Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

Dati e *webapps*

Rocco Tripodi

rocco@unive.it

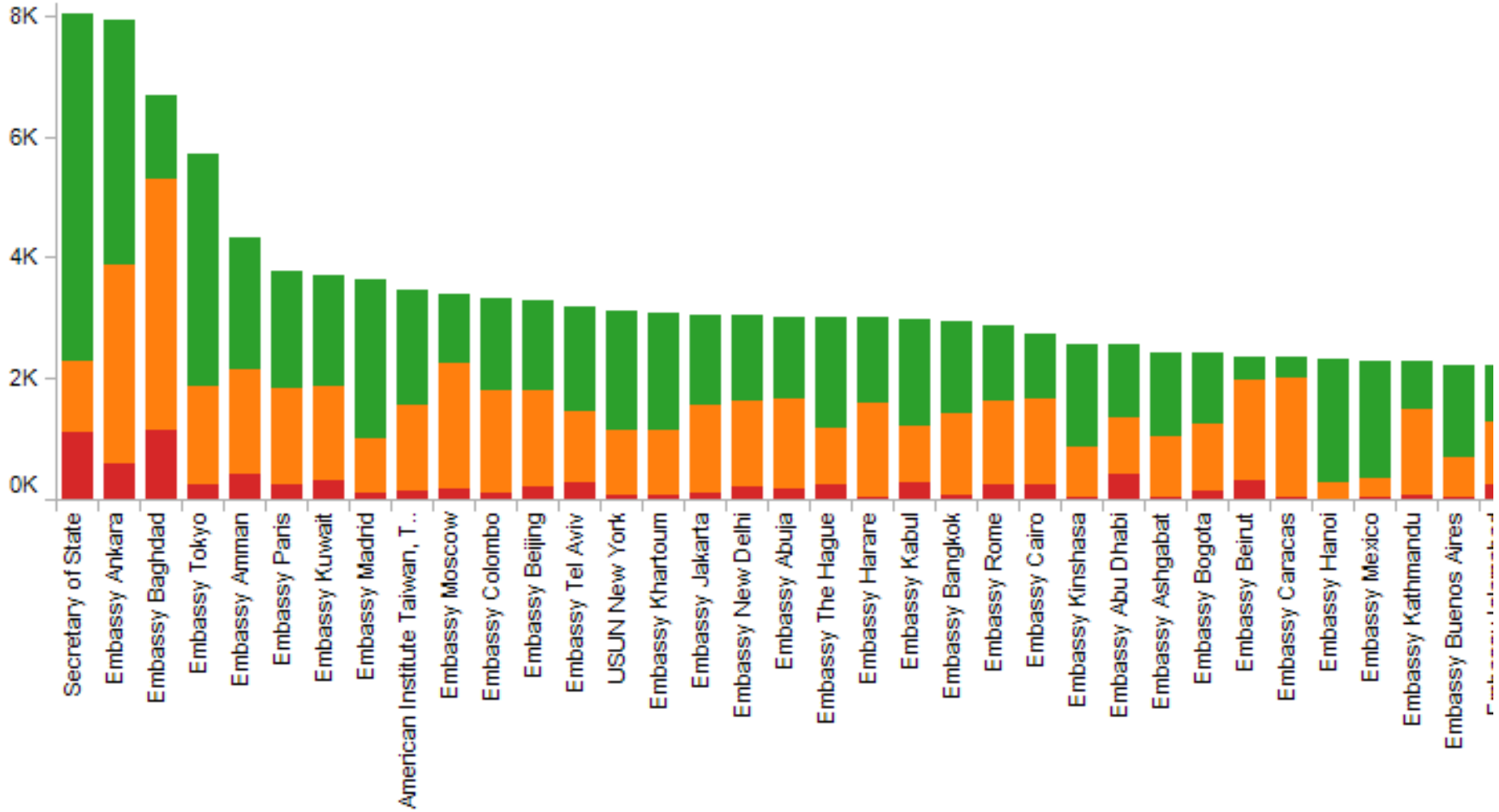
Il caso wikileaks come esempio

51 287 documenti 261 276 536 parole

Classificazione
15 652 secret
101 748 confidential
133 887 unclassified

Ogni documento oltre al *body text* comprende una dei metadati e un riassunto
Il sistema dei metadati ha dei *tags* standard usati dal Dipartimento di Stato
External political relations – 145 451
Internal government affairs – 122 896
Human rights – 55 211
Economic Conditions – 49 044
Terrorists and terrorism – 28 801

Visualizzazione 1



Visualizzazione 2

Visualizzazione creata con Tableau (Link), un software largamente usato nell'ambiente giornalistico per creare progetti di *data visualization*

Pre-processing dei documenti

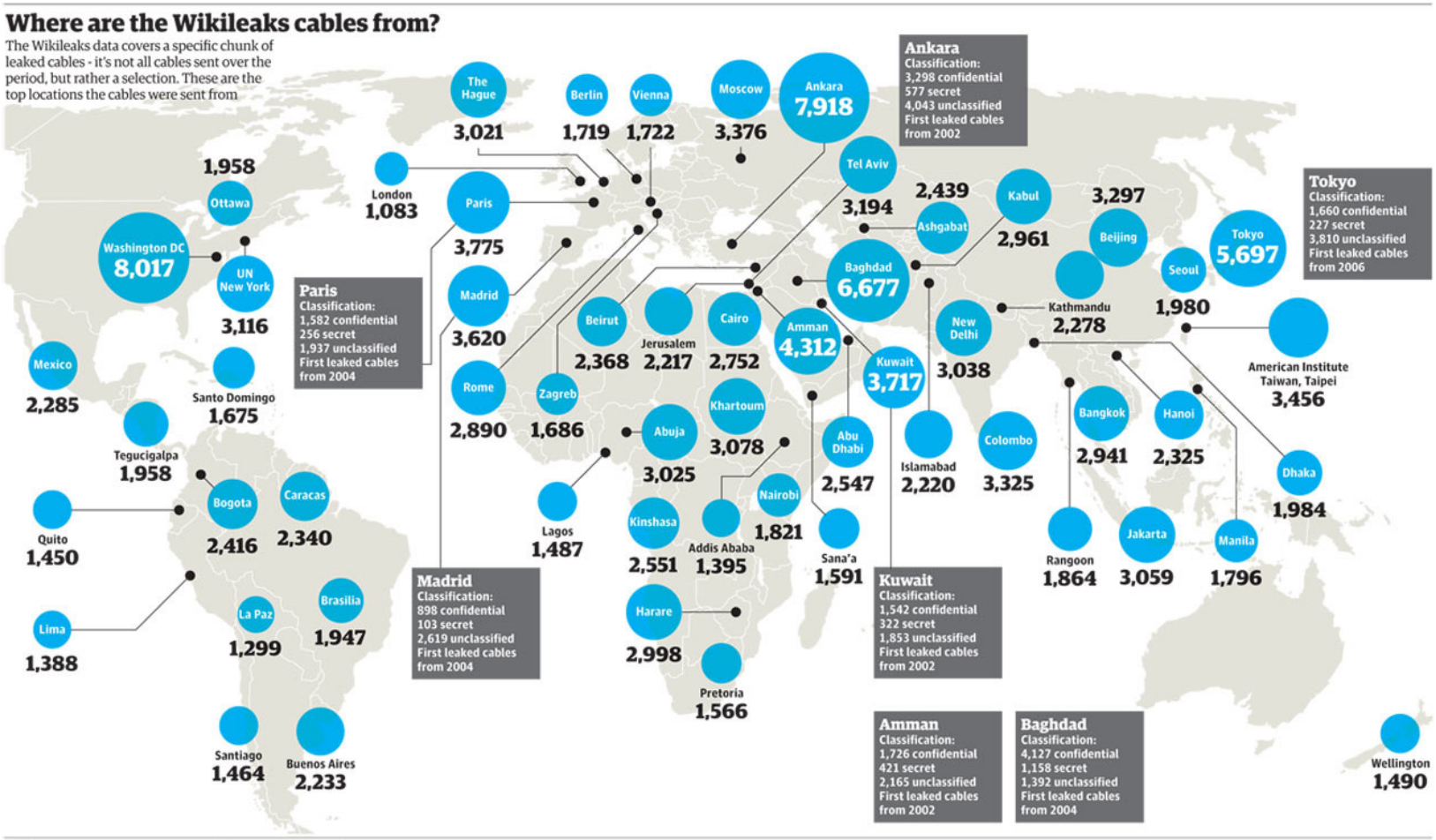
Analisi dei metadati

Visualizzazione basata sui metadati e non sul testo dei documenti

Es:

Saturday, 28 March 2009, 02:24
C O N F I D E N T I A L STATE 030049
EO 12958 DECL: 3/24/2019
TAGS OVIP">OVIP (CLINTON, HILLARY), PREL, AS, PK, AF, CH,
RS">RS,JA">JA">JA, KS, ASEAN
SUBJECT: (U) Secretary Clinton's March 24, 2009

Infografica del Guardian (Link)



Dataset (Link)

Google fusion tables All cables with tags

Discussions (3) Saved links (214) Get li

File View Edit Visualize Merge

Current view: **All** - Show options

1 - 100

Date	Source	Address	Tags
Feb 28, 2010	Embassy Tokyo	1-10-5 Akasaka Minato-ku, Tokyo 107-8420 Japan	ENRG, EPET, ETTC, IR, PREL JA
Feb 28, 2010	Consulate Vancouver	1095 W. Pender St., Vancouver, B.C. V6E 2M6 Canada	ASEC, KOLY, CASC, CA, PTE AMGT, PINR, OVIP
Feb 28, 2010	Secretary of State	U.S. Department of State 2201 C Street NW Washingt...	AEMR, ASEC, CASC, KFLO, MARR, PREL, PINR, AMGT, CI
Feb 28, 2010	Embassy Pretoria	US Embassy PO Box 9536, Pretoria 0001 877 Pretoriu...	PARM, MNUC, KNNP, PREL
Feb 28, 2010	Embassy Pretoria	US Embassy PO Box 9536, Pretoria 0001 877 Pretoriu...	KNNP, MNUC, PARM, IR, SF
Feb 28, 2010	Embassy Pretoria	US Embassy PO Box 9536, Pretoria 0001 877 Pretoriu...	PARM, PREL, SY, IAEA., SF
Feb 28, 2010	Consulate Monterrey	Ave. Constitución 411 Pte. Monterrey, Nuevo León. ...	ASEC, KCRM, SNAR, CASC, PGOV, MX
Feb 28, 2010	Secretary of State	U.S. Department of State 2201 C Street NW Washingt...	CASC, ASEC, KPAO, PREL, M
Feb 28, 2010	Secretary of State	U.S. Department of State 2201 C Street NW Washingt...	AEMR, ASEC, CASC, KFLO, MARR, PREL, PINR, AMGT, HA
Feb 28, 2010	Mission Geneva	IZD Tower Wagramerstrasse 17-19 1220 Vienna	PARM, KACT, MARR, PREL, RS, US

Web applications

Accesso tramite il web ad una applicazione

Esempi visti nella prima lezione

Interattività

Personalizzazione

Siti web dinamici (utilizzo di linguaggi come JavaScript e database)

Webapp del Guardian (Link)

Ricerca ed espolazione (di base) dei *cables*

Ricerca per parola chiave e per località geografica

Oltre ai metadati l'applicazione entra all'interno del testo per individuare le citazioni significative

Strutturazione dei documenti

Documenti strutturati VS documenti non strutturati

Es: txt e html/xml/xcl

I metadati aiutano a classificare i documenti

Es: gli *spider* dei motori di ricerca ricavano importanti informazioni leggendo le sezioni <meta> delle pagine html. Con le informazioni ricavate riescono a indicizzare le pagine (assegnando un id) in modo da ordinarle e renderne facilmente rintracciabili

Ma oltre alle meta informazioni, un testo contiene informazioni non strutturate che hanno bisogno di essere individuate e rese esplicite

Tecniche di Information Retrieval, Information Extraction, Text Summarization, Text Categorization e text Mining, unite alle tecniche di NLP, consentono di ricavare informazioni strutturate da documenti non strutturati