



BIG DATA ANALYTICS

Chairman: Prof. Gianluca Cubadda (Università degli Studi di Roma "Tor Vergata")



h. 14:00 "Big Data Analytics e machine learning"

Relatore: Dott. Marco Scarnò (Senior Researcher CINECA)



Marco Scarnò (m.scarno@cineca.it)



<http://www.dialogic.nl/documents/2010.080-1226.pdf>

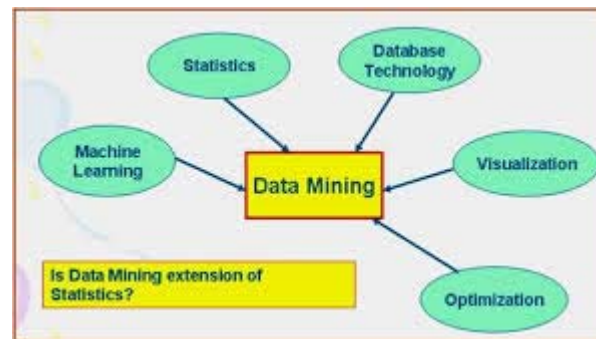
Internet as data source

Feasibility Study on Statistical Methods on Internet as a
Source of Data Gathering

Nel 2013 abbiamo cominciato a «lavorare» considerando il World Wide Web come una possibile fonte informativa capace di fornire dati statistici relativi a particolari fenomeni.



Un tempo...



Con l'idea che il data base era *proprietario* (comunque di grandi dimensioni, spesso sparso e a volte non strutturato)

Ora... il riferimento è divenuto quanto viene identificato con il termine Big Data, ossia tutta quell'informazione sparsa, non strutturata, volatile, di grandi dimensioni, ecc.

Ma, in questo caso comunque, si tratta d'informazione, più o meno, liberamente accessibile

Regulation (EU) No 99/2013 of the European Parliament and of the Council of 15 January 2013 on the European statistical programme 2013-17, OJ L 39, 9.2.2013, p. 12–29

Scheveningen Memorandum
Big Data and Official Statistics

The DGINS

CONSIDERING

1. Recent innovations in the information and communication technologies have been leading to an increasing degree of digitization of economies and societies at all levels that offer new opportunities for the compilation of statistics.
2. The use of Big Data for statistical purposes challenges the European Statistical System to effectively address a variety of issues.
3. The demand for timely and cost efficient production of high quality statistical data increases, as well the need for new solutions to declining response levels.
4. Official statistics should incorporate as much as possible all potential data sources, including Big Data, into their conceptual design.
5. The distinguishing aspect of many Big Data sources is that they are not confined to national borders and, as such, represent unique opportunities for collaboration at European level as well as on global level.
6. Many European initiatives have a link to Big Data, including the European Commission's ambition for developing a strategy for the European data value chain, the on-going EU Data Protection reform and the Horizon2020 program.
7. The implementation of new methods of production of European statistics represents an objective of the European Statistical Programme 2013-2017¹ and aims at efficiency gains and quality improvements, including increased timeliness.



1. Acknowledge that Big Data represent new opportunities and challenges for Official Statistics, and therefore encourage the European Statistical System and its partners to effectively examine the potential of Big Data sources in that regard.
2. Recognise that Big Data is a phenomenon which is impacting on many policy areas. It is therefore essential to develop an 'Official Statistics Big Data strategy' and to examine the place and the interdependencies of this strategy within the wider context of an overall government strategy at national as well as at EU level.
3. Recognise that the implications of Big Data for legislation especially with regard to data protection and personal rights (e.g. access to Big Data sources held by third parties) should be properly addressed as a matter of priority in a coordinated manner.
4. Note that several NSIs are currently initiating or considering different uses of Big Data in a national context. There is a momentum to share experiences obtained from concrete Big Data projects and to collaborate within the ESS and beyond, on a global level.
5. Recognise that developing the necessary capabilities and skills to effectively explore Big Data is essential for their integration into the European Statistical System. This requires systematic efforts like appropriate training courses and establishing dedicated communities including academics for sharing experiences and best practice.
6. Acknowledge that the multidisciplinary character of Big Data requires synergies and partnerships to be effectively built with experts and stakeholders from various domains including government, academics and owners of private data sources.
7. Acknowledge that the use of Big Data in the context of official statistics requires new developments in methodology, quality assessment and IT related issues. The European Statistical System should make a special effort to supports these developments.
8. Agree on the importance of following up the implementation of this memorandum by adopting an ESS action plan and roadmap by mid-2014 that should be further integrated into the Statistical Annual Work Programmes of Eurostat.

Ricordi «universitari»

Nel PARADIGMA INFORMATIZIONE
 il Processo di Analisi
Induttiva si realizza
 tramite meccanismi di
combinazione di diverse
FONTI INFORMATIVE.

“PRODUZIONE DI INFORMA-
 ZIONE A MEZZO DI
 INFORMAZIONE”

R. Coppi, Appunti sul Data Mining

Ciò con l'obiettivo di ridurre l'incertezza iniziale su un determinato fenomeno reale

$$I_{add}=F(I_i, I_e)$$

- I_i =Informazione iniziale
- I_e =Informazione empirica
- I_{add} =Informazione addizionale
- F =Funzione (algoritmo di apprendimento, cluster, ecc.)

$$I_{add}=$$

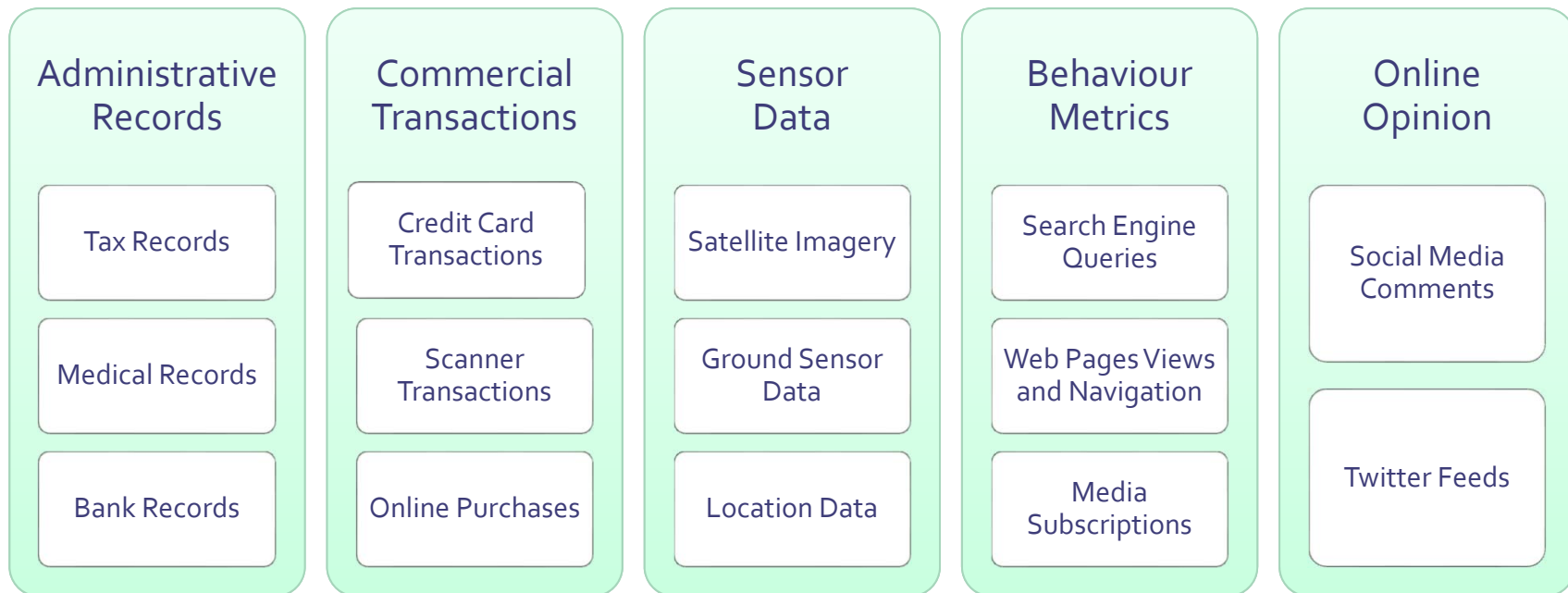
Unità di produzione (UP)	risparmio medio (RM)	costo medio (CM)	costo medio (CM)
1	10	10	10
2	8	14	8
3	6	24	6
4	7	28	4
5	6	30	2

$$+ F(I_e)$$



Possibili fonti *BIG DATA* per l'Informazione Empirica (e loro categorizzazione)

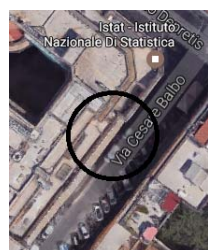
Nota: BIG non riferisce necessariamente al «volume»; più in generale si tratta di dati «non strutturati»



Questioni aperte:

- Trasformare l'Informazione empirica in dati statistici strutturati
- Definire un nuovo paradigma inferenziale

Un esempio: immagini satellitari



Pixels



Polygons

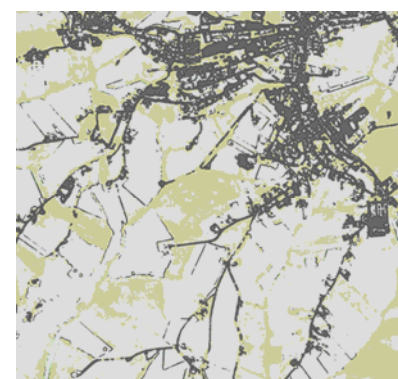
Variabili che rappresentano la luminosità dei pixel

Unit	$\underline{X}$	$\underline{Y}$
1	$\underline{X}_1$	$\underline{Y}_1$
...		
n	$\underline{X}_n$	$\underline{Y}_n$

Y: land cover
derivante da
sorgenti esterne
(registri, indagini,
ecc.)

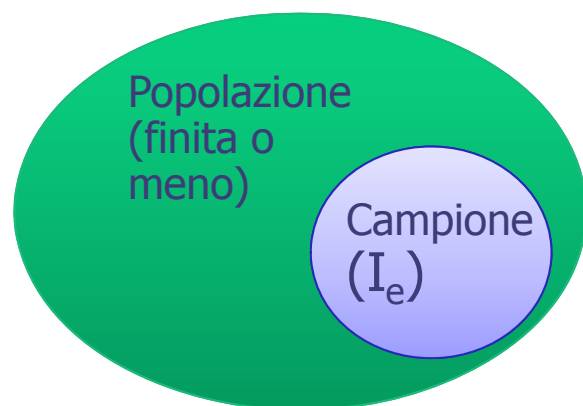


- 25 mila immagini per l'Italia scaricate da Google;
- Conoscenza del land cover per il punto centrale di ogni immagine;
- Regressione logistica lineare per "addestrarsi" sul riconoscimento

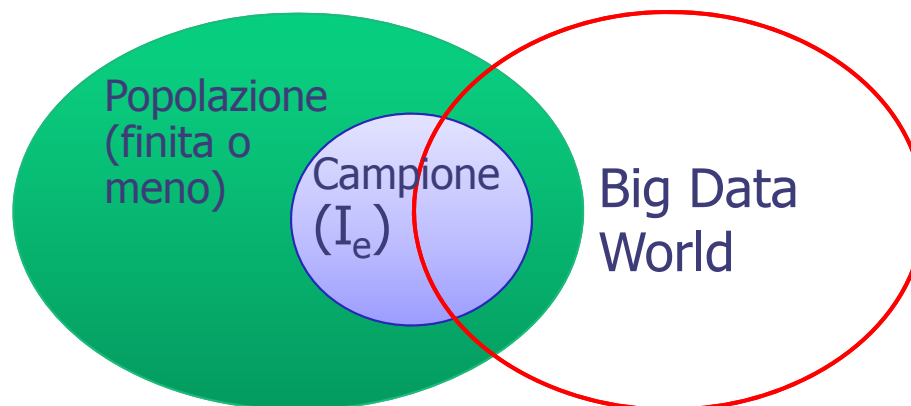




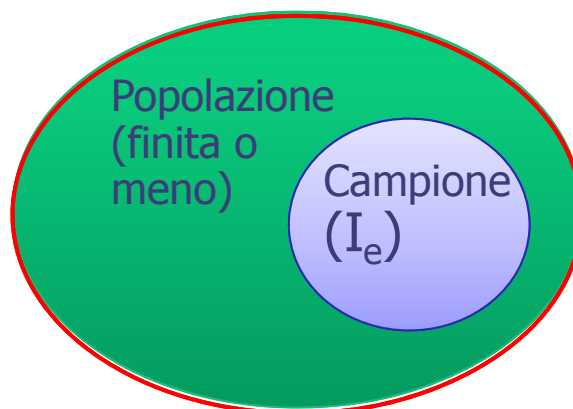
Il contesto inferenziale



Il paradigma classico



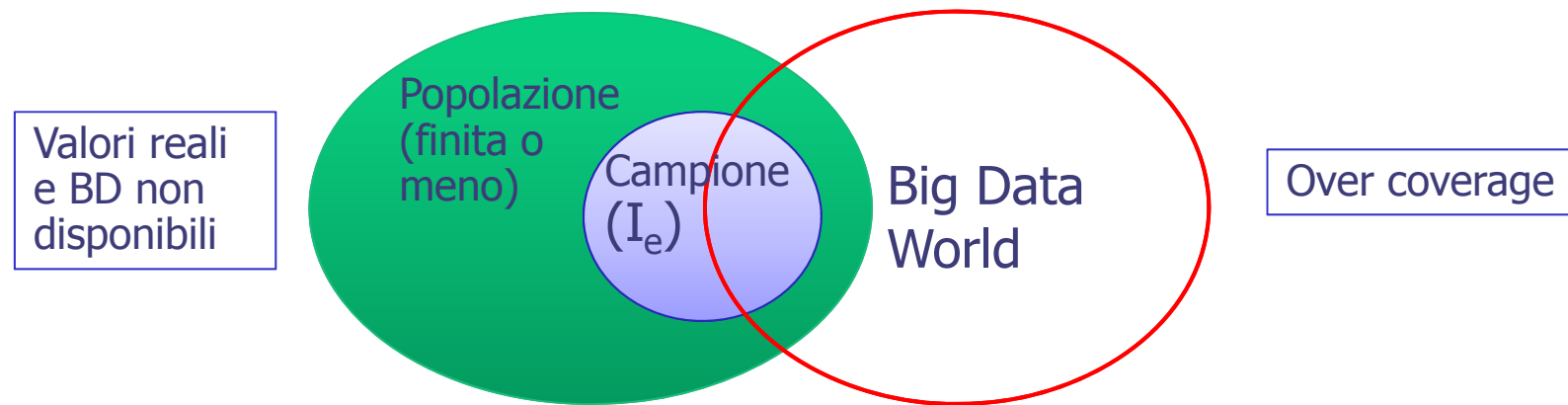
Il paradigma Big Data con overlap parziale
(E-commerce, mobile phone data, Social media, satellite images, etc.)



Il paradigma Big data con overlap perfetto
(Satellite images per un determinato livello di risoluzione e per un problema specifico)



Il dettaglio del contesto inferenziale nei BD



Fonti di distorsione (bias):

- Coverage bias – la popolazione Big non è quella d'interesse
- Self selection bias – I BIG data possono non essere rappresentativi della vera popolazione
- Measurement error – I dati rilevati potrebbero non riferirsi all'oggetto di studio
- L'incremento d'informazione acquisita NON reduce gli errori non-campionari

Esempio 1: sostituzione indagine e ampliamento delle informazioni sulle aziende (con ISTAT)

Archivio (ASIA) delle informazioni sulle aziende (con più di 10 dipendenti)

**RILEVAZIONE SULLE
TECNOLOGIE DELL'INFORMAZIONE
E DELLA COMUNICAZIONE (ICT)
NELLE IMPRESE - ANNO 2018**

C8 Indicare quali servizi offre il sito Web dell'impresa:

Rispondere ad ogni riga

	<i>Si</i>	<i>No</i>
a. Possibilità di effettuare ordinazioni o prenotazioni on-line (es. carrello della spesa on-line).....	☐	☐
b. Tracciabilità on line dell'ordine.....	☐	☐
c. Accesso a cataloghi di prodotti o listini prezzi.....	☐	☐
d. Possibilità di personalizzare i contenuti del sito per i visitatori abituali.....	☐	☐
e. Possibilità per i visitatori del sito di personalizzare o progettare prodotti.....	☐	☐
f. Avvertenze sulla politica in materia di privacy, marchio di certificazione della tutela della privacy o certificazione della sicurezza del sito.....	☐	☐
g. Annuncio di posti di lavoro vacanti o possibilità di effettuare domande di impiego on line.....	☐	☐
h. Collegamenti o riferimenti ai profili dell'impresa sui <i>social media</i>	☐	☐
i. Possibilità di presentare on-line dei reclami (tramite email, modulo web, ecc.).....	☐	☐

Campione di circa 12 mila aziende

Scopo del progetto:

- Verificare la possibilità di sostituire l'indagine con una raccolta diretta sul web;
- Generalizzare la stima a tutte le aziende per le quali è noto il sito web

Esempio 2: arricchimento archivio agriturismi (con ISTAT)

I_i =Informazione iniziale I_e =Informazione empirica

Regione	Denominazione	Provincia	+	Numero camere	Prezzo
Abruzzo	AZ. AGRIT. RUELLA	L'Aquila		30	10
Abruzzo	AZ. AGR. LA VILLA	L'Aquila		30	20
Abruzzo	CASALE DELLE CAMPITELLE	L'Aquila		6	10
Abruzzo	L'APE	L'Aquila		0	10
Abruzzo	CUPELLO	L'Aquila		14	20
Abruzzo	IL CAPRICCIO DI GIOVE	L'Aquila		12	10

Scopo del progetto:

- Arricchire l'archivio con le informazioni sugli agriturismi (di natura amministrativa);
- Verificare la presenza di strutture non adeguate e proporre aggiornamenti degli archivi;
- Produzione di nuove statistiche (anche sulla base dei prezzi)

Esempio 3: estrazione del Sentiment da testi (Banca D'Italia)

ULTIME NOTIZIE SU BANKITALIA

NEWS24 | 22 settembre 2018

Visco: il rapporto debito-Pil rischia l'insostenibilità

Così il Governatore di *Bankitalia* Ignazio Visco al 64/o Convegno della Corte di Conti a Varenna (Lecco).
– di Nicola Barone

AGORA | 21 settembre 2018

Di Maio dà buca al convegno del private equity e venture capital: il deputato M5S Carabetta per capire cosa succederà al settore

C'è attesa, nel mondo del private equity e del venture capital, per capire quale è il 5Stelle-pensiero sul capitale di rischio. Così si intensificano i contatti fra l'Aifi, che rappresenta il settore in Italia, con a capo Anna Gervasoni e Innocenzo Cipolletta, ed esponenti del Governo giallo-verde. E' inutile dirlo. Fra i maggiori fondi di private equity e di venture capital c'è grande preoccupazione, all'interno del nuovo orizzonte politico Lega - Stelle, per le sorti di un settore che continua...

– Carlo Festa

NEWS24 | 21 settembre 2018

Bankitalia, nuovo mandato da vice dg per Panetta. Ritratto dell'economista e divulgatore

Già nel 2011, con Draghi ancora governatore di via Nazionale, gli fu affidato il compito di coordinare le attività connesse alla partecipazione di *Bankitalia* all'Eurosistema e l'analisi della stabilità finanziaria; e durante la permanenza al... Poi, quando per il nostro Paese si è posta l'esigenza che le banche accelerassero il recupero dei non performing loans, nei suoi interventi il dirigente di *Bankitalia* ha spiegato che esiste un limite alla velocità di questo rientro, oltre il...

– di Rossella Bocciarelli

Bollettino Economico



Condividi     

(trimestrale - date di pubblicazione per il 2018: 19 gennaio, 13 aprile, 13 luglio e 19 ottobre)

vai al livello superiore
PUBBLICAZIONI

ARCHIVIO

2018	2017	2016	2015	2014
2013	2012	2011	2010	2009
2008	2007	2006	2005	2004
2003	2002	2001	2000	1999

► Anni precedenti

Scopo del progetto (poi divenuto uno strumento «general purpose» che consente di analizzare i testi identificando un Sentiment):

- Verificare la «percezione» dell'Istituzione;
- Identificare «correlazioni» tra Sentiment e altre variabili;
- Identificare i termini/concetti rilevanti (topic analysis).

Esempio 4: estrazione del Sentiment da Twitter-FB-Tripadvisor (Reggia di Caserta)

```

"obs","text","favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","replyToUID","statusSource","screenName","retweetCount","isRetwee
"1","Pasta Reggia Caserta-Sidigas Avellino: streaming e diretta live https://t.co/UMN8XJr5vc","FALSE","0",NA,"2017-01-01 10:01:03","FALSE",NA,"815498059073208
"2","Capodanno alla Reggia di Caserta - Blasting News https://t.co/MKclUCntEo #Caserta #news","FALSE","0",NA,"2017-01-01 09:57:33","FALSE",NA,"815497180035354
"3","Maria Carolina attente il visitatore alla Reggia di Caserta per la vestizione. Due gli eventi in programma - #Napoli è su NapoliTime - Eveni","FALSE","0
"4","Campania: #Reggia di #Caserta, interni Foto di m0nicu #regg... (volgocampania) https://t.co/ju0M8onqEf","FALSE","0",NA,"2017-01-01 07:56:31","FALSE",NA
"5","Pasta Reggia Caserta - Sidigas Avellino - https://t.co/Mz0Dffxx0J https://t.co/GTfyQnZ17z #pastachannel","FALSE","0",NA,"2017-01-01 01:06:10","FALSE",NA
"6","Reggia di Caserta. Palace of Caserta #happynewyear #flight #picoftheday #caserta #reggia #napoli https://t.co/qohUKc2hs3","FALSE","1",NA,"2016-12-31 19:
"7","Maria Carolina attente il visitatore alla Reggia di Caserta per la vestizione. Due gli eventi in programma https://t.co/Dx7n1DAUd4","FALSE","0",NA,"2016-
"8","Luci di Salerno e Reggia di Caserta https://t.co/2KC2G7nLV3","FALSE","0",NA,"2016-12-31 19:13:29","FALSE",NA,"815274696707080192",NA,"<a href= http://www
"9","31 dicembre 2016 **** Reggia di Caserta **** Felice Anno Nuovo!!!!","FALSE","0",NA,"2016-12-31 18:55:29","FALSE",NA,"815270168670728193",NA,"<a href= h
"10","Reggia di Caserta. Palace of Caserta #happynewyear #flight #picoftheday #caserta #reggia #napoli https://t.co/rsRphRCaR0","FALSE","1",NA,"2016-12-31 18
"11","Just posted a photo @ Reggia di Caserta https://t.co/UI05Bluna1","FALSE","0",NA,"2016-12-31 17:15:13","FALSE",NA,"815244932868804608",NA,"<a href= http:
    
```



- Scopo del progetto:
- Verificare la «percezione» dell'Istituzione;
 - Identificare punti di forza e debolezza;
 - Marketing territoriale.

Esempio 5: classificazione automatica documenti (SOGEI)

Dato un contenzioso del quale è noto la «questione controversa» trattata...

direzione provinciale i di milano _____ ufficio controlli milano , 13 ottobre 2011 alla commissione tributaria provinciale di milano
 prot . n . 2011 287734 controdeduzioni ricorso n . 1045428 - t9b notificato in data 30 06 2011 e ricevuto dall'ufficio in data 04 07 2011 [...] ritenendo
 di poter beneficiare della disposizione di cui all'art . 15 dlgs 218 1997 [...] dell'art . 15 dlgs 218 1997 , [...] il periculum in mora l'art . 47 del dlgs 546
 1992 richiede espressamente [...]

Studiare un metodo in grado di riconoscere automaticamente la «questione controversa» e
 altre eventuali

Trovati riferimenti a: 3 questione/i controversa/e

Dettaglio (distanza, descrizione, macroarea, area):

0.29, Definizione delle sanzioni a seguito di rinuncia all'impugnazione dell'avviso di accertamento o liquidazione ai sensi dell'art. 15 DLGS n. 218 del 1997 - Inammissibilità del ricorso successivamente proposto., SANZIONI TRIBUTARIE, DISPOSIZIONI GENERALI

0.24, Sospensione dell'atto impugnato - Presupposti. Art. 47 DLGS n. 546 del 1992, PROCESSO TRIBUTARIO, GIUDIZIO IN CTP

0.15, Riduzione delle sanzioni per omessa impugnazione. Art. 15 DLGS n. 218 del 1997, SANZIONI TRIBUTARIE, DISPOSIZIONI GENERALI

Perché gli esempi citati sono progetti Big Data?

Usano informazioni non strutturate:

- Pagine web, tra cui:
 - Siti;
 - Aggregatori;
 - Motori di ricerca;
- Social media;
- Documenti o testi.

E il machine learning?

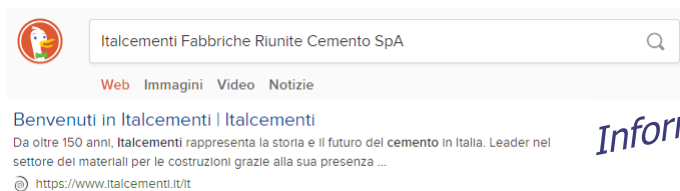
Algoritmi di classificazione supervisionata (estrazione di regole di classificazione)

Algoritmi di classificazione non supervisionata (topic analysis)

Ricerca «intelligente» delle informazioni rilevanti (sfruttando i risultati dei motori di ricerca)

La strategia per l'uso dell'Informazione dal WWW

Uso automatico motori di ricerca (DuckDuckGo, Bing, Istella, *no Google*)



Informazione semi-strutturata

*Tramite metodo
addestrato a
riconoscere se il
risultato coincide
con quello cercato*

Estrazione informazione da siti HUB



Informazione semi-strutturata



Estrazione informazione da siti generici



Informazione non strutturata



La strategia per il trattamento preliminare dei testi

- Eliminazione/sostituzione caratteri non validi;
- Riconoscimento frasi;
- Identificazione parole;
- Lemmatizzazione;
- Selezione lemmi.

document_reference	document_content
2008_01	...nica. Le tensioni sui mercati hanno registrato andamenti alterni, reagendo al susseguirsi di notizie sulle perdite di alcuni intermediari, sul la caduta del mercato immobiliare statunitense, sulla crescita ec...
2008_02	...nto del credito bancario negli Stati Uniti 12 Il recente rincaro delle materie prime alimentari 17 L'indebitamento delle famiglie italiane 23 La recente revisione statistica dei valori medi unitari e delle quantità...
2008_03	...a 24 L'offerta e la domanda di credito in Italia 30 Le ipotesi alla base del quadro previsionale 33 Il documento di programmazione economico-finanziaria per gli anni 2009-2013 e la manovra triennale 36 ADV...
2008_04	...elaborazioni, salvo diversa indicazione, sono eseguite dalla Banca d'Italia. Segni convenzionali: - il fenomeno non esiste; il fenomeno esiste, ma i dati non si conoscono; ... i dati non raggiungono la cif...
2009_01	...ncario nei principali paesi industriali 10 Le misure adottate in Italia per la stabilità del sistema finanziario 31 Il mercato interbancario collateralizzato 33 L'offerta e la domanda di credito in Italia 36 Il decret...

vardescriptor	word	type	lemma
2014_01_32	insieme	ADV	insieme
2014_01_32	con	PRE	con
2014_01_32	le	DET:DEF	il
2014_01_32	autorità	NOM	autorità
2014_01_32	di	PRE	di
2014_01_32	vi	PRO:DEMO	vi

Original record descriptor	Sentence (rebuilt from words in different records)
2008_01_14	rispettare autunno sviluppo stare determinando revisione ribasso previsione crescita rialzo inflazione
2008_01_15	peggioramento prospettive anno corso avere riguardare soprattutto stato unire riserva federale avere reagire tre successivo riduzione tasso ufficiale sce coi
2008_01_16	ulteriore riduzione essere attendere operatore mercato prossimo riunione federal open market committee fomc
2008_01_17	area euro aumento attività produttivo essere superiore crescita potenziale medio anno ma decelerazione quarto trimestre
2008_01_18	quadro previsionale eurosistema pubblicare scorso dicembre segnalare crescita appena linea potenziale inflazione rimanere superiore cento gran parte anno

I risultati

Esempio 2: arricchimento archivio agriturismi (con ISTAT)

Conference Paper Full-text available

ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms

October 2016

DOI: 10.1481/icasVII.2016.f29d

Conference: Modernization of Agricultural Statistics in Support of the Sustainable Development

Agenda · At: FAO Headquarter, Rome

Project: [Agricultural Statistics](#)

Giulio Barcaroli · Daniela Fusco · Paola Giordano · [Show all 7 authors](#) · Marco Scarnò

Table 1 – *main results obtained after the execution of the query related to the search of each AF*

Result of the query	Frequency
Not found by the search engine	1,737
Found by the search engine but the retrieved URLs are not referred neither to Tripadvisor nor as a correct link to the AF specific website	10,202
URLs recognized as correct specific websites of AFs not found in TripAdvisor	1,354
AF found in Tripadvisor, without a recognized specific website	3,590
AFs with both recognized specific websites and found in TripAdvisor	1,935
Total	18,818

The total number of AFs for which it is possible to gather information by the Internet is therefore 6,879 (36.6% of the total).

Table 2 – *Confusion matrix related to the availability of restaurant as in the master data set and as recognized by referring to the AFs identified only in the hub, only by considering their web sites or jointly*

Presence of restaurant as in the master dataset	Information derived only from y the hub		Information derived from the specific websites		Information derived both from the hub and from specific websites	
	No	Yes	No	Yes	No	Yes
No	1667	640	537	475	325	812
Yes	741	542	148	194	141	657
<i>Rate of success</i>	61.5% (100*2209/3590)		50.6% (100*685/1354)		50.7% (100*982/1935)	

Informazioni spesso non coerenti tra archivio e siti; circa 1/3 le aziende identificate sul web



I risultati

Esempio 3: estrazione del Sentiment da testi (Banca D'Italia) (1/2)

- Un insieme di procedure per il trattamento del testo;
- Un nuovo dizionario per il Sentiment (costruito mediante un algoritmo «di apprendimento» capace di calcolare peso e intensità dei termini)

Nota: i pesi e le intensità permettono di centrare il termine rispetto i sinonimi e i contrari (al primo passo i pesi sono stati assegnati sia usando precedenti valutazioni, sia manualmente)

La formula per il Sentiment= $100 \cdot \frac{\sum W_{PT}}{\sum W_{PT} + \sum W_{NT}}$

Il test del nostro dizionario considerando circa 4000 mila giudizi da Amazon (già classificati in *positivi* e *negativi*) e il confronto con il dizionario OpenER

Table 2: Confusion matrix obtained by testing our dictionary on pre-classified sentences

Our dictionary		Classification using our dictionary			
		Negative	Positive	Not assigned	Total
Initial classification	Negative	754	852	10	1616
	Positive	197	2111	9	2317
	Total	951	2963	19	3933

Table 3: Confusion matrix obtained by testing the OpenER dictionary on pre-classified sentences

OpenER dictionary		Classification using OpenER dictionary			
		Negative	Positive	Not assigned	Total
Initial classification	Negative	927	676	13	1616
	Positive	426	1865	26	2317
	Total	1353	2541	39	3933

OpenER: Open Polarity Enhanced Name Entity Recognition" project, funded by the European Commission under the 7th Framework Program (FP7)

From Tables 2 and 3 we notice that:

- for our dictionary: $CR = 100 \cdot \frac{(f_{nn} + f_{pp})}{\sum_{i,j} f_{i,j}} = 100 \cdot \frac{(754+2111)}{3914} = 73.2$ and $\phi^2 = 0.19$
- for the OpenER dictionary: $CR = 100 \cdot \frac{(927+1865)}{3894} = 71.7$ and $\phi^2 = 0.16$

Valori ottenuti con l'uso anche delle negazioni e degli intensificatori



- for our dictionary: $CR = 100 \cdot \frac{(966+2123)}{3914} = 78.9$, and $\phi^2=0.32$
- for the OpenER dictionary: $CR = 100 \cdot \frac{(1010+1864)}{3894} = 73.8$, and $\phi^2=0.20$



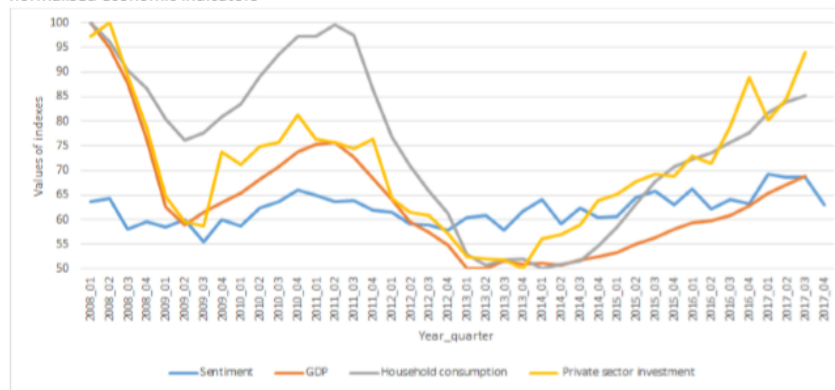
Valori ottenuti con l'uso del calcolo semplificato

19

I risultati

Esempio 3: estrazione del Sentiment da testi (Banca D'Italia) (2/2)

Figure 10: Sentiment index for each quarterly Bulletin (from January 2008 to October 2017) and of other normalised economic indicators



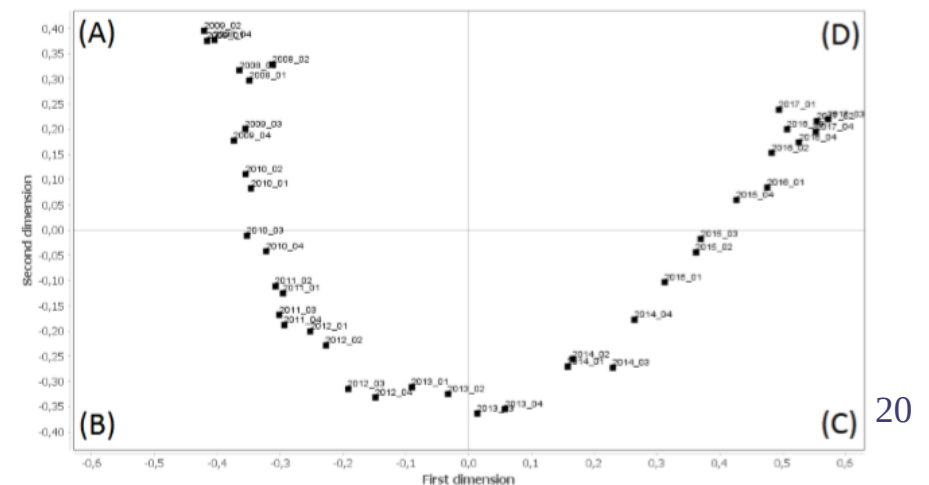
Sentiment e variabili macro-economiche:
esiste una correlazione significativa con il
GDP a LAG -1

Il metodo di Analisi delle Corrispondenze applicato alla tabella *termini* (righe)-*bollettini* (colonne) consente di effettuare una *topic analysis*

Table 8: terms that characterize the four quadrants identified by the LCA

Quadrant (A) Bulletin 2008-2010	Quadrant (B) Bullettins 2011-2013	Quadrant (C) Bullettins 2013-2015	Quadrant (D) Bullettins 2015-2017
rincaro materia aumento prezzo rincaro inflazione sostegno sistema ricapitalizzazione banca incentivo rottamazione riduzione disavanzo restrizione offerta credito contrazione esportazione operazione ristrutturazione debito correggere cartolarizzazioni	crisi debito manovra finanziaria riduzione debito riduzione spesa spread titolo stato riforma governance caduta pil taglio spesa ridurre tasso	costo credito impresa segnale rafforzamento legge stabilità ristrutturazione debito grezia liquidità banca credito impresa contributo european stability crescita esportazione corso greggio	Esito referendum brexit Banca monte paschi Stress test proseguire crescita crescita retribuzione diminuzione disoccupazione proseguire espansione credito ribasso inflazione sostenere crescita

Figure 15: Projections of the Economic Bulletins issued by the Bank of Italy from 2008 to 2017 in the first two dimensions of the Lexical Correspondence Analysis



I risultati

Esempio 5: classificazione automatica documenti (SOGEI)

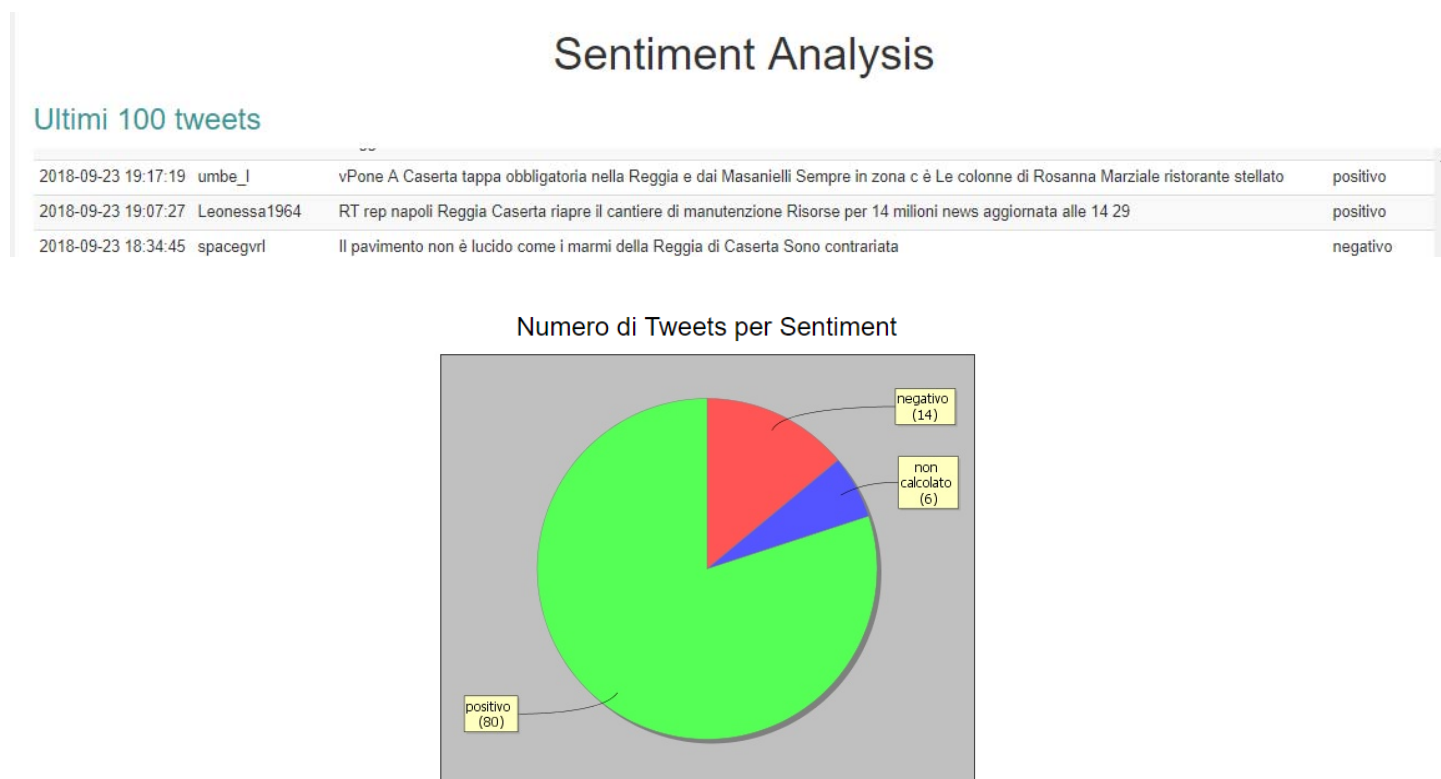
Il metodo sviluppato considera:

- Un opportuno trattamento del testo e delle questioni controverse da identificare;
- La possibilità di utilizzare il metodo dell'analisi delle corrispondenze (riferendosi alla tabella *termini-questioni controverse*) in un'ottica «predittiva», dove ogni nuovo documento viene proiettato considerandolo come una «variabile supplementare»

I risultati hanno evidenziato che nell'80% dei casi il solo trattamento del testo consente di riconoscere la questione controversa effettivamente trattata. Questa percentuale sale a circa il 97% nel caso in cui si consideri anche l'analisi delle corrispondenze.

I risultati

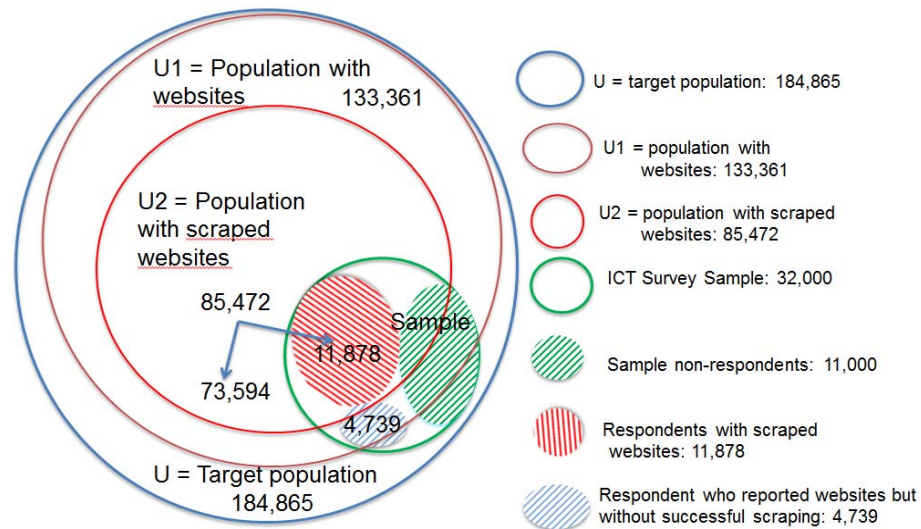
Esempio 4: estrazione del Sentiment da Twitter-FB-Tripadvisor (Reggia di Caserta)



L'applicazione viene eseguita automaticamente, aggiornandosi in tempo reale

I risultati

Esempio 1: sostituzione indagine e ampliamento delle informazioni sulle aziende (con ISTAT)



I risultati, noti, provenienti dall'Indagine sulle 11,878 aziende consentono la stima dei parametri di un metodo di classificazione che poi viene applicato all'intera popolazione di aziende (per le quali è stato possibile risalire al sito web, ossia 85,472).

In seguito si confrontano le stime totali con quelle derivanti dall'Indagine (*design based estimated*), tenendo conto anche dell'incertezza a queste associata.

Table 1- Web-ordering estimates comparison

DOMAIN		Sample	Population	Websites	Design based estimate	Lower limit C.I.	Upper limit C.I.	Model based estimate
Size class of persons employed								
cl1	from 10 to 49	13485	163041	114880	14.61	13.36	15.85	15.26
cl2	from 50 to 99	2402	12224	10164	15.34	13.39	17.28	16.35
cl3	from 100 to 249	2675	6443	5549	18.25	16.41	20.09	17.63
cl4	from 250 and more	2830	3157	2836	26.56	24.82	28.3	23.66

Non concordanza?

Cosa emerge dai risultati precedentemente illustrati?

- L'aggiunta d'informazione «empirica» da BD pone problemi di «copertura» e di «compatibilità»;
- Il Sentiment e l'analisi dei testi in generale offrono la possibilità di «comprendere» alcuni aspetti di un fenomeno reale, sebbene anch'essi limitati al contesto e con «copertura» limitata (gli utenti dei Social Media sono una parte di quelli generali);
- La sostituzione dell'informazione mediante BD potrebbe risultare «accettabile» entro certi «margini», ma non è «perfetta».

Ciò consente di definire la sequenza *Big Data* e *Machine learning* come un processo capace di ridurre l'incertezza informativa iniziale e di offrire un supporto alle decisioni.

Ma è possibile riferire tale processo a un appropriato contesto inferenziale? In altri termini... è possibile valutare anche l'incertezza associata ai «risultati» ottenuti con il *BD&ML*?

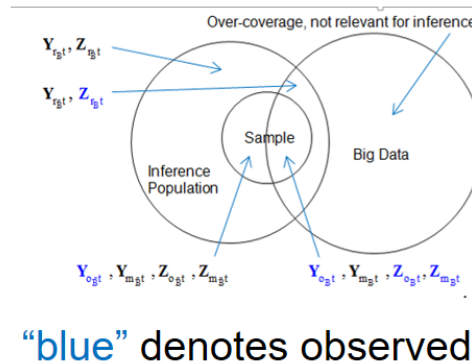
La valutazione dell'incertezza associata al BD e al ML

Tecniche statistiche di «ricampionamento» (tipo bootstrap) possono contribuire a determinare il livello d'incertezza delle stime ottenute con BD e ML. Tuttavia si pone il problema comunque della differente «copertura».

Un risultato tuttavia interessante sulla valutazione dell'incertezza è derivabile dai lavori di Siu-Ming Tam, il quale propone di considerare i valori noti per «calibrare i big data».

- Partition the Variable of Interest vector into $\mathbf{Y}_t = (\mathbf{Y}_{ot}, \mathbf{Y}_{mt}, \mathbf{Y}_{rt})'$
- Then the model $\mathbf{Y}_t = \mathbf{Z}_t \boldsymbol{\beta}_t + \mathbf{e}_t$ can be written as

$$\begin{bmatrix} \mathbf{Y}_{oB^t} \\ \mathbf{Y}_{mB^t} \\ \mathbf{Y}_{rB^t} \\ \mathbf{Y}_{o\tilde{B}^t} \\ \mathbf{Y}_{m\tilde{B}^t} \\ \mathbf{Y}_{r\tilde{B}^t} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{oB^t} \\ \mathbf{Z}_{mB^t} \\ \mathbf{Z}_{rB^t} \\ \mathbf{Z}_{o\tilde{B}^t} \\ \mathbf{Z}_{m\tilde{B}^t} \\ \mathbf{Z}_{r\tilde{B}^t} \end{bmatrix} \boldsymbol{\beta}_t + \begin{bmatrix} \mathbf{e}_{oB^t} \\ \mathbf{e}_{mB^t} \\ \mathbf{e}_{rB^t} \\ \mathbf{e}_{o\tilde{B}^t} \\ \mathbf{e}_{m\tilde{B}^t} \\ \mathbf{e}_{r\tilde{B}^t} \end{bmatrix}$$



Big Data e Machine Learning al CINECA

- Sviluppo e mantenimento software Free e Open Source (ADaMSoft) che implementa tutti i processi relativi ai processi precedentemente descritti. Esso può essere impiegato su qualunque elaboratore/sistema operativo (dal PC al Supercomputer);
- Test di pacchetti/strumenti/metodologie innovative (*deep learning*) su Supercomputer:
 - TensorFlow: machine learning and deep learning;
 - Algoritmi Word2vec: ricevono in ingresso un corpus e restituiscono un insieme di vettori che rappresentano la distribuzione semantica delle parole;
 - PyThorch: librerie per il natural language processing scritte in Python;
 - Keras: un'altra libreria per il machine learning.
- Organizzazione di corsi, docenze, laboratori, tirocini a tesi
- Partecipazione al BDVA (Big Data Value Association) e alla Commissione Big Data dell'ISTAT)

Alcune delle procedure specifiche create per il BD e il ML in ADaMSoft

- ☐ Data Management (DBMS interactions)
- ☐ Data Management (automatic changes on variables)
- ☐ Data Management (delete, copy, crypt, export, manage code labels, transpose, etc.)
- ☐ Data Management (sort, merge, append)
- ☐ Data Management (utility for learning procedures)
- ☐ Data set transfer
- ☐ Decision trees
- ☐ Error localization and data imputation (qualitative variables, requires Concord/SCIA)
- ☐ Error localization and data imputation (quantitative variables, FELLEGI-HOLT)
- ☐ Evaluate the result of classification models
- ☐ Evaluate the results of function approximation
- ☐ Fitting/interpolation (polynomial, spline, loess, etc.)
- ☐ Function fitting (user defined)
- ☐ Fuzzy cluster analysis (k-means method)
- ☐ Hierarchical cluster analysis
- ☐ Import a data table (file) in ADaMSoft (create a dictionary)
- ☐ Interface with R package
- ☐ Item analysis (Cronbach alpha, etc.)
- ☐ Linear algebra
- ☐ Linear discriminant analysis
- ☐ Linear logistic regression
- ☐ Linear regression
- ☐ Linear regression with fuzzy dependent variable
- ☐ Neural network (MLP)
- ☐ Principal component analysis
- ☐ Record linkage
- ☐ Replacement of numerical missing data values
- ☐ Statistics on quantitative variables
- ☐ Text mining (create term-document matrix and similar)
- ☐ Text mining (evaluate the sentiment)
- ☐ Text mining (interaction with external files, like import, export, etc.)
- ☐ Text mining (lemmatization, part of speech, language detection, etc., requires TreeTagger)
- ☐ Text mining (lexical correspondence analysis, topic analysis and related)
- ☐ Text mining (search for a word, annotate a data set considering specific distances for strings, etc.)
- ☐ Text mining (transform a sentence in a series of terms or vice versa)
- ☐ Text mining (tweets analysis, creating different matrixes of terms, hashtags, etc.)
- ☐ Utility (load a web log file or a generic log file)
- ☐ Utility (send an email, work with files, directories or FTP servers, copy a file from an URL, etc.)
- ☐ Web mining (execute queries on a web search engine)
- ☐ Web mining (spider, scraper, get links, parse pages, etc.)

Per approfondire:

The Sentiment Hidden in Italian Texts Through the Lens of A New Dictionary, July 2018, Conference: CARMA 2018 - 2nd International Conference on Advanced Research Methods and Analytics

ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms, October 2016, Conference: Modernization of Agricultural Statistics in Support of the Sustainable Development Agenda

Internet as Data Source in the Istat Survey on ICT in Enterprises. Austrian Journal of Statistics, 44(2), 31-43.

Use of Artificial Intelligence And Web Scraping Methods To Retrieve Information From The World Wide Web, Int. Journal of Engineering Research and Application

D. Sforzini, Reggia di Caserta: un Patrimonio in Rete, July 2017, Conference: Conferenza GARR 2016 - The CreActive Network. Uno spazio per condividere e creare nuova conoscenza,

G. Barcaroli, N. Golini, P. Righi, Quality evaluation of experimental statistics produced by making use of Big Data, July 2018, Conference: Q2108 - European Conference on Quality in Official Statistics

Tam, S.-M., and Clarke, F. (2015) Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. International Statistical Review, 83: 436–448. doi: 10.1111/insr.12105.

