watson Workload Optimized System (Power 750)

- 90 x IBM Power 750[1] servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops
- Runs IBM DeepQA software
- Scales out with and searches vast amounts of unstructured information with UIMA & Hadoop open source components
- SUSE Linux provides a cost-effective open platform which is performance-optimized to exploit POWER 7 systems
- 10 racks include servers, networking, shared disk system, cluster controllers

Linux

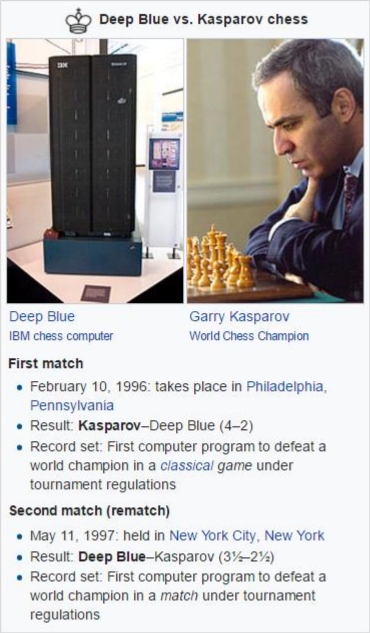
POWER7™ BUILT ON Power™

[1] Note that the Power 750 featuring POWER7 is a commercially available server that runs AIX, IBM i and Linux and has been in market since Feb 2010

© 2010 IBM Corporation

IBM Deep Blue – Chess Machine

Deep Blue, at the Computer History Museum



Deep Blue vs. Kasparov chess

Deep Blue
IBM chess computer

Garry Kasparov
World Chess Champion

First match

- February 10, 1996: takes place in Philadelphia, Pennsylvania
- Result: **Kasparov**–Deep Blue (4–2)
- Record set: First computer program to defeat a world champion in a *classical game* under tournament regulations

Second match (rematch)

- May 11, 1997: held in New York City, New York
- Result: **Deep Blue**–Kasparov (3½–2½)
- Record set: First computer program to defeat a world champion in a *match* under tournament regulations

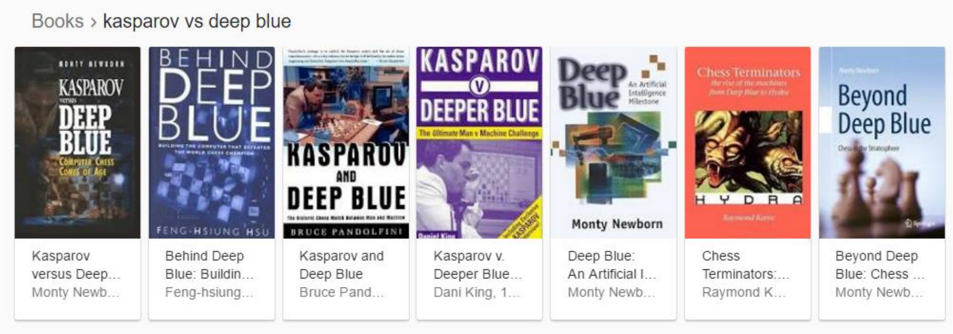
IBM Deep Blue – Chess Machine

IBM Deep Blue – Chess Machine

Deep Blue–Kasparov, 1997 game 6

r 7...h6; Deep Blue continued

Media Coverage and Books



Books › kasparov vs deep blue

Kasparov versus Deep… Monty Newb…

Behind Deep Blue: Buildin… Feng-hsiung…

Kasparov and Deep Blue Bruce Pand…

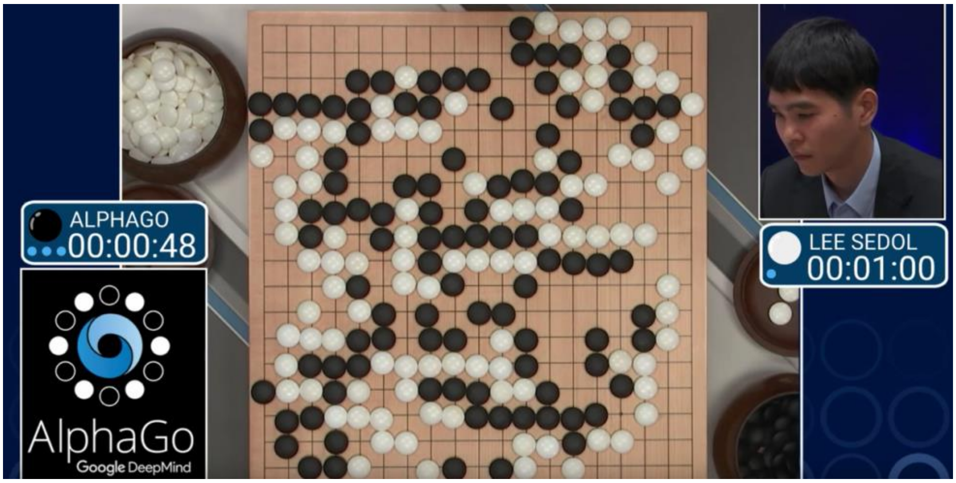
Kasparov v. Deeper Blue… Dani King, 1…

Deep Blue: An Artificial I… Monty Newb…

Chess Terminators:… Raymond K…

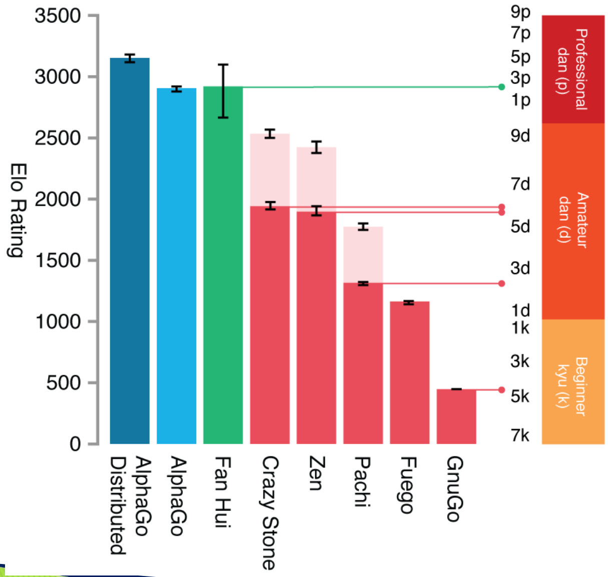
Beyond Deep Blue: Chess … Monty Newb…

More Recently: Alpha Go



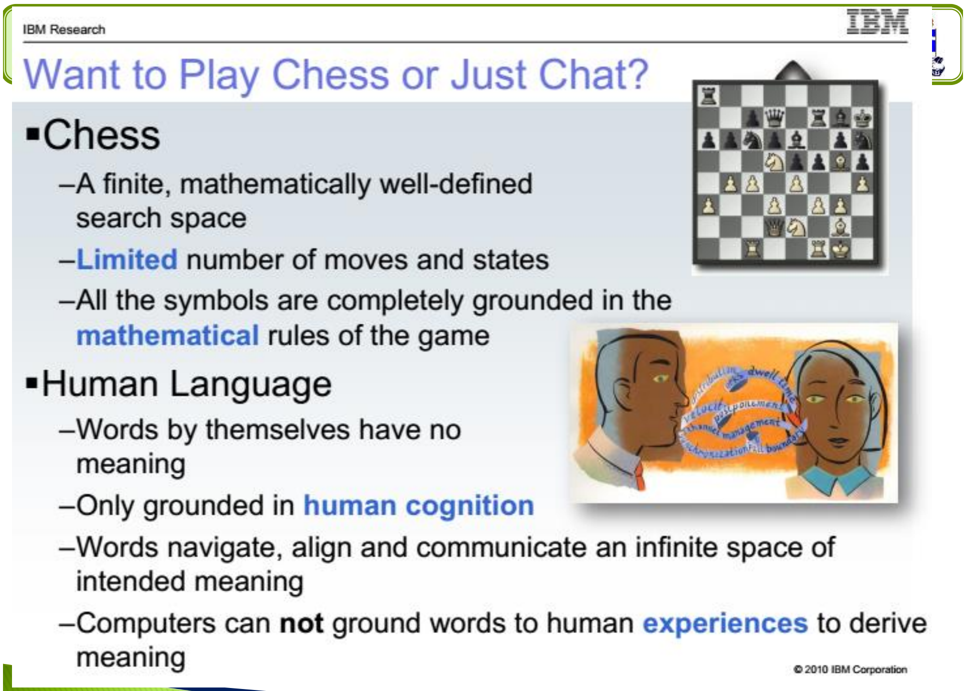
4 x 1

Alpha Go: better than any human!



AlphaGo’s greatest challenge

0 x 3

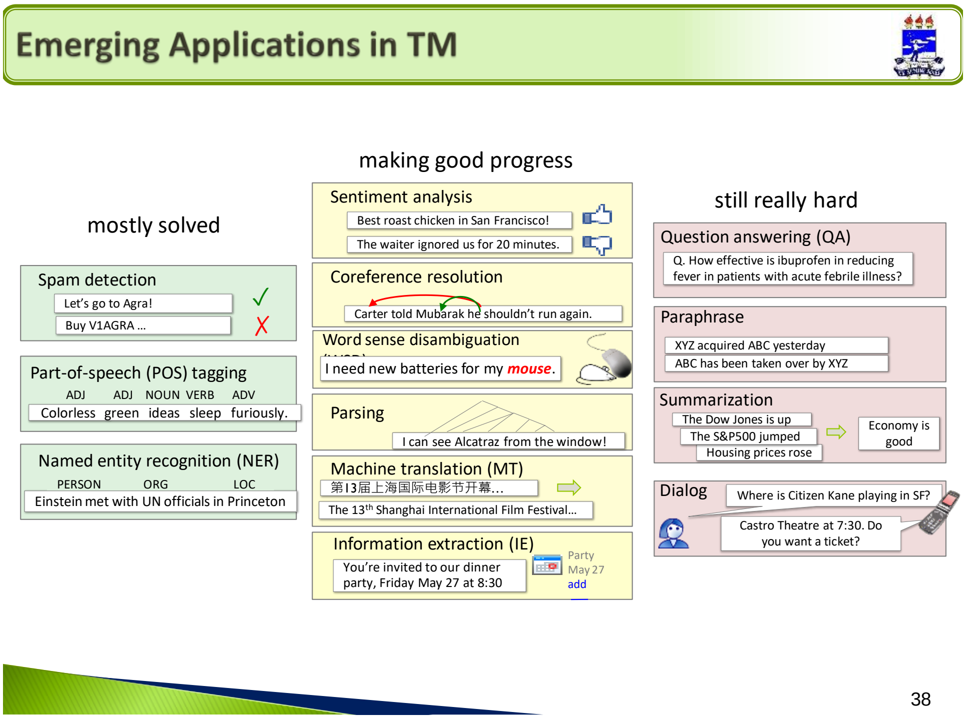


IBM Research

IBM

Want to Play Chess or Just Chat?

- Chess
  - –A finite, mathematically well-defined search space
  - –**Limited** number of moves and states
  - –All the symbols are completely grounded in the **mathematical** rules of the game
- Human Language
  - –Words by themselves have no meaning
  - –Only grounded in **human cognition**
  - –Words navigate, align and communicate an infinite space of intended meaning
  - –Computers can **not** ground words to human **experiences** to derive meaning



Emerging Applications in TM

making good progress

Sentiment analysis
Best roast chicken in San Francisco!
The waiter ignored us for 20 minutes.

Coreference resolution
Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)
I need new batteries for my ***mouse***.

Parsing
I can see Alcatraz from the window!

Machine translation (MT)
第13届上海国际电影节开幕…
The 13th Shanghai International Film Festival…

Information extraction (IE)
You're invited to our dinner party, Friday May 27 at 8:30
Party May 27 add

mostly solved

Spam detection
Let's go to Agra! ✓
Buy V1AGRA … ✗

Part-of-speech (POS) tagging
ADJ ADJ NOUN VERB ADV
Colorless green ideas sleep furiously.

Named entity recognition (NER)
PERSON ORG LOC
Einstein met with UN officials in Princeton

still really hard

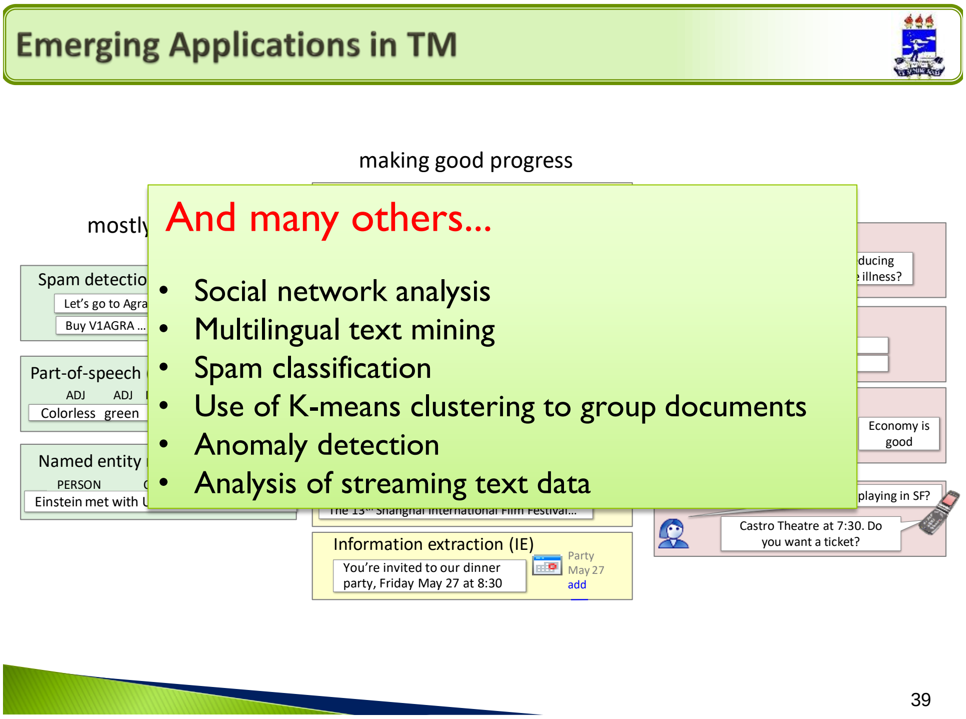
Question answering (QA)
Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
XYZ acquired ABC yesterday
ABC has been taken over by XYZ

Summarization
The Dow Jones is up
The S&P500 jumped
Housing prices rose
Economy is good

Dialog
Where is Citizen Kane playing in SF?
Castro Theatre at 7:30. Do you want a ticket?

38



Emerging Applications in TM

making good progress

mostly

And many others…

- Social network analysis
- Multilingual text mining
- Spam classification
- Use of K-means clustering to group documents
- Anomaly detection
- Analysis of streaming text data

Spam detectio
Let's go to Agra
Buy V1AGRA …
Part-of-speech (
ADJ ADJ
Colorless green
Named entity
PERSON
Einstein met with U
ducing
illness?
Economy is good
playing in SF?
Castro Theatre at 7:30. Do you want a ticket?
The 13th Shanghai International Film Festival…
Information extraction (IE)
You're invited to our dinner party, Friday May 27 at 8:30
Party May 27 add

39

Próxima Aula...

Introdução ao Processamento de Linguagem Natural