Federico Neresini and **Andrea Lorenzet**
University of Padua and Observa *Science in Society* – Italy

Corpus construction and topic detection on Italian news S&T coverage 1990-2013

Bilgi University, ISTANBUL 9-12.1.13

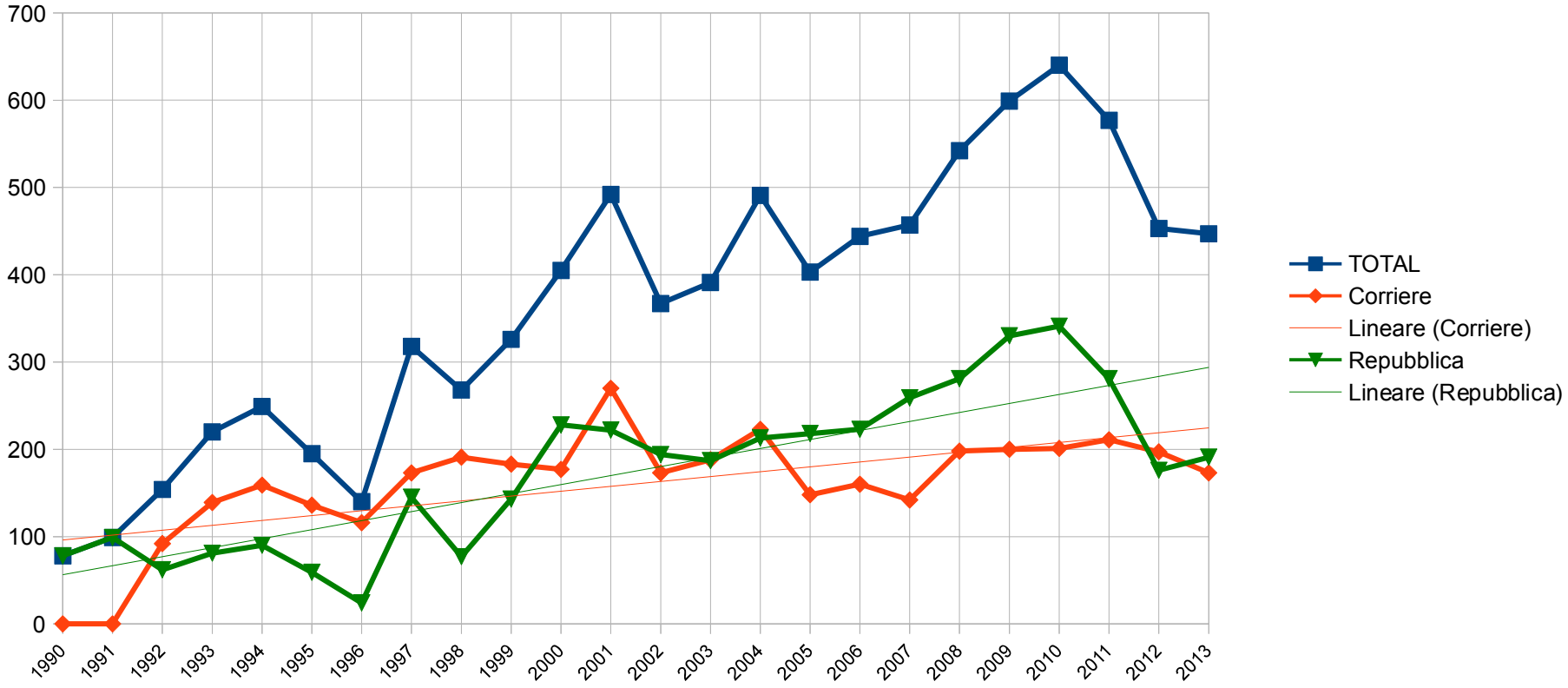
OVERVIEW

- Corpora construction 1990-2013;
- Applying LSE keywords to Italian corpora;
- Automated history of S&T in the news with LDA topic detection;
- Indicators & controversies

S&T coverage 1990-2013

La Repubblica and Il Corriere della Sera

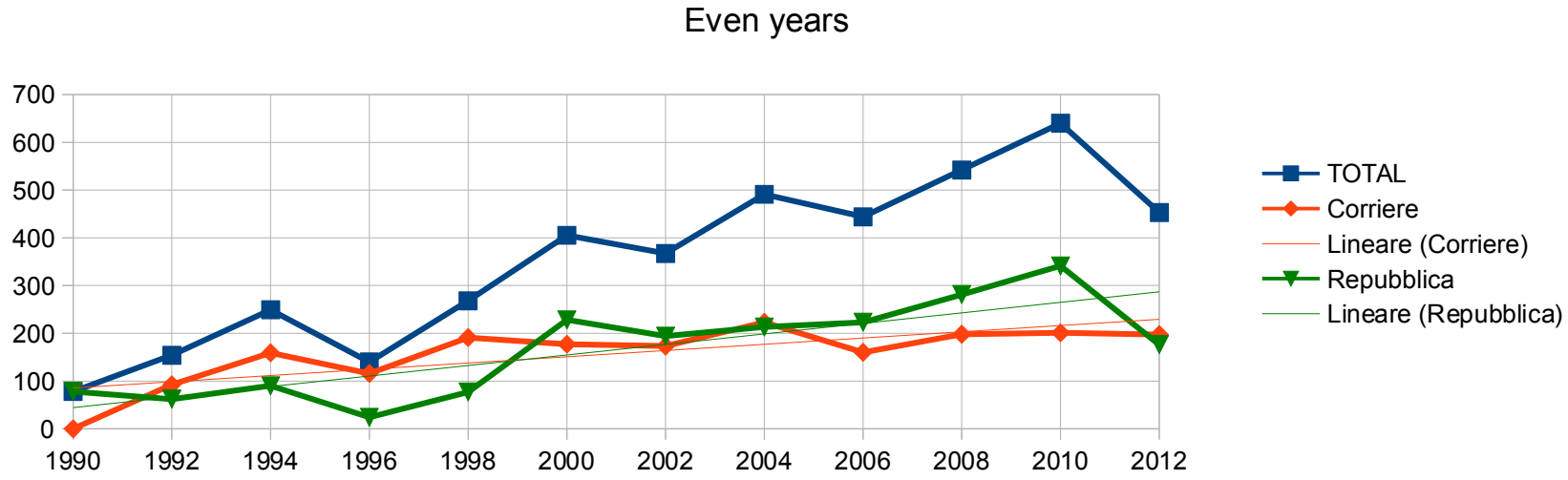
(total relevant articles n=9330, correlation between newspapers 0.54)

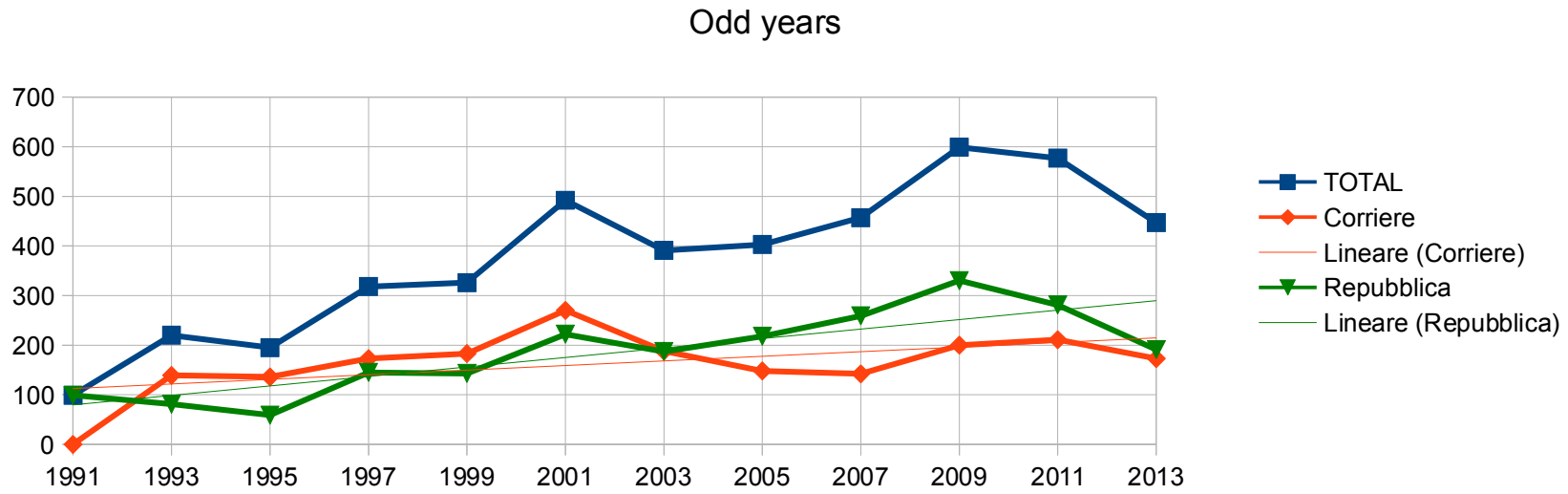


Corpora construction

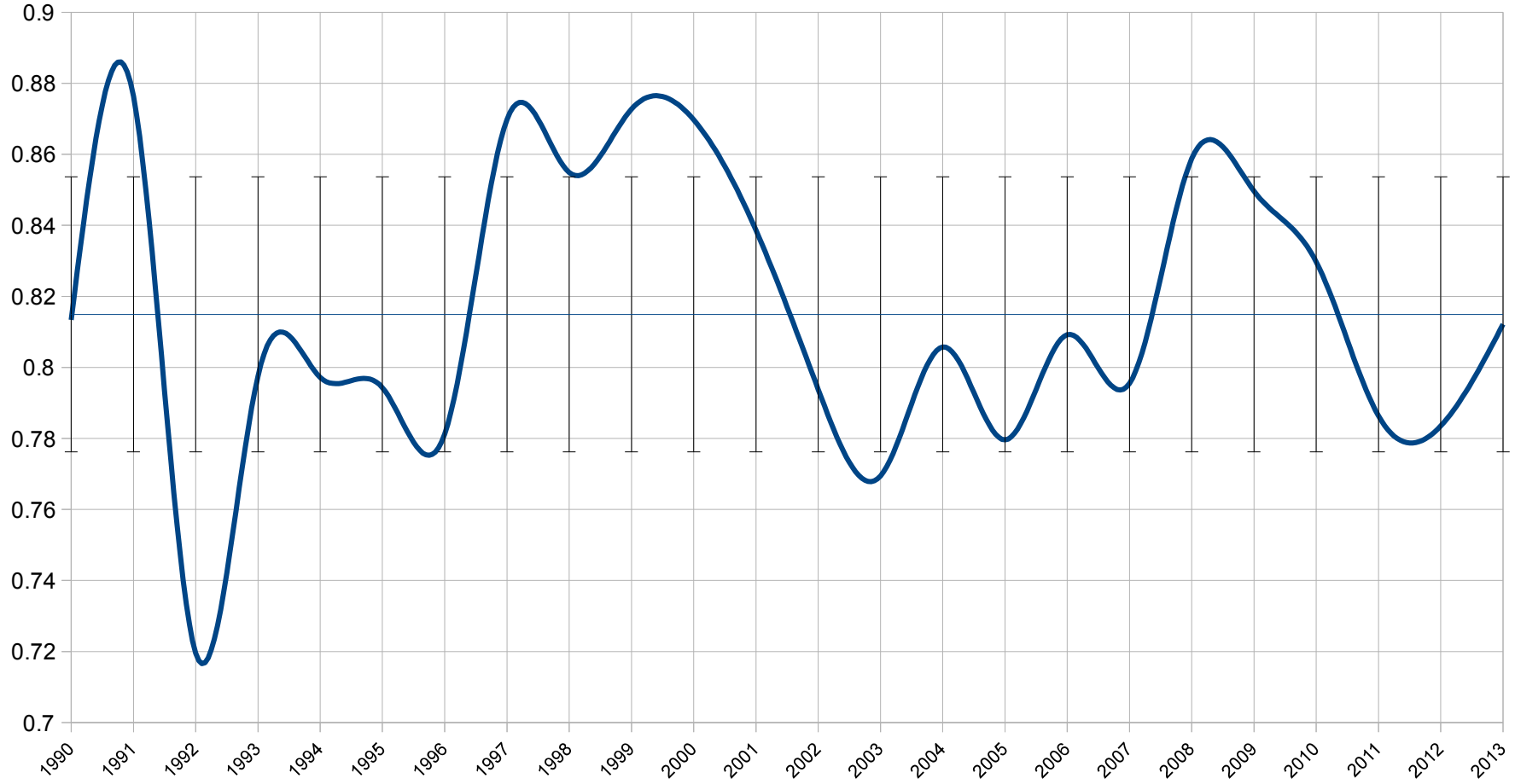
- Period considered 1990-2013
- Two sub-corpora (even and odd years) 1990-2013;
- Main Italian newspapers: Il Corriere della Sera (Milan) and La Repubblica (Rome);
- Collection of all articles for the selected days in the LSE artificial week schedule, one for even and one for odd years;
- Relevant articles detected through Science in the Media Monitor machine relevance scores;

Two sub corpora: even and odd years (corr.=0.90)

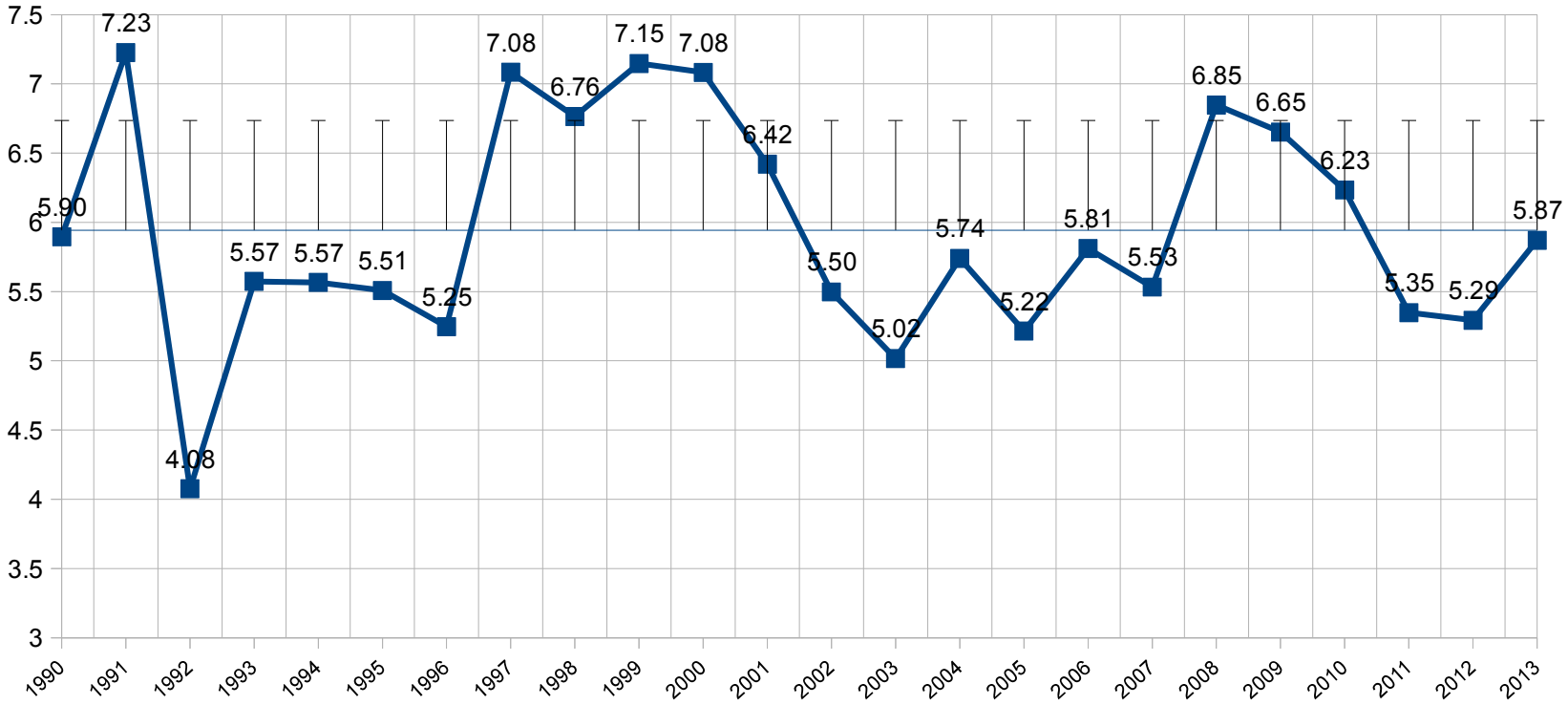




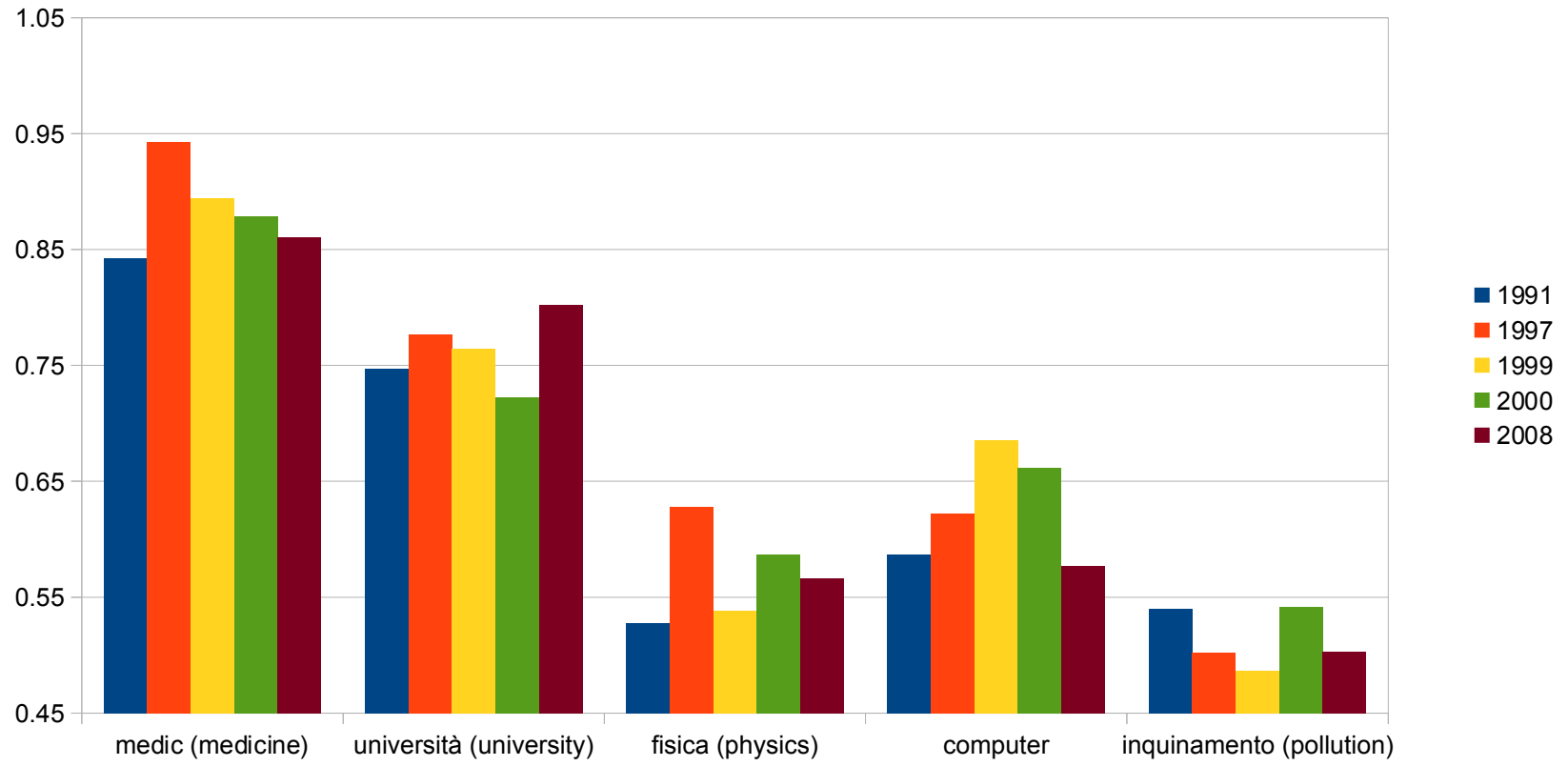
S&T salience 1990-2013 - 1/log(tot.art/rel.art)



S&T salience 1990-2013 - % of relevant art. on total



Peaks of coverage, an analysis of the main keywords/issues



Applying LSE keywords to SMM corpora (total 1990-2013=148290; even yrs.= 71532; odd yrs. 76758)

	Even	% (n=71532)	Odd	% (n=76758)	Total sample	% (n=148290)
Observa classifier	4680	6.45	4650	6.06	9330	6.29
LSE keywords	34621	42.33	36649	42.12	71270	48.06
LSE with Observa classifier	4340	6.07	4322	5.63	8662	5.84

SMM English language version

retrieved items: 548309

scraped items: 528712

after dedupping: 256103

Two online newspapers:

NYTIMES: from Jan 2013 194820 items

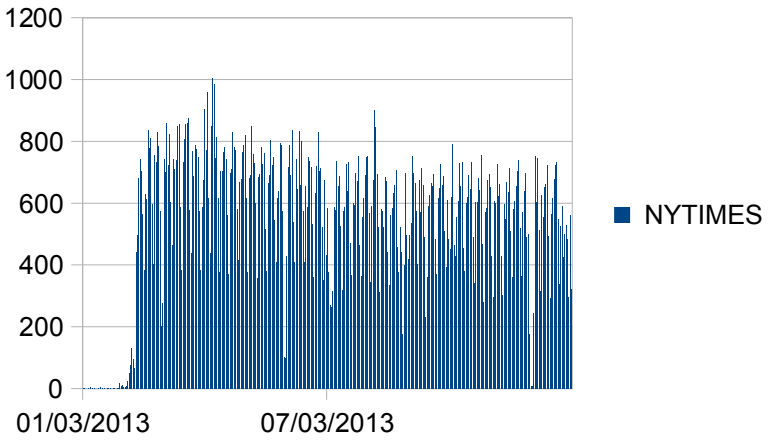
GUARDIAN: from Sep 2013 31716 items

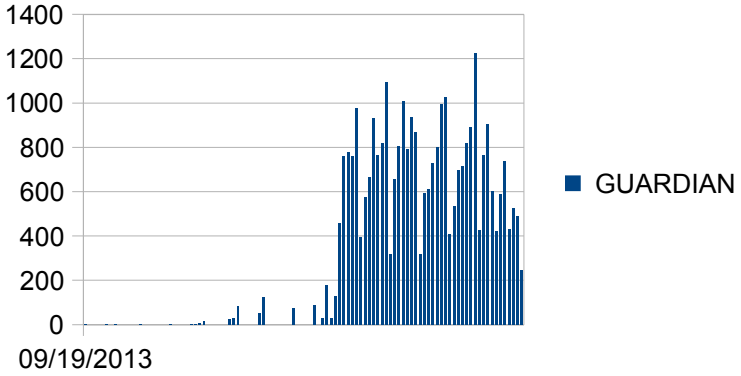
FINANCIAL TIMES: from Jan 2013

21709 items

SMM French version is also being tested

SMM blogs covers 500 sources with
1000 posts retrieved per day





LDA (automatic topic detection) all rel. articles 1992-2012 (n=8131)

FUTURE	.53	mondo vita fatto ricerca esempio caso spesso grande punto senso possono scienza politica problema vero idea possibile futuro volta
POLICY	.31	fatto ieri casa giorno italia presidente volta famiglia bene ore detto mesi aver perche citt legge vita sera visto
ECONOMICS	.30	italia milioni euro anno presidente miliardi cento governo regione citt imprese sviluppo paesi europa settore ministro mercato energia progetto
EDUCATION	.19	universit ricerca studenti scuola anno istituto corso laurea scienze giovani facolt formazione studio scuole corriere corsi ricercatori ingegneria sede
HEALTH	.18	salute medici malattia pazienti casi malattie medico rischio farmaci ospedale cellule ricerca bambini pagina studio paziente caso sangue possono
WEB/ICT	.16	internet computer rete mercato milioni sistema tecnologia mondo software web digitale dollari tv societ italia microsoft servizi azienda sito
SPACE	.17	acqua auto terra metri mare chilometri anno nuova grande aria motore spazio circa ambiente serie verso mondo km specie
HUMANITIES	.14	storia film libro vita grande mondo uomo libri amore padre ala secolo regia morte autore cinema opera guerra giovane
FOREIGN	.12	guerra usa america presidente americani uniti nucleare bush cina washington armi americano new paesi paese sicurezza americana mondo ieri
ARTS	.10	ore teatro musica piazza san arte euro mostra ingresso museo domani regia lire sala palazzo libero info festival spettacolo

LDA (automatic topic detection) highly rel. articles 1992-2012 (n=3293)

FUTURE	.90	ricerca esempio fatto punto italia volta caso alcuni vita possono infatti mondo cento problema momento possibile grande potrebbe importante
ECONOMICS	.23	milioni italia mercato miliardi anno aziende paesi sviluppo societ settore mondo imprese europa ricerca euro economia presidente industria innovazione
POLICY	.22	presidente legge ministro ieri governo politica consiglio diritto italia commissione sicurezza fatto nazionale cittadini unione direttore sanit diritti caso
EDUCATION	.22	ricerca universit studenti anno laurea corso scienze istituto ingegneria scuola corriere formazione studio facolt giovani corsi scadenza sede laureati
HEALTH	.20	salute cellule malattia malattie pazienti farmaci rischio casi medico ricerca medici virus tumori bambini ospedale sangue terapia donne cancro
ENVIRONMENT	.18	acqua energia ambiente terra inquinamento gas nucleare ambientale emissioni citt chilometri rifiuti metri aria milioni pianeta anno circa pagina
HISTORY	.18	casa fatto guerra giorno mondo vita storia padre film grande figli aver visto bene america volta figlio morte mesi
HUMANITIES	.16	scienza storia uomo libro mondo vita secolo scienziati cultura grande libri specie cervello opera fisica nobel scienziato idea universo
WEB/ICT	.14	internet computer rete software tecnologia web microsoft sistema mercato digitale nuova google auto servizi mondo sito tv informazioni online
ARTS	.11	ore teatro musica piazza lire mostra museo san euro arte domani ingresso citt palazzo centro bambini sala libero festival

Topics over years

	YRS.	92-96	97-01	02-06	07-12	92-12
	N	382	766	821	1324	3293
1.	FUTURE	1.3	.98	1.7	.97	.90
2.	EDUCATION	.15	.20	.22	.28	.22
3.	HEALTH	.15	.19	.17	.16	.20
4.	POLICY	.17	.18	.19	.11	.22
5.	ENVIRONMENT	.13	.13	.13	.19	.18
6.	WEB/ICT	.12	.17	.15	.13	.14
7.	HUMANITIES	.13	.11	.12	.12	.16
8.	GENETICS	N.A.	.14	.16	.14	N.A.
9.	FOREIGN	N.A.	.17	.09	N.A.	N.A.
10.	SPACE	.10	N.A.	N.A.	.10	N.A.
11.	HEALTH POLICY	.10	N.A.	N.A.	N.A.	N.A.
12.	ARTS	.09	.09	N.A.	.06	.11
13.	GMOs	N.A.	N.A.	.08	N.A.	N.A.

Topic “Health” over years

92-96	.15	virus malattia sangue aids cellule casi malattie epatite sonno gene geni stress punti organismo bambino farmaci genetica professor
97-01	.19	salute medico rischio malattia malattie casi pazienti possono donne terapia perche ospedale farmaci sangue medici paziente tumori
02-06	.17	farmaci pazienti medici malattia ricerca salute malattie ospedale medicina medico casi bambini paziente cura terapia tumori test rischio cuore
07-12	.16	pazienti malattia cellule malattie farmaci cancro medici tumori ricerca salute rischio donne test casi tumore medico paziente ospedale studio

Ranking of recurrent words for “health” topic

	1997-2001	2002-2006	2007-2012
1.	Health	Patients	Patients
2.	Physician	Physician	Physician
3.	Risk	Health	Cancer
4.	Patients	Hospital	Health
5.	Hospital	Cancer	Risk
6.	Cancer	Risk	Hospital

Some ideas for three different types of CAS indicators in the media

a) SALIENCE of general or specific issues:

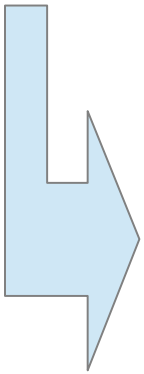
$$\frac{\text{n° articles on general or specific issue}}{\text{tot. n° of articles}}$$

b) RELEVANCE of controversial dimensions within a general or specific issue

average IDF of keywords classifiers for each dimension

c) FREQUENCY of terms related to a controversial dimension

For example: risk classifier TF on each article (n. risky terms/tot words in the article);



features of most significant articles within a general or specific issue (metadata)

Preliminary results (b type CAS indicators) on GMOs case study

La Repubblica and Il Corriere della Sera, period considered 1998-2013, n=4450

Dimension	*Indicator*	Avg IDF
Risk	Avg. IDF of selected risky terms	0.62
Antagonism	Avg. IDF of selected antagonistic terms	0.55
Moralization	Avg. IDF of selected morality terms	0.39
Politicization	Avg. IDF of selected politicization terms	0.69
Legal	Avg. IDF of selected legal terms	0.61

For example, risky terms are: risk, dread, danger, fear, illness, contamination, accident, alarm, catastrophe, tragedy, pollution, emission, cancer, disaster, accident, emergency