

Corso di *Web Mining e Retrieval*

- Introduzione al Corso -

Corso di Laurea in Informatica
(a.a. 2012-2013)

Roberto Basili



Overview

- WM&R: Motivazioni e prospettive
- Modalità di erogazione del Corso
- Prerequisiti
- Forma e struttura delle prove d'esame



Did you know?

To whom were these questions addressed B.G.? (Before Google)



If MySpace were a country,
it would be the **5th-largest** in the world
(between Indonesia and Brazil)



Did you know?



A Web of people and opinions

- **31.7%** of the more than 200 million bloggers worldwide blog about opinions on products and brands (Universal McCann, July 2009)
- **71%** of all active Internet users read blogs.
- 2009 Survey of **25,000** Internet users in **50** countries: **70%** of consumers trust opinions posted online by other consumers (Nielsen Global Online Consumer, 2010).



Social Media Analytics

- Complex process for Social Media Analytics are necessary whereas ...
- ... Opinion Mining and Sentiment Analysis play a crucial role



WM&R: Motivazioni

- *Cos'è il Web Mining?*
- *Perché IR?*
- *Perché Apprendimento Automatico?*
- *Quale contributo l'IR fornisce alle tecnologie di sfruttamento delle informazioni del Web?*
- *Quali sono le prospettive per l'impiego di tali tecnologie?*



Cos'è il Web Mining?

- *Web Mining* attualmente si riferisce ad un insieme di tecnologie necessarie per lo sfruttamento delle informazioni pubblicamente disponibili nel Web
 - Contenuti: dati ma anche ... persone, luoghi, eventi, concetti, ...
 - Relazioni:
 - Link strutturali
 - Collegamenti tematici, concettuali e interpersonali
 - Ridondanze/analogie
 - Multilingualità
 - Trend e comportamenti collettivi
 - Opinioni



Perché IR?

- La taglia delle informazioni in gioco pone il problema della *localizzazione*
- Accedere in modo automatico è possibile solo governando il problema di sapere **dove** si trova una informazione *rilevante*
- La ricerca corrisponde al calcolo di una funzione *aleatoria* di mapping tra requisiti e informazione utile



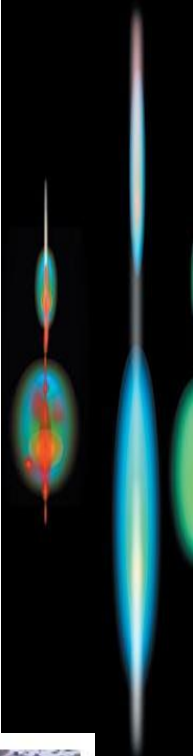
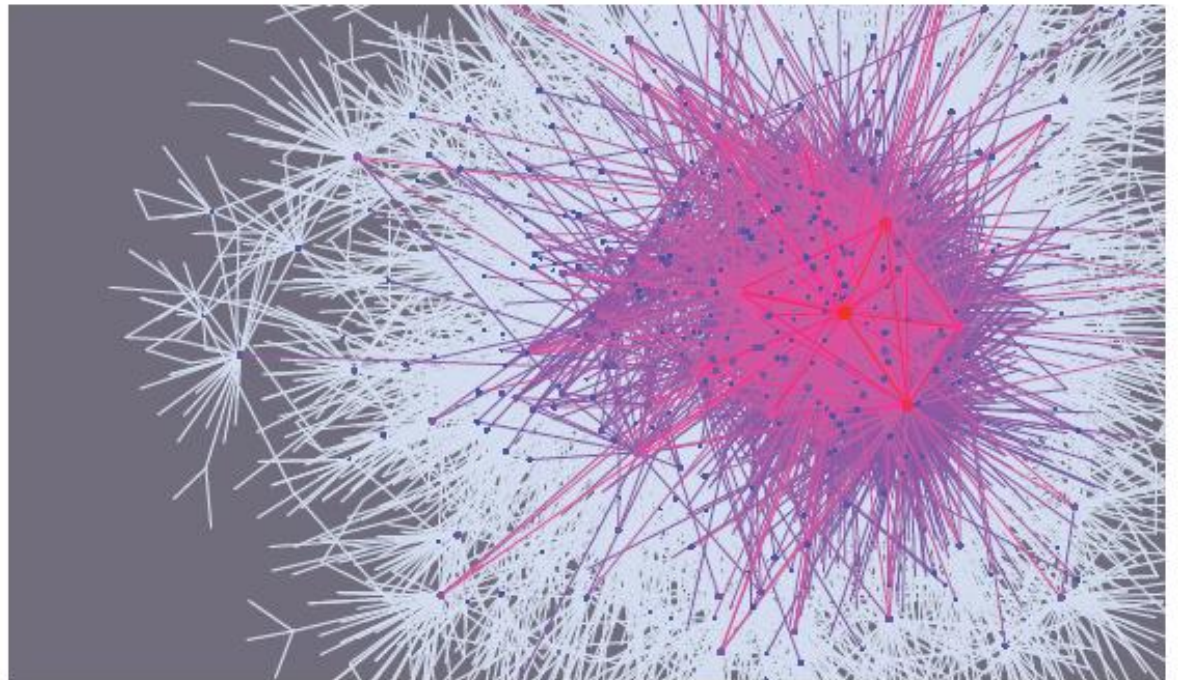
Machine Learning vs IR?

- La eterogeneità delle informazioni produce significativi effetti di incertezza nel processo di ricerca
 - Incompletezza della informazione
 - Ricchezza di dati, formati e modalità di accesso
 - Requisiti vaghi
 - Aspetti soggettivi
 - Tempestività



ML vs. IR

- La pervasività degli elementi di incertezza rende impraticabile la ricerca di soluzioni esaustive (ottimi globali)
- “*Finding diamonds in the rough*”
(Fan Chung, UCSD)



ML vs. IR

- Le tecniche di ML propongono una ampia serie di algoritmi, strategie e tecniche per la produzione di soluzioni *sub-ottime* effettive
- Nel processo di *learning* i dati suggeriscono la ipotesi risolutiva per la funzione di *mapping*
- Tale ipotesi è attesa migliorare la prestazione complessiva del sistema di base
 - Accuratezza
 - Efficienza computazionale



Machine Learning

- (Langley, 2000): l'Apprendimento Automatico si occupa dei meccanismi attraverso i quali un agente intelligente migliora nel tempo le sue prestazioni P nell'effettuare un compito C .
- La prova del successo dell'apprendimento è quindi nella capacità di misurare l'incremento ΔP delle prestazioni sulla base delle esperienze E che l'agente è in grado di raccogliere durante il suo ciclo di vita.
- La natura dell'apprendimento è quindi tutta nella caratterizzazione delle nozioni qui primitive di *compito*, *prestazione* ed *esperienza*.

Esperienza ed Apprendimento

- L'esperienza, per esempio, nel gioco degli scacchi può essere interpretata in diversi modi:
 - i dati sulle vittorie (e sconfitte) pregresse per valutare la bontà (o la inadeguatezza) di strategie e mosse eseguite rispetto all'avversario.
 - valutazione fornita sulle mosse da un docente esterno (oracolo, guida).
 - Adeguatazza dei comportamenti derivata dalla auto-osservazione, cioè dalla capacità di analizzare partite dell'agente contro se stesso secondo un modello esplicito del processo (partita) e della sua evoluzione (comportamento, vantaggi, ...).



Apprendimento senza supervisione

- In assenza di un oracolo o di conoscenze sul task esistono ancora molti modi di migliorare le proprie prestazioni, ad es.
 - Migliorando il proprio modello del mondo (acquisizione/*discovery* della conoscenza)
 - Migliorando le proprie prestazioni computazionali (ottimizzazione)



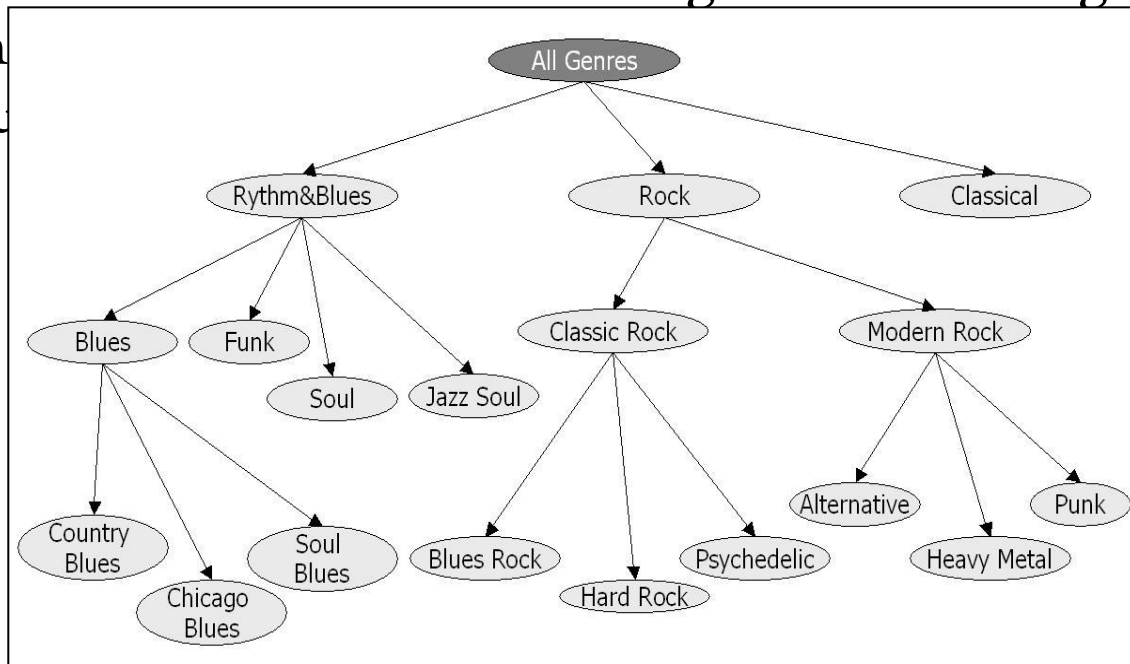
Apprendimento senza supervisione

- Esempio:

- una collezione mp3 può essere organizzata in generi attraverso il raggruppamento di brani simili secondo proprietà audio (*clustering*): tale organizzazione è naturalmente gerarchica
- Il miglioramento avviene quindi almeno rispetto agli algoritmi di ricerca: la organizzazione gerarchica

con
alcu

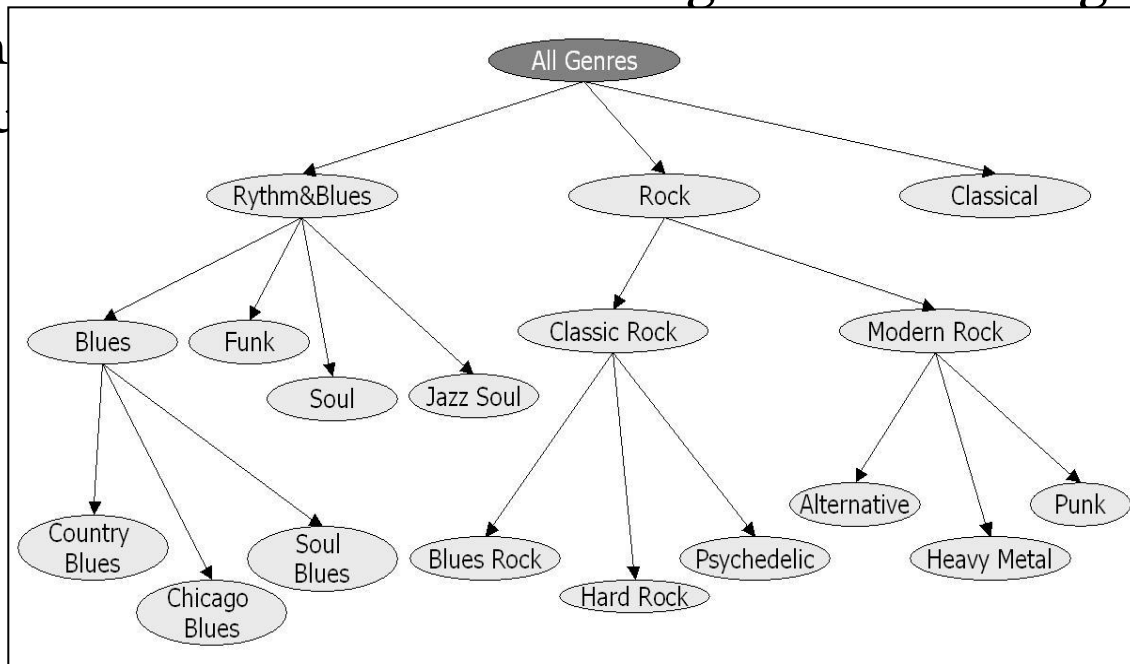
eme in



Apprendimento senza supervisione

- E Al termine del processo di acquisizione il sistema
– dispone di un sistema di classi e relazioni indotti che
migliora la sua interazione futura con l'ambiente
operativo (ad es. l'utente)
- Il miglioramento avviene quindi almeno rispetto agli
algoritmi di ricerca: la organizzazione gerarchica

con
alcu



L'apprendimento automatico

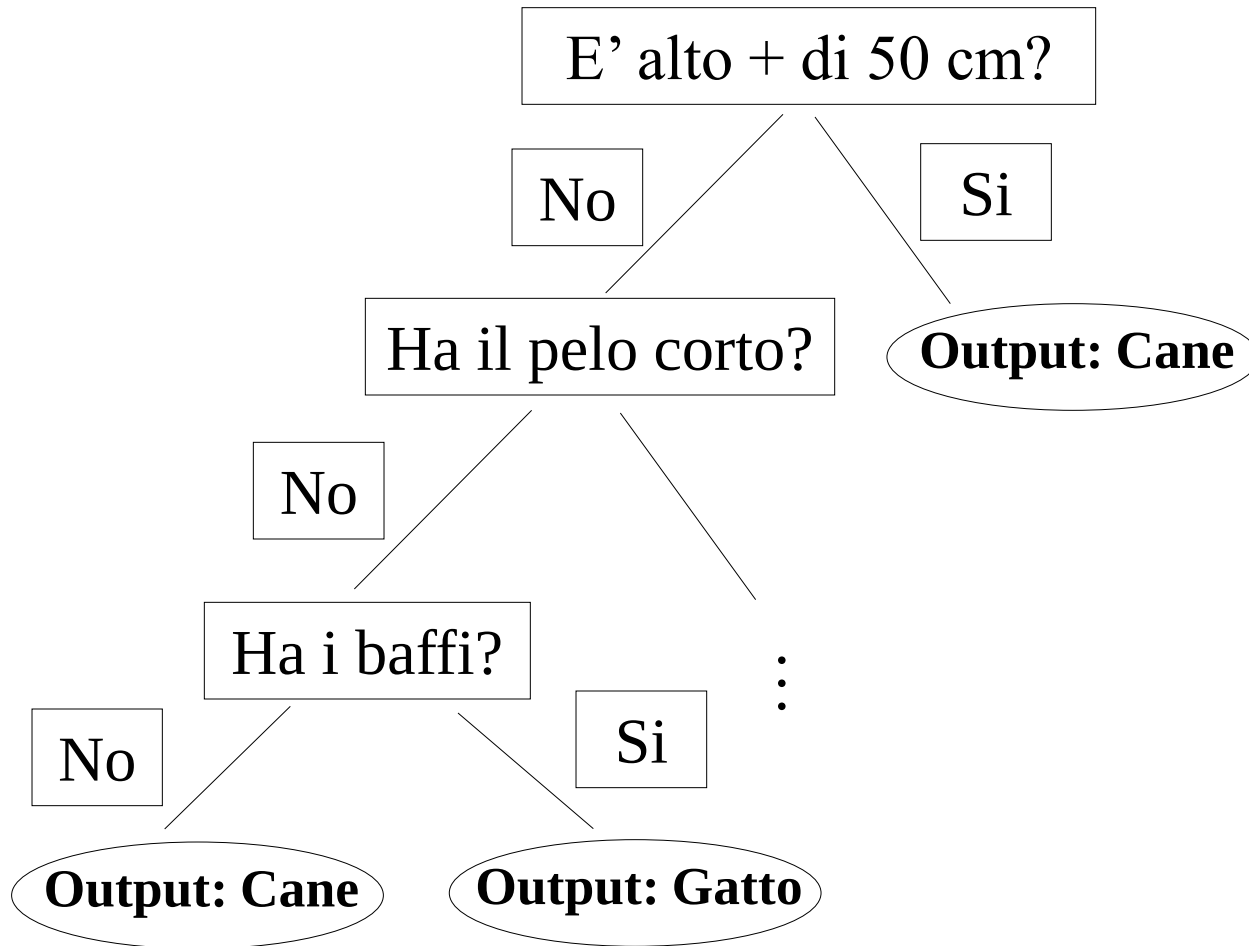
- Apprendere una funzione da esempi:
 - a valori reali, *regression*
 - a valori interi finiti, *classification*
- Supponiamo di volere apprendere una funzione intera:
 - 2 classi, *gatto e cane*
 - $f(x) \rightarrow \{\text{gatto}, \text{cane}\}$
- Dato un insieme di esempi per le due classi
 - Si estraggono le features (*altezza, baffi, tipo di dentatura, numero di zampe*).
- Si applica l'algoritmo di learning per generare f

Algoritmi di Apprendimento

- Funzioni logiche booleane, (ad es., alberi di decisione).
- Funzione di Probabilità, (ad es., classificatore Bayesiano).
- Funzioni di separazione in spazi vettoriali
 - Non lineari: KNN, reti neurali multi-strato,...
 - **Lineari**, percettroni, **Support Vector Machines**,...
- Trasformazioni di spazi: embeddings, analisi spettrale



Alberi di decisione (Gatti/Cani)



Web IR?

- I processi di IR studiati in domini antecedenti all'affermarsi del Web debbono essere estesi ed adattati rispetto alla maggiore ricchezza ed ai problemi maggiori che tali scenari presentano
 - Complessità strutturale: contenuti, topologia e uso
 - Affidabilità dell'informazione
 - Multimodalità, Multimedialità
 - Partecipazione (aspetti sociali)



Web IR

- Processing Web data: content detection, link detection, ...
- Web Crawling
- Web Search: indici, link analysis
- Ranking: weighting contents, links and formats, authority, timeliness
- Meta-search
- Link Analysis

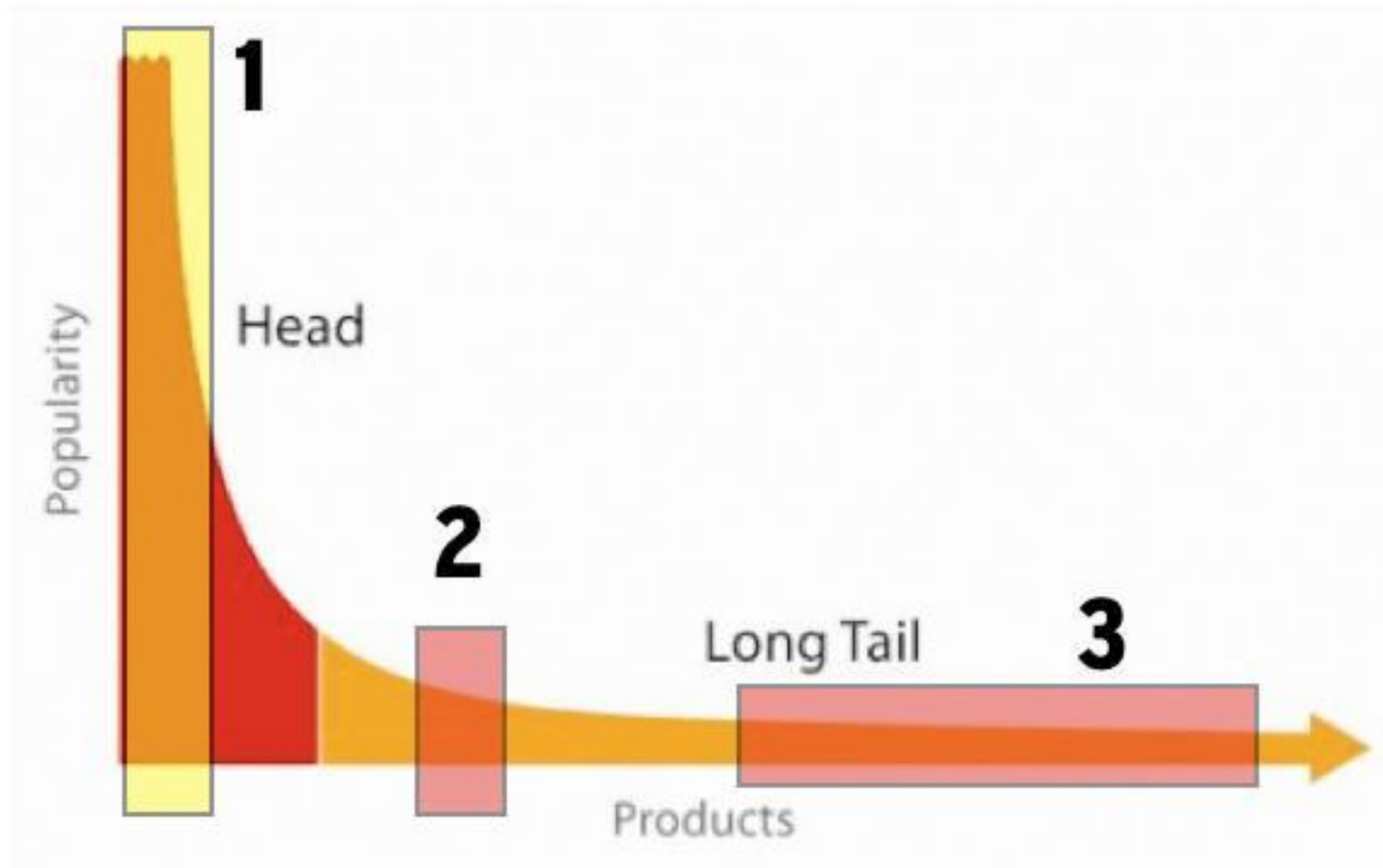


Prospettive delle tecnologie WM&R

- Crescita esponenziale della taglia dei problemi
- Crescente interesse verso processi di IR agenti su dati complessi (multimediali, sociali)
- Web partecipativo: Web 2.0
- Ruolo crescente della mediazione degli strumenti informatici
 - Software as a Service
 - Personalizzazione



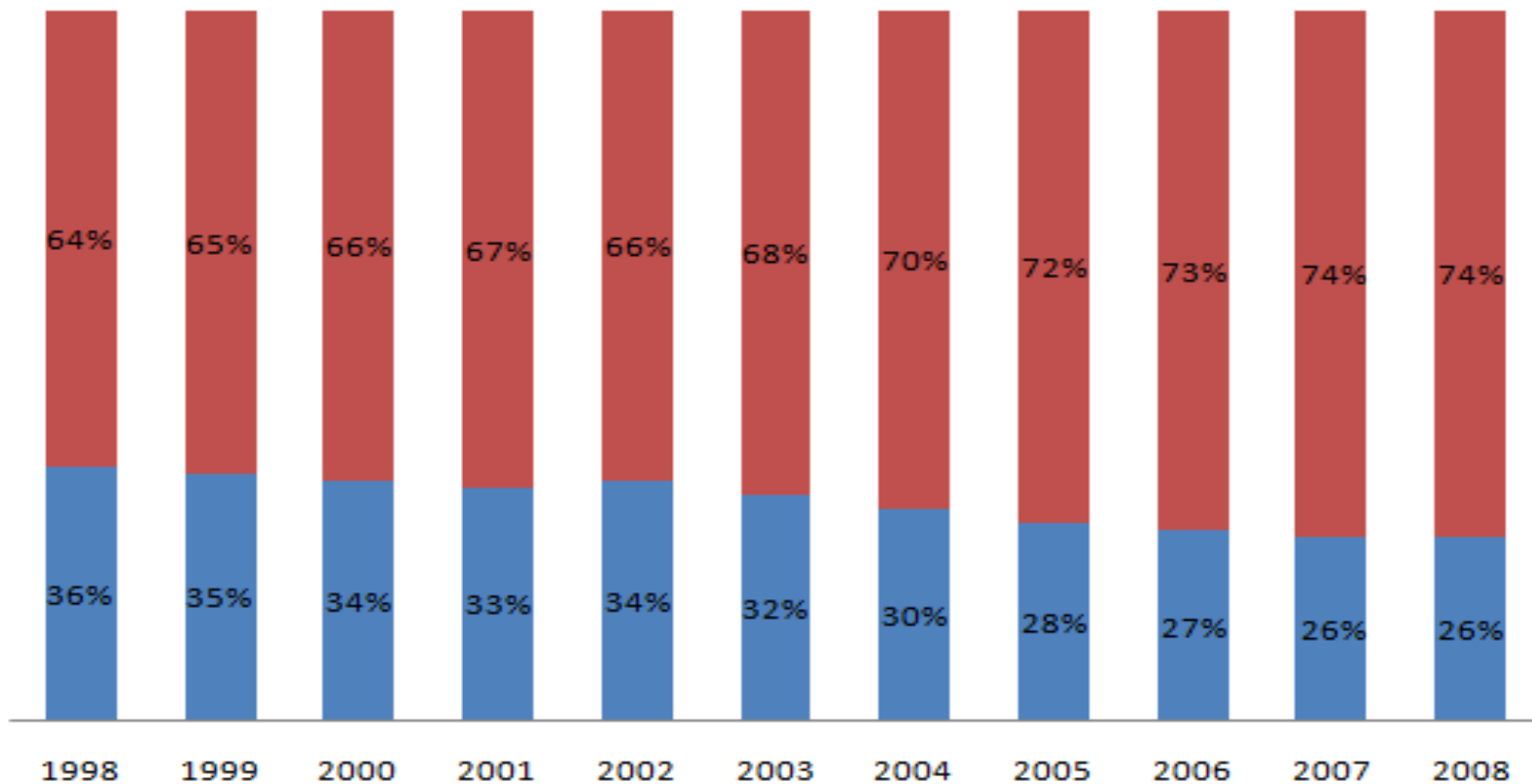
La lunga Coda



The Long Tail

The Long Tail of Travel

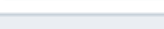
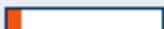
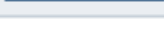
■ Top 50 travel destinations ■ Everywhere else

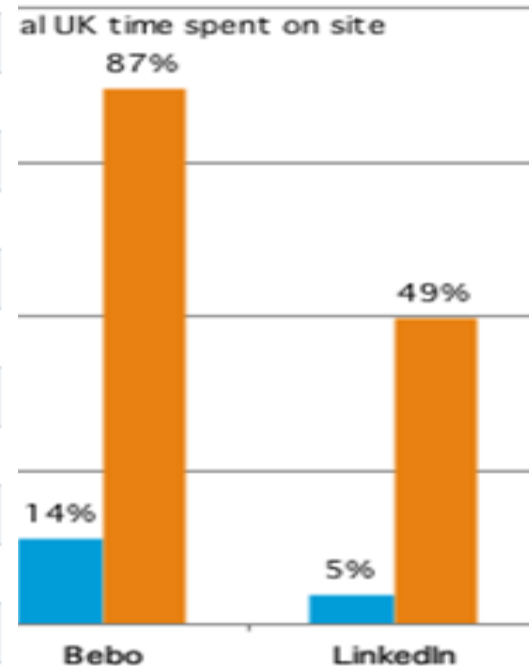


Maren Jinnett over data compiled by the [UK's Civil Aviation Authority](#). (Wired Blog network, Oct 2009)

Top 20 Websites

The following report shows **websites** for the industry 'All Categories', ranked by **Visits** for the week ending **06/05/2010**.

Rank	Website	Visits
1.	Facebook	8.63% 
2.	Google	7.23% 
3.	Yahoo!	3.76% 
4.	Yahoo! Mail	3.68% 
5.	YouTube	2.68% 
6.	MySpace	1.88% 
7.	msn	1.80% 
8.	Windows Live Mail	1.63% 
9.	Yahoo! Search	1.30% 
10.	Bing	1.25% 
11.	Gmail	0.86% 
12.	AOL	0.79% 
13.	eBay	0.79% 
14.	Aol Mail	0.57% 
15.	My Yahoo!	0.43% 
16.	Wikipedia	0.42% 
17.	Yahoo! News	0.39% 
18.	Amazon.com	0.35% 
19.	Pogo	0.32% 
20.	Google Maps	0.29%



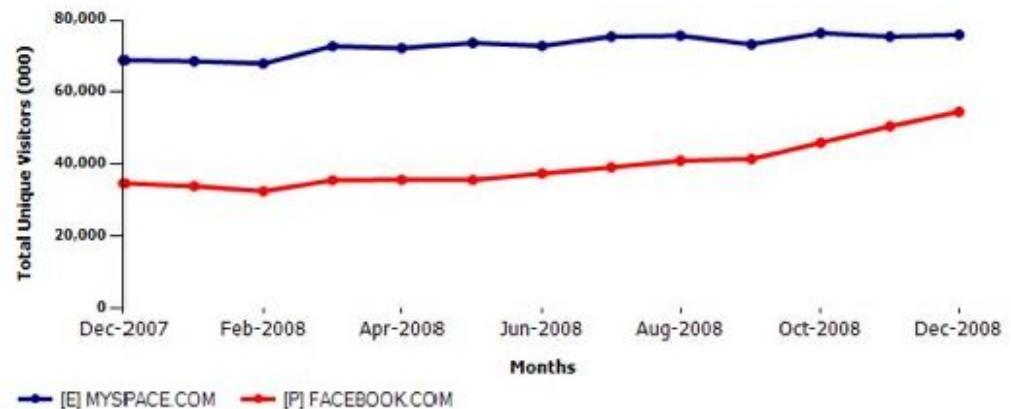
Source: The Nielsen Company/UKOM

Social Web

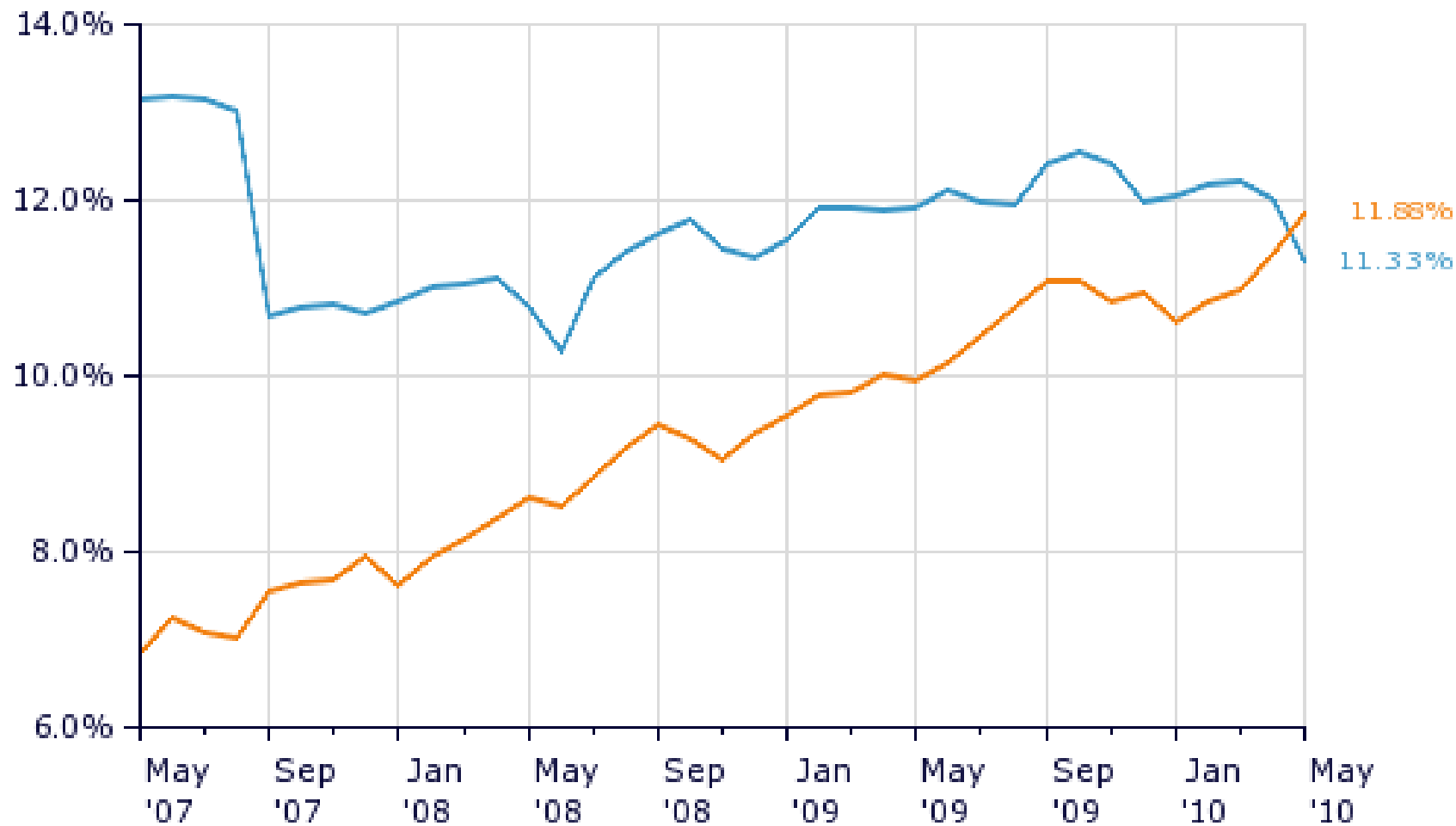
2008 U.S. Social Network Usage (Comscore)

	12/1/2007 (millions)	12/1/2008 (millions)	Yearly Growth	Monthly Growth
MySpace	69	76	10%	0.8%
Facebook	35	55	57%	3.8%
Classmates	10	16.6	66%	4.3%
LinkedIn	2.9	6.3	117%	6.7%
Bebo	NA	4.9		
Ning	0.8	3.9	388%	14.1%
Friendster	1.8	1.7	-6%	-0.5%

MEDIA TREND REPORT



UK Internet visits to Social Networks and Search Engines



- Computers and Internet - Search Engines
- Computers and Internet - Social Networking and Forums

Monthly market share in 'All Categories', measured by visits, based on UK usage.

Created: 03/06/2010. © Copyright 1998-2010 Hitwise Pty. Ltd. Source: Experian Hitwise UK



BLAH



Subject: {Disarmed} Amazon.com recommends "Collective Intelligence in .

amazon.com**Recommended for You**

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[Collective Intelligence in Action](#)



[Introduction to Statistical Relational Learning \(Adaptive Computation and Machine Learning\)](#)



[Quantitative Methods in Linguistics](#)

► [See More Recommendations](#)

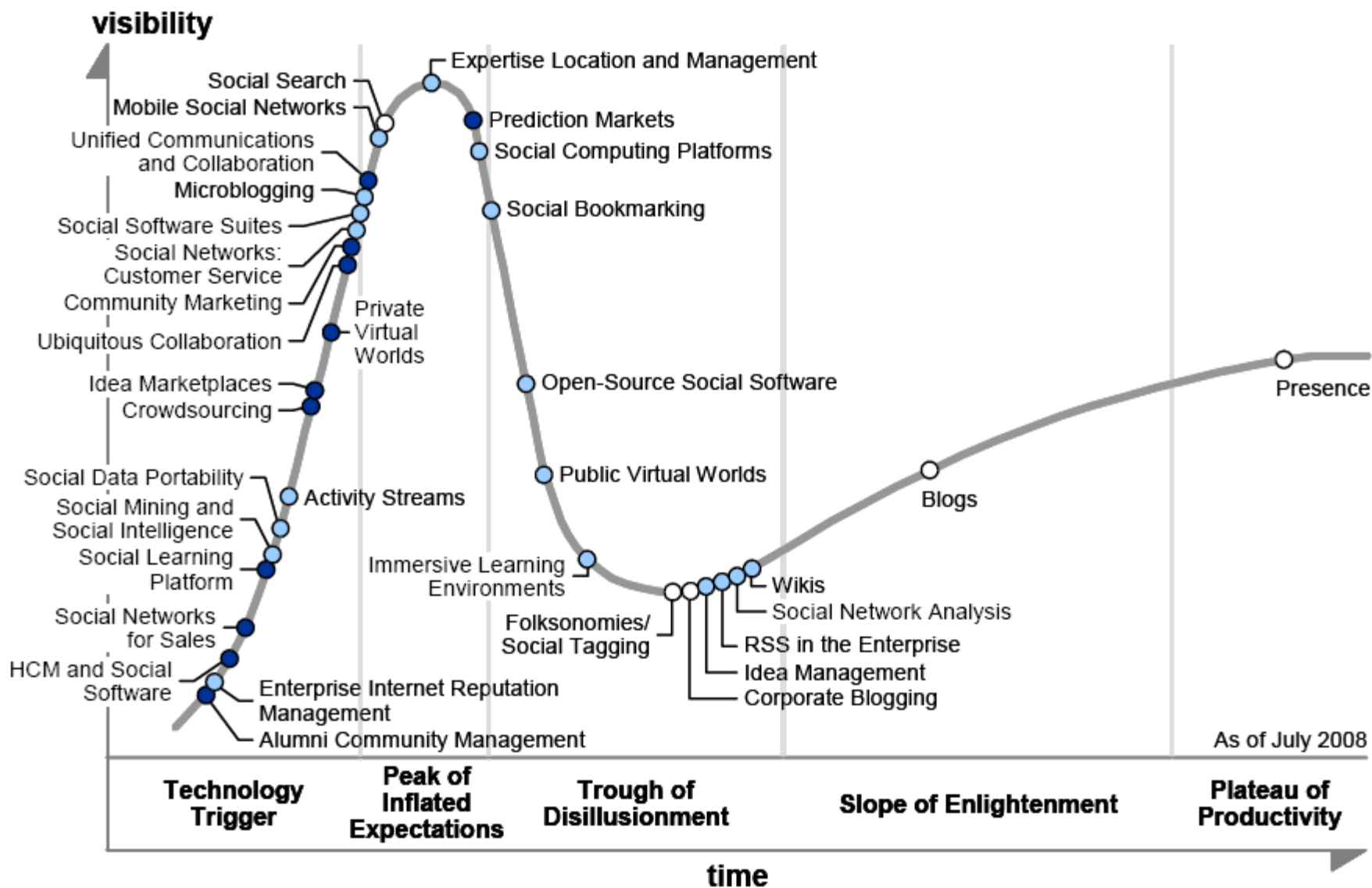


[Collective Intelligence in Action](#)
by Satnam Alag
Average customer review: ★★★★★

Add to cart

Add to Wish List

Figure 1. Hype Cycle for Social Software, 2008



Source: Gartner (July 2008)

Natural Language Parsing

UK Economy News Headlines - FT.com - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

http://www.ft.com/world/uk/economy

Più visitati Corso: Basi di dati Gruppi Posta :: Benvenuto a H... ClustrMaps - map of vi... UniversitàCedol Tree Kernels in SVM-lig... Net RicercaAteneo Keysrc Calls EMEROTECA GEMS2010

Ripr

1.00 V_Sog 1.00 V_PP 1.00 V_Adv 1.00 Adv_PP 0.50 NP_PP

Mortgage_approvals type Nom

fell type VerFin Sentence

sharply type Adv

in_June type Prep

lending_yet_more_weight type Nom

to_the_theory type Prep

Mortgage type NNP morph mas.fem.sing.

approvals type NNS morph mas.fem.plur.

in type IN morph invariante

'June' type NNP morph mas.fem.sing.plur.

lending type NN morph mas.fem.sing.

yet type RB morph invariante

more type JJR morph mas.fem.plur.sing.

Britain's place in the world, and how far it has travelled since 1947 - Jul-29

With Alex Barker and Jim Pickard

Mechanical & Electrical Engineering

Deputy Director of Finance

London Ambulance Service

RECRUITERS

Italy (Italia)

Today is: 2006-07-06 17:14:55



TIME 00:00:02:39

TG1 - 2005-10-23

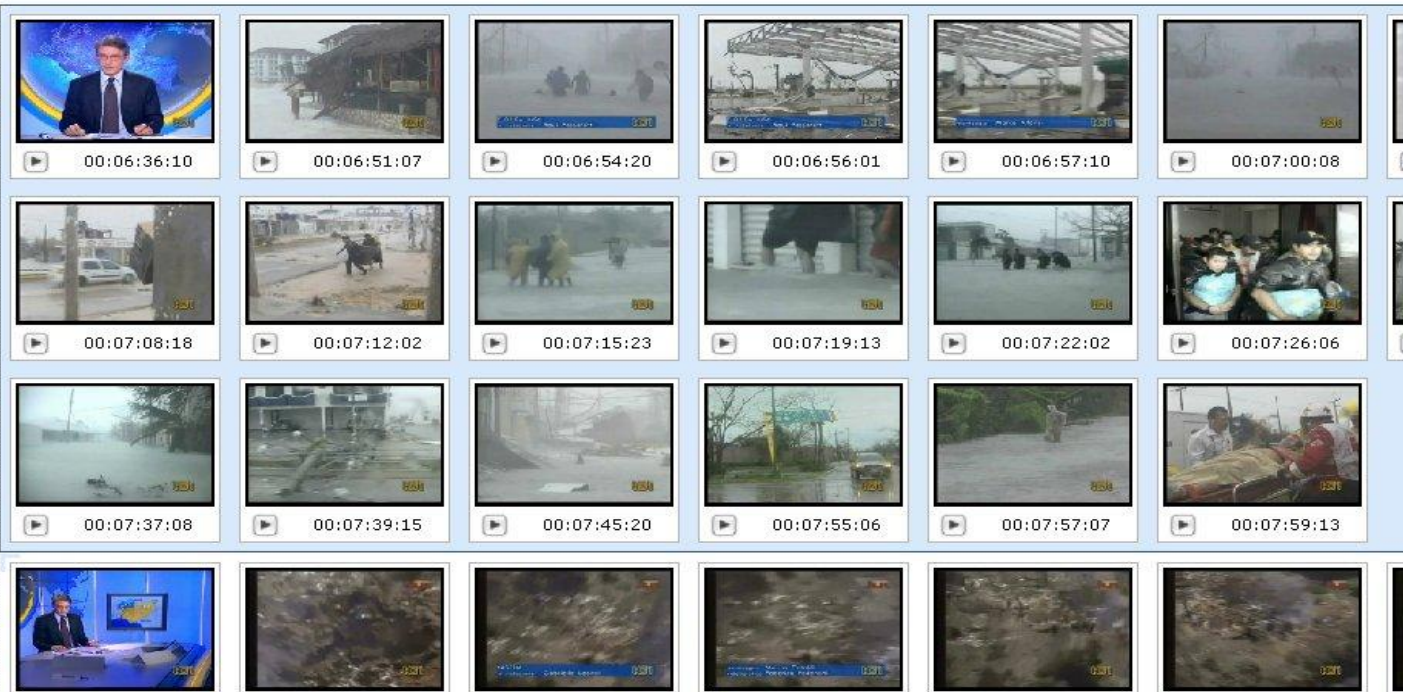
- Other Classification
- Other Classification
- Other Classification
- Ambiente, Natura e Territorio
- Ambiente, Natura e Territorio**
- Other Classification
- Politica, Partiti, Istituzioni e Sindacati
- Politica, Partiti, Istituzioni e Sindacati
- Other Classification
- Other Classification
- Other Classification
- Usi e costumi
- Other Classification
- Sanita' e Salute
- Giustizia, Criminalita' e Sicurezza
- Other Classification
- Other Classification
- Giustizia, Criminalita' e Sicurezza
- Other Classification
- Other Classification
- Musica e Spettacolo
- Sport
- Sport

Info Transcription Semantic Analysis Content Analysis

Search

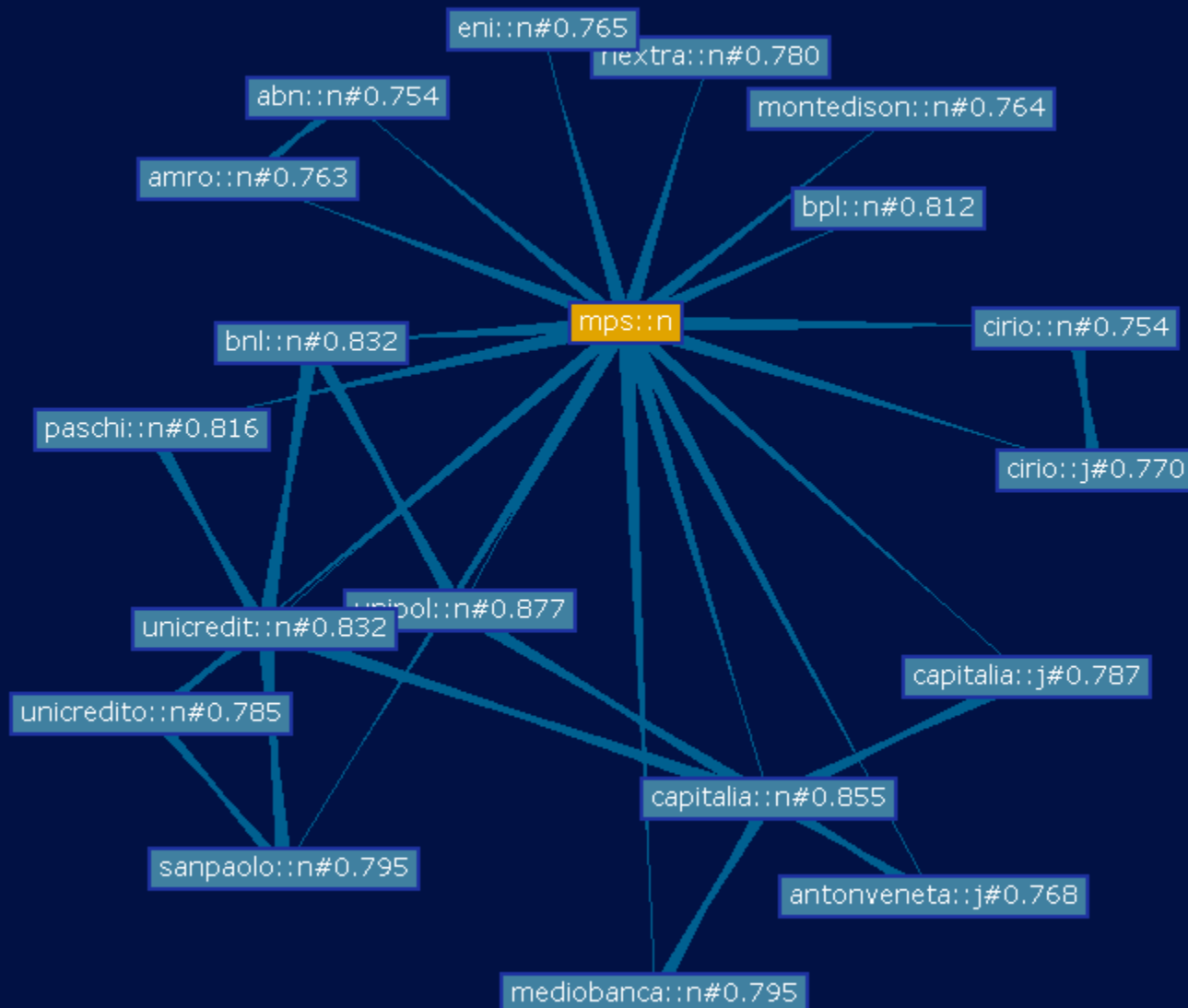
00:06:36	chamonix america dove perde forza ma fa sempre paura l' uragano di mallarme italia andiamo nel centro
00:06:41	che in florida riguardasse cento km di costa sull' atlantico si e' formata nel frattempo un' altra tempesta tropicale
00:06:52	ha lasciato una riviera messicana dello jucker puntando verso la florida l' uragano delle corde wilma il dodicesimo ciclone di una stagione eccezionale dell' atmosfera piu' di qualcuno lavatrici su strada a festeggiare lo scampato pericolo mentre dall' altra l' emergenza ha segnato l' inizio dei sacrifici scarseggiano cibo e acqua si e' costretti a fare i conti con la sopravvivenza ad attraversare queste strade inondate sferzate dal
00:07:22	vento la pioggia per raggiungere i centri della croce rossa vengono distribuiti ieri alla popolazione dino risi ma ha lasciato otto vittime soltanto migliaia di casi devastato la rete ospedaliera abbattuto centrali elettriche che ha causato danni a un milione di persone in florida e' attesa per
00:07:44	e nelle isole di kiss e' gia' iniziata la grande fuga non bastasse sull' atlantico a sud di porto rico si e' formata falla venti di hemingway le
00:07:52	nessuna tempesta tropicale della stagione la buona notizia che dovrebbe essere innocua la brutta notizia che la stagione degli uragani
00:07:59	non e' ancora finita nulla fino al trentanove e c' e' stato una sciagura
00:08:05	in nigeria

TIMELINE

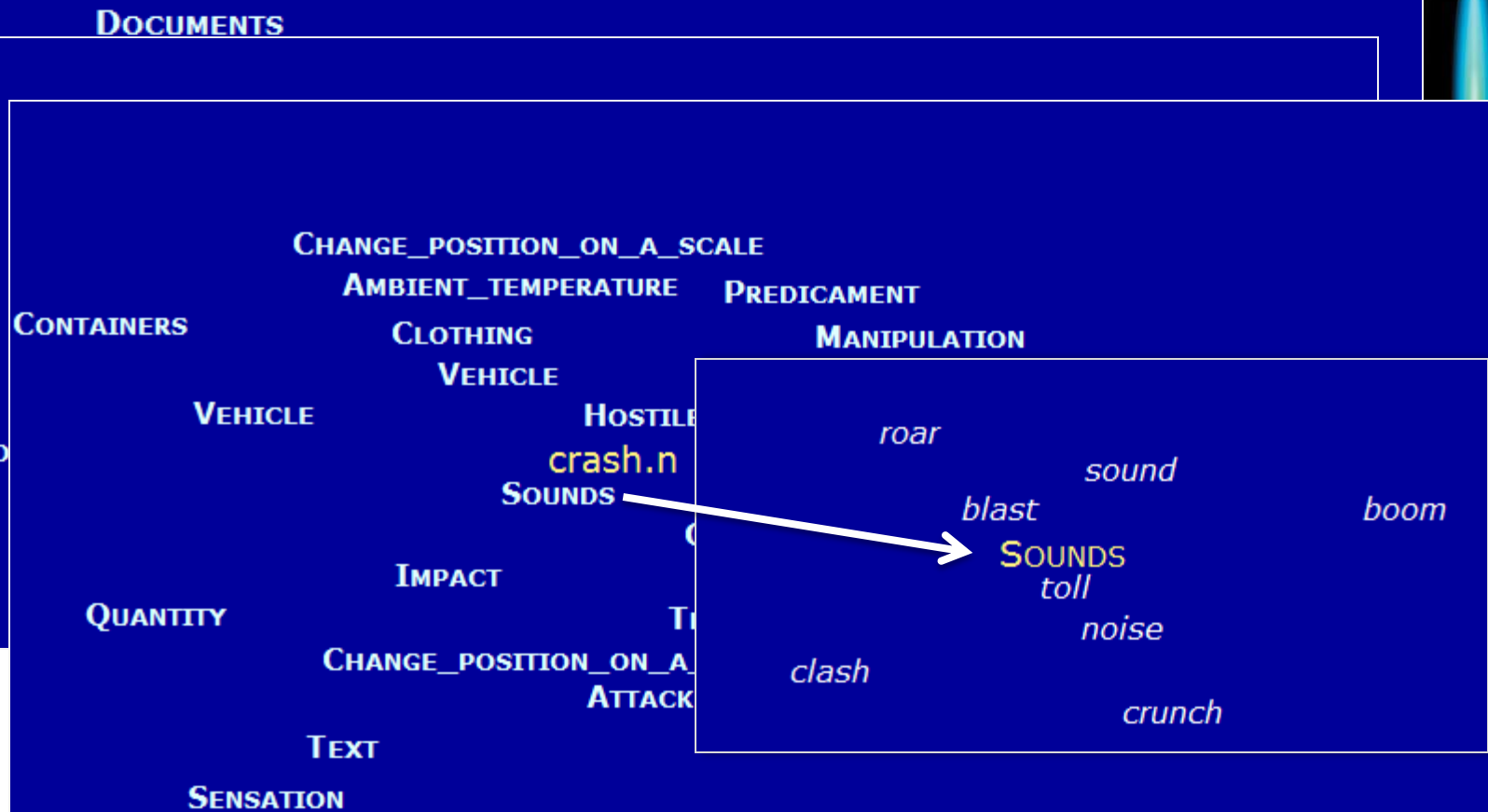


Vector Spaces for Lexical Semantics

er\wac2_w3_f200_d10000_downcase-LSA_d100.txt



Spaces for NL predicates





Home Preferences Links Documents Contact

Roberto Basili

Search

All results (99)

- » [alessandro moschitti \(27\)](#)
- » [maria teresa pazienza \(25\)](#)
- » [tor vergata \(12\)](#)
- » [department of computer science systems \(9\)](#)
- » [artificial intelligence and human oriented \(6\)](#)
- » [database of individual seismogenic \(2\)](#)
- » [vigna murata 605 it 00143 \(2\)](#)
- » [fault mapper \(2\)](#)
- » [musical genre a machine learning \(2\)](#)
- » [roberto basili fabio \(2\)](#)

[More clusters](#)

YOU ARE IN "ALESSANDRO MOSCHITTI" CLUSTER WITH 27 DOCUMENTS

DBLP: ROBERTO BASILI

Paolo Annesi, **Roberto Basili**: Cross-Lingual Alignment of FrameNet Annotations through Hidden Markov Models. ... **Roberto Basili**, Cristina Giannone, Chiara Del Vescovo, Alessandro ...

<http://dblp.uni-trier.de/db/indices/a-tree/b/Basili:Roberto.html>

PUBZONE - ROBERTO BASILI

Roberto Basili, Danilo Croce, Cristina Giannone, Diego De Cao. ... **Roberto Basili**, Cristina Giannone, Chiara Del Vescovo, Alessandro Moschitti, Paolo Naggari. Kernel-Based ...

<http://www.pubzone.org/pages/publications/showAuthor.do?userId=79.2000>

LEARNING DOMAIN-SPECIFIC FRAMENETS FROM TEXTS

Marco Pennacchiotti, Diego De Cao, Paolo Marocco, **Roberto Basili**, ... Marco Pennacchiotti, Diego De Cao, **Roberto Basili**, Danilo Croce, Michael Roth, Automatic ...

<http://olp.dfki.de/olp3/Basili.pdf>

DIEGO DE CAO HOME PAGE

Roberto Basili, Danilo Croce, Diego De Cao, and Cristina Giannone. ... **Roberto Basili**, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. ...

<http://art.uniroma2.it/decao/>

[Home](#)[Preferences](#)[Links](#)[Documents](#)[Contact](#)

All results (98)

- » [computer](#) (33)
- » [learning algorithm for machine](#) (11)
- » [machine learning course](#) (6)
- » [machine learning group](#) (6)
- » [introduction to machine learning](#) (3)
- » [machine learning applications](#) (3)
- » [artificial intelligence](#) (7)
- » [applying machine learning](#) (3)
- » [challenges machine learning](#) (3)
- » [machine learning the study](#) (5)

[More clusters](#)

TOP 98 RESULTS OF RETRIEVED FOR THE QUERY **MACHINE LEARNING**

MACHINE LEARNING - WIKIPEDIA, THE FREE ENCYCLOPEDIA

Machine learning is a scientific discipline that is concerned with ... A major focus of **machine learning** research is to automatically learn to recognize complex ...

http://en.wikipedia.org/wiki/Machine_learning

MACHINE LEARNING: DEFINITION FROM ANSWERS.COM

machine learning (m??sh?n ?l?rni?) (computer science) The process or technique by which a device modifies its own behavior as the result of its

<http://www.answers.com/topic/machine-learning>

CMU 15-859 MACHINE LEARNING THEORY

Course description: This course will focus on theoretical aspects of **machine learning**. We will examine questions such as: What kinds of guarantees ...

<http://machinelearning.com/>

THE INTERNATIONAL MACHINE LEARNING SOCIETY - ABOUT

The International **Machine Learning** Society is a non-profit organisation whose main aim is to foster **machine learning** research and whose main activity ...

<http://www.machinelearning.org/>

MACHINE LEARNING: WEEKLY STUDY GUIDE

Weekly study guide for the course on **Machine Learning** taught by Vasant Honavar at Iowa

Semantics and News

Applicazioni Risorse Sistema mar 27 lug, 23.47 dan

file:///home/danilo/Downloads/SRL_ITA/sorgente/Economia%20-%20Repubblica.it.html

Telefilm in stream... Telefilm in stream... Flash Forward pri... Telefilm in stream... Cronologia Altri Pro



L'ad punta a nuove regole sulla base del modello Pomigliano. L'annuncio, che prevede l'uscita da Federmeccanica, domani al vertice con il governo o giovedì con una lettera a Bombassei. Potrebbe avvenire assieme alla decisione di creare una new company per

Pomigliano di SALVATORE TROPEA

Cisl-Uil: "L'accordo di categoria non si tocca" di S. PAROLA

Sacconi: "Su Fiat partita aperta"

Nasce Fabbrica Italia Pomigliano



**Si dimette il capo di Bp
buonuscita un milione di sterline**

Oggi l'annuncio: a Tony Hayward subentrerà il direttore esecutivo Robert Dudley. **I costi legati al disastro sono saliti a 32,2 miliardi di dollari,** ma la società li deterrà evitando di versare al fisco Usa 10 miliardi



**Manager Usa, è Ellison di Oracle
il più pagato del decennio**

Ha guadagnato 1,84 miliardi di dollari. Nella classifica del *Wall Street Journal* sui leader delle società quotate, secondo con 1,14 miliardi il capo di Expedia, terzo Irani di Occidental Petroleum. Solo quarto Steve Jobs



Il nemico alle porte

La Consob e la mano invisibile

Altri articoli

PICCOLE GRANDI IMPRESE
DI LUCA PAGNI

La grande sfida del teleshopping

**La crisi colpisce anche i porti turistici
ma siamo sicuri che sia un male?**

Altri articoli

PERCENTUALMENTE
DI ROSARIA AMATO

La prova del 9

L'export risolveva il Pil, ma non le famiglie

Altri articoli

GLI ESPERTI RISPONDONO

CASA
A cura di Antonella Donati

Compenso extra, quando ne ha diritto l'amministratore

Mia moglie ed il fratello sono proprietari di un appartamento in condominio. Allo stato

24ORE AGI

Roma 19:04
ACEA: NEL I SEMESTRE UTILE NETTO +52,1% A 82, MLN

Parigi 18:42
AIR FRANCE-KLM: TORNA IN UTILE NEL PRIMO TRIMESTRE

Le altre not

CREDITO ALLE IMPRESE

Microimprese: con la crisi aumenta il rischio di credito

IN COLLABORAZIONE CON



Semantic [role] LABELING



I costi legati al disastro sono saliti a 32,2 miliardi di dollari.

Submit

Clear

Change_position_on_a_scale: *[I costi]_{Item} legati al disastro sono saliti [a 32,2 miliardi di dollari.]_{Final_value} .*

[Show CONLL format](#)

1	I	RD	det	2	_	Item
2	costi	S	subj	7	_	Item
3	legati	V	mod	2	_	_
4	al	EA	comp	3	_	_
5	disastro	S	prep	4	_	_
6	sono	VA	aux	7	_	_
7	saliti	V	ROOT	0	Change_position_on_a_scale	Target
8	a	E	comp	7	_	Final_value

Completato

- I costi legati al disastro sono saliti a 32,2 miliardi di dollari.

Semantic Role Labeling

Architecture | How To | Know-All Service | **Semantic Role Labeling Service** | Credits



Semantic [role] LABELING

L'aumento dei costi legati al disastro è stato di 32,2 miliardi di dollari.

Submit

Clear

Change_position_on_a_scale: L'aumento [dei costi]_{Item} legati al disastro è stato [di 32,2 miliardi di dollari]_{Final_value}.

Telegraph.co.uk

SEARCH

[Home](#) [News](#) [Sport](#) [Finance](#) [Lifestyle](#) [Comment](#) [Travel](#) [Culture](#) [Technology](#)
[UK](#) [World](#) [Politics](#) [Celebrities](#) [Obituaries](#) [Weird](#) [Earth](#) [Science](#) [Health News](#) [Education](#)
[USA](#) [Barack Obama](#) [Asia](#) [China](#) [Central Asia](#) [Europe](#) [Australasia](#) [Middle East](#) [Africa](#)

SRL_EN - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

http://10.6.0.103:3320/Know_All/srl_en.html

Google

[UtiliLinks](#) [Più visitati](#) [Corso: Basi di dati](#) [Gruppi](#) [Posta :: Benvenuto a H...](#) [ClustrMaps - map of vi...](#) [UniversitaCedol](#) [Tree Kernels in SVM-lig...](#) [Net](#) [RicercaAteneo](#) [Keysrc](#) [Calls](#) [EMEROTECA](#) [GEMS2010](#)

[URGENT - Towards...](#) [Inbox | LinkedIn](#) [Italy news, all the la...](#) [SRL_EN](#) [AMI2 - Italian Fram...](#) [Satoshi Sekine: Co...](#) [Deposito File: Web ...](#) [Deposito File: Web ...](#)

[Architecture](#) [How To](#) [Know-All Service](#) [Semantic Role Labeling Service](#) [Credits](#)



Semantic [role] LABELING


 ARTIFICIAL INTELLIGENCE
 AT ROMA TOR VERGATA

Silvio Berlusconi won a crucial confidence vote in parliament amid accusations that his party bought the support of some opposition with money and political favours.

Finish_competition: [Silvio Berlusconi]_{Competitor} **won** [a crucial confidence vote in parliament]_{Competition} amid accusations that his party bought the support of some opposition with money and political favours.

Importance: Silvio Berlusconi won [a **crucial** confidence vote in parliament]_{Factor} amid accusations that his party bought the support of some opposition with money and political favours.

Commerce_buy: Silvio Berlusconi won a crucial confidence vote in parliament amid accusations that [his party]_{Buyer} **bought** [the support of some opposition with money and political favours]_{Good}.

[Show CONLL format](#)

Question Answering

Sistema di Prenota... Know-All AMI2 - Italian Fram... Satoshi Sekine: Co... Deposito File: V

Architecture How To **Know-All Service** Semantic Role Labeling Service Credits

Know-All

1.0b

Where is Taj Mahal?

Semantic Role Labeling: disabled: ☒ enabled: ☐

Search Engine I: ☒ Search Engine II: ☐

Identifinder: ☒ Chaos: ☐

Submit

Clear

Answer 0001 (offers by Identifinder)

0001: India

0002: Agra

0003: Brooklyn

Download Results

Know-All status: idle

References

- Mitchell, Tom. M. 1997. *Machine Learning*. New York: McGraw-Hill.
- [Kernel machines, neural networks and graphical models](#), P. Frasconi, A. Sperduti, A. Starita, Rivista AI*IA Numero speciale per i “50 anni di IA”, 2007.
- Nice Video lectures by Andrew Ng (Stanford)
<http://academicearth.org/courses/machine-learning>