(Technical Aspects of)
Harvesting Data from Social Network Sites

aivars glaznieks & egon w. stemle
<aivars.glaznieks, egon.stemle}@eurac.edu>

Institute for Specialised Communication and Multilingualism

European Academy of Bozen/Bolzano
(EURAC)

February 14th, 2013
The Researchers’ Night is a mega event taking place every year on a single September night in about 300 cities all over Europe.

(Among other things) to see what researchers really do and why it matters for our daily life.

Different events offer a wide variety of fun-learning activities, e.g.

- behind-the-scenes guided tours of research labs (that are normally closed to the public),
- interactive science shows, and
- hands-on experiments or workshops.
Looking for things to do on Friday, 28 September 2012?

Interested in some fun learning – with your family, school or friends – through an engaging science show?

Then don't miss this year's Researchers' Night events!

From the events guide on the right, choose among about 300 cities all over Europe to find a science show that appeals to you most.

The events offer a wide variety of fun-learning activities, from behind-the-scenes guided tours of research labs that are normally closed to the public, through interactive science shows, to hands-on experiments or workshops.

See you there!
LUNA - Die Lange Nacht der Forschung 3.0

Was für eine Nacht!!! 13.000 Besucher holten sich den Energiekick bei der LUNA 3.0


Findest heraus, wer die tollen Preise gewonnen hat.
Goal

We wanted to show how HLT researchers

- process,
- analyse, and
- visualise data.

Means

To this end, we

- collected text snippets (FB Messages, Twitter Tweets) from parties participating in the Researchers’ Night 2012,
- processed the data (added Language ID, and POS tags),
- analysed the data (extracted POS distributions, and identified salient terms), and
- used the data for visualisation.
NotteGiardini
MontanelliRicercatori
Indro Pubblici Ricercatori
video Lombardia
MEETmeTONIGHT ricerca aggiunto
notte Città Di Piazza foto Pisa Milano Toscana pubblico Facebook Light caricato scienza Palermo pubblicato UnShine playlist preferiti Music mondo
Bolzano album domani città sarà Firenze proposte ai Festival Piazza settembre  pmi evento eco m progetto science Roma Congresso Central Europe Science visita uso senza usoColibri

LEGEND: Adverb, Adjective, Noun, Verb, Determiner, Pronoun, Other

Italiano
Deutsch
English
Français
all messages

servus wie gehts
ottima idea, continuate all'infinito
the kryostasis thingythings round the corner are a blast!
A Pisa, tutti gli eventi con tantissima gente. A la Limonaia lunghe file agli stand per "toccare" con mano alcune ... http://t.co/GeJYtWb7

A Pisa tutte le iniziative con un grande successo di pubblico: lunghe file agli stand dei laboratori per "toccare" http://t.co/vPFClp1z


aivars, egon (ComMul@EURAC)
harvesting social networks
February 14th, 2013
We collected data from
- Facebook (messages) and Twitter (tweets),
- parties participating in the Researchers’ Night 2012,
- people posting 'in the vicinity' of the city Bolzano,
- initially, in an asynchronous, and then, a synchronous way.

We used
- the Compact Language Detector embedded in Google’s Chromium browser for language identification,
- the IMS TreeTagger for POS tagging, and
- the WaCky corpora (i.e. frequency lists) for detecting salient words.
- Finally, we used readily available (mostly Google Chart) tools for visualisation.
Twitter and Facebook APIs

- The documentation of (and the discussions about) APIs were in dis-synchronisation with the ’current’ version of the API.
- We encountered difficulties in ’following too many users’ at the same time (+ vicinity restrictions).
The term "Web 2.0" was coined in January 1999 by Darcy DiNucci, a consultant on electronic information design (information architecture). In her article, "Fragmented Future", DiNucci writes:

The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven.”
It used to be Search Engines

In 2003 Adam Kilgarriff and Gregory Grefenstette put it like this:

The default means of access to the Web is through a search engine such as Google. Although the Web search engines are dazzlingly efficient pieces of technology and excellent at the task they set for themselves, for the linguist they are frustrating wrt. for example

- maximum number of queries,
- syntactic restrictions on formulating queries,
- obscure(d) selection criteria of results, and
- obscure(d) result figures.

Well, then download the pages (i.e. the former results)

but then, you’re in the business of web-page cleaning...
It used to be Search Engines II

...and in 2007 Adam Kilgarriff:

Working with commercial search engines makes us develop workarounds. We become experts in the syntax and constraints of Google, Yahoo, Altavista, and so on. We become ‘googleologists’. The argument that the commercial search engines provide low-cost access to the Web fades as we realize how much of our time is devoted to working with and against the constraints that the search engines impose.