

Δουλεύοντας με messy data - cleaning - filtering

Ανδρέας Βέγλης – καθηγητής



Αριστοτέλειο
Πανεπιστήμιο
Θεσσαλονίκης



OPEN
KNOWLEDGE
GREECE



ΣΥΛΛΟΓΟΣ ΑΠΟΦΟΙΤΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΚΡΗΤΗΣ

Περιεχόμενο δεδομένων

- Σπάνια τα δεδομένα που συλλέγουμε μπορούν να χρησιμοποιηθούν άμεσα.
- Συνήθη προβλήματα
 - Κωδικοποίηση χαρακτήρων
 - Δεκαδικό διαχωριστικό
 - Ορθογραφικά λάθη



Κωδικοποίηση χαρακτήρων

- Μη σωστή κωδικοποίηση → δεν είναι δυνατή η ανάγνωση των δεδομένων
- Ορίζοντας την κωδικοποίηση χαρακτήρων:
 - Υπολογιστές αναπαριστούν τα πάντα με μηδενικά και μονάδες (αριθμούς και γράμματα).
 - Χαρακτήρας → 1 byte με τιμή από το 1 έως το 255 (00000001- 11111111).
 - Κάθε γράμμα της αλφαβήτου και τα άλλα σύμβολα αναπαριστούνται με ένα τέτοιο byte.
 - a → 97, b → 98, c → 99 έως το 127 είναι τα αγγλικά
- Διαφορές στην κωδικοποίηση των χαρακτήρων ανάμεσα σε διαφορετικές χώρες (από το 128 έως το 255):
 - Στα Γαλλικά το character code 201 είναι το É
 - Ενώ στα Ρωσικά ο character code 201 το Ш



1	Name	Net worth	Sources of wealth
2	Bernard Arnault	21.24 billion	LVMH
3	Gérard Mulliez & family	21.00 billion	Auchan
4	Liliane Bettencourt & family	17.50 billion	L'Oréal
5	Bertrand Puech & family	12.20 billion	Hermès
6	François Pinault & family	8.10 billion	PPR
7	Serge Dassault & family	7.50 billion	Dassault
8	Louis-Dreyfus family	6.60 billion	Louis-Dreyfus Group
9	Alain Wertheimer & family	4.50 billion	Chanel
10	Pierre Castel & family	4.50 billion	Groupe Castel
11	Vincent Bolloré	3.80 billion	Bolloré

12
13 Source: Wikipedia

Western (ISO 8859-1)



```
0103 - Richest French families - ISO-8859-1.tsv *  
1 Name      Net worth  Sources of wealth  
2 Bernard Arnault      21.24 billion  LVMH  
3 Gerard Mulliez & family  21.00 billion  Auchan  
4 Liliane Bettencourt & family  17.50 billion  L'Oréal  
5 Bertrand Puech & family  12.20 billion  Hermès  
6 François Pinault & family  8.10 billion  PPR  
7 Serge Dassault & family  7.50 billion  Dassault  
8 Louis-Dreyfus family  6.60 billion  Louis-Dreyfus Group  
9 Alain Wertheimer & family  4.50 billion  Chanel  
10 Pierre Castel & family  4.50 billion  Groupe Castel  
11 Vincent Bolloré      3.80 billion  Bolloré  
12
```

13 Source: Wikipedia

Cyrillic (ISO 8859-5)

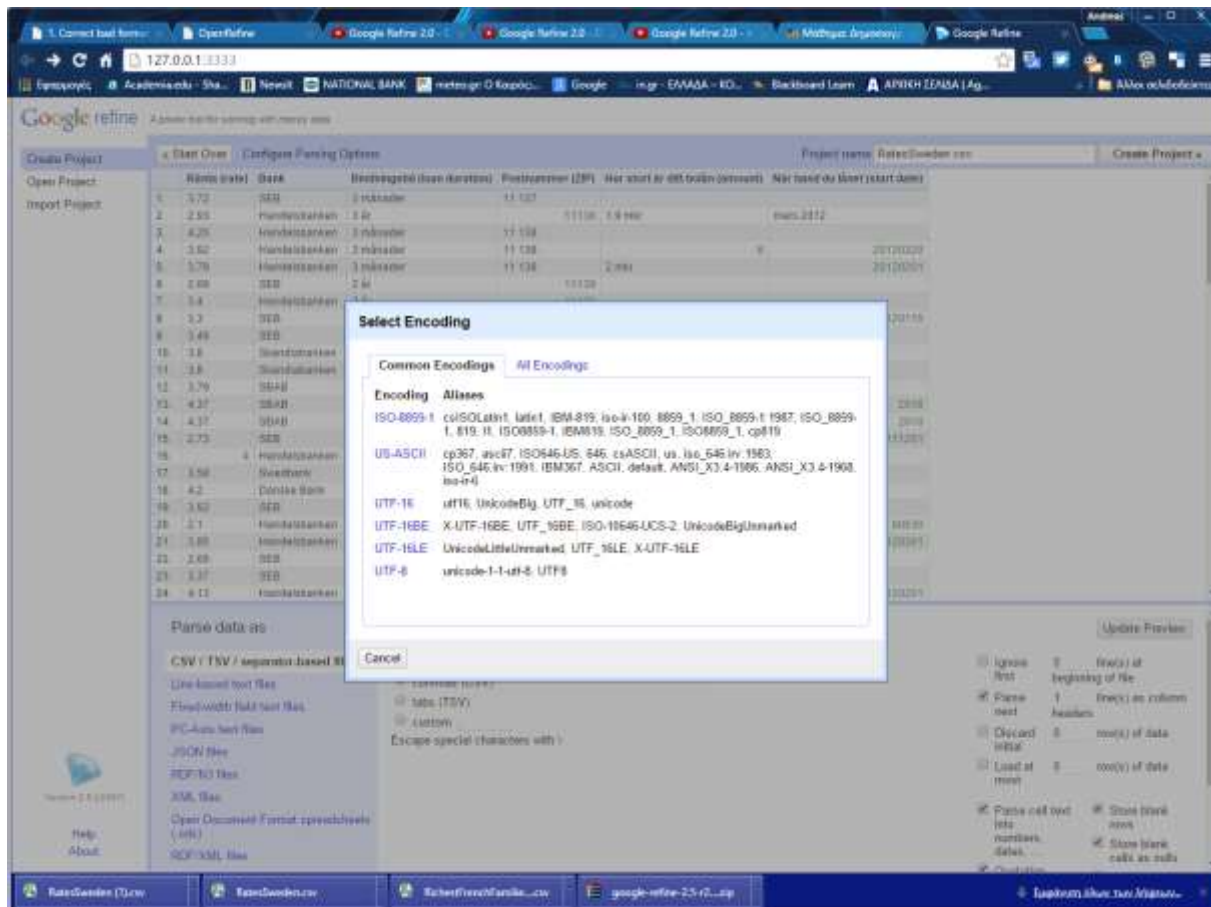
Εύρεση της σωστής κωδικοποίησης

- UTF-8 η στάνταρ κωδικοποίηση που χρησιμοποιείται σήμερα.
- Χρήση του εργαλείου ανοικτού κώδικα Open Refine.

Open Refine

- Open source εργαλείο
- <http://openrefine.org>
- Portable εφαρμογή (δεν απαιτεί εγκατάσταση)
- Προσφέρει πολλές δυνατότητες επίλυσης προβλημάτων messy data

Αλλαγή κωδικοποιήσεων άμεσα



Δεκαδικό διαχωριστικό

- Άλλες χώρες χρησιμοποιούν το , και άλλες την .
- Χρήση function μετατροπής στο Open Refine:
 - `Replace(value, “,”,”.”)`
- Δήλωση αριθμών στο Refine

Εισαγωγή αρχείου .csv με κωδικοποίηση UTF-8 στο Excel

- Δεδομένα → Λήψη εξωτερικών δεδομένων από κείμενο → οδηγός εισαγωγής (καθορισμός παραμέτρων)



Διόρθωση ορθογραφικών λαθών

- Αποτέλεσμα:
 - Καθαρισμός των δεδομένων
 - Δυνατότητα για συγκρίσεις



Παράδειγμα Data project

- Slate gun data (Slate)
- Periscopic Gun deaths in the USA data



Μη σωστές τιμές και διπλοεγγραφές

- Πως εντοπίζουμε μη σωστές τιμές στα δεδομένα
- Πως διορθώνουμε τις μη σωστές τιμές καθώς και τις διπλοεγγραφές.



Μη σωστές τιμές

- Πρόβλημα στη διαδικασία εισαγωγής δεδομένων. Ορθογραφικά λάθη.
- Παράδειγμα:
 - [Mapped: Every Protest on the Planet Since 1979](#)



Διπλοεγγραφές

- Παράδειγμα, αριθμός μεταλλίων σε Ολυμπιακούς αγώνες.
- [Assessment of Italian schools at risk in case of an earthquake](#)

Προχωρημένες τεχνικές καθαρισμού δεδομένων

- Χρήστη τεχνικών πληθοπορισμού για τον καθαρισμό δεδομένων.
- Σωστή οργάνωση των δεδομένων σε μορφή (format) βάσης δεδομένων



Παραδείγματα χρήσης πληθοπορισμού για καθαρισμό δεδομένων

- [Free the Files: Help ProPublica Unlock Political Ad Spending](#) (καλή εφαρμογή)
- Διαμοιρασμός δεδομένων προς καθαρισμό μέσω google spreadsheet (μη καλή εφαρμογή).

Καλή εφαρμογή

Tell Us About This Document

DOCUMENT TEXT Zoom Search

p. 1

WiredMedia Station Order Revised
Center Forward 2012

Market: Salt Lake City-Ogden Estimate ID
Flight Dates: Monday, September 10, 2012 to Sunday, September 16, 2012 11541

Contact: Diane Downey Phone: 801-275-4444
Email: diane.downey@wiredmedia.com Fax: 801-275-4525
Direct:

KUCW-TV

Program Name	DP	Length	Days	Gross Rate	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Total
					Sep 10	Sep 11	Sep 12	Sep 13	Sep 14	Sep 15	Sep 16	
Rules of Engagement	SP	30	M-F	\$71.00	1			1				2
5:00 PM - 6:30 PM				\$193.00								
Two & Half Men	1A	30	M-F	\$400.00	1		2	1	1			5
5:00 PM - 6:30 PM				\$1,200.00								
Big Bang Theory	1A	30	M-F	\$500.00	1		1	1	1			4
6:30 PM - 7:00 PM				\$1,500.00								
Station Gross Totals				\$4,400.00	3		3	2	2			10

1. Who Bought It? [HELP](#)

Was **CENTER FORWARD** the candidate/committee who bought these ads?

Yes

No

2. What Agency? [HELP](#)

What is the name of the advertising agency that bought these ads?

3. Contract Number? [HELP](#)

What is the contract number on this ad buy? (Sometimes labeled "Rep. Order")

4. How Much? [HELP](#)

What was the gross total cost? (Sometimes labeled "Report Totals")

Is there something else notable about this file?

Κακή εφαρμογή

manifestation du 7 septembre

File Edit View Insert Format Data Tools Help Last edit was made on June 16, 2011 by anonymous

123 Arial 10



	A	B	C	D	E	F	G	H	I	J	K
1								? = on ne sait pas de qui vient le chiffre			
2		Abbeville									
3											
4		Agen	7000	4000	34	0.6120948				9000	4400
5		Ajaccio						Total organisateurs	2251150		
6		Albertville						Total police	1113725		
7		Albi									
8		Alençon	6000	4300	46.47651	0.08359				4000	4000
9		Alès	3500	3500	44.1273866	4.0790267	police				



Εργαλεία για χρήση πληθοπορισμού για καθαρισμό/διόρθωση δεδομένων

- [Crowdcrafting](#)
 - Open Knowledge Foundation.
 - Υποστηρίζει τη δημιουργία μικρών διεργασιών για το κοινό.
 - Επιτρέπει την αντιγραφή και τη προσαρμογή γνωστών διεργασιών
- [Amazon Mechanical Turk](#)
 - Παρόμοια λογική
 - Υποστηρίζει την αμοιβή των χρηστών που συμμετέχουν.



Crowdcrafting

Oops! You are an anonymous user and will not get any credit for your contributions. Sign in now!

PDF Transcription: Contribute

Transcribe the following page

Important This is just a demo application. You can re-use the code to create your own application.



link. On this iteration, TraceMonkey calls T_{10} . Because T_{10} has a statement on line 2 in scope, this branch was not taken in the original case, so this causes T_{10} to fall against and take a side exit. The goal is not yet met, so TraceMonkey returns to the interpreter, which executes the continue statement.

And, TraceMonkey calls T_{11} , which in turn calls the second case T_{12} . T_{12} loops back to its case header, setting the first location where a case containing in the monitor.

link. On this iteration, the only job on line 2 is taken again. This time, the side exit becomes hot, so a trace T_{13} is installed that covers line 3 and returns to the loop header. Thus, the end of T_{12} jumps directly to the start of T_{13} . The side exit is passed to the on-failure location, it never directly to T_{13} .

At this point, TraceMonkey has completed enough traces to cover the entire second loop iteration, so the rest of the program runs entirely as before.

3. Trace Trees

In this section, we describe how, trace trees, and how they are used at run time. Although our techniques apply to any dynamic language interpreter, we will describe them assuming a bytecode interpreter to keep the exposition simple.

3.1 Traces

A trace is simply a program path, which may cross function call boundaries. TraceMonkey focuses on loop traces, that originate at a loop edge and represent a single iteration through the monitored loop.

Similar to an annotated basic block, a trace is only covered at the run time but may have many exits. In contrast to an annotated basic block, a trace can contain side exits. Since a trace always only follows one single path through the original program, however, new exits can be assigned to an exit in a trace and have a single predecessor with the original block.

A special case is a basic annotated with a type for every variable (including instructions in the trace). A special trace also has an entry type map giving the expected types for variables used on the trace before they are defined. For example, a trace could have a type map like: $\{x:1, \text{ for } \text{function}\}$, meaning that the trace may be entered only if the value of the variable x is of type int , and the value of for is of type function . The entry type map is much like the signature of a function.

In this paper, we only discuss typed loop traces, and we will refer to them simply as "traces". The key property of typed loop traces is that they can be compiled to efficient machine code using the same techniques used for typed languages.

© 2006-2009 Mozilla Foundation. All rights reserved.

You are working now on task: 20368

You have completed: 0 tasks from 14

Write here the transcription

Submit transcription!

Our Favorite!



Sound Cloud

If you're into music, this is your project. The demo that also shows you how to create your own sound recognition project.

 Author

C'mon!

Find a project that fascinates you,
and help researchers analyse their
data.

FIND A PROJECT

Create a Project and get other
volunteers to help you analyse your
data.

CREATE A PROJECT

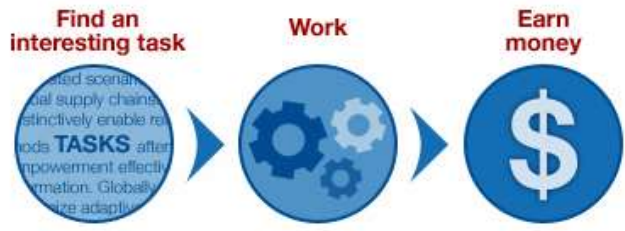
Mechanical Turk is a marketplace for work.
 We give businesses and developers access to an on-demand, scalable workforce.
 Workers select from thousands of tasks and work whenever it's convenient.
275,829 HITS available. [View them now.](#)

Make Money by working on HITS

HITS - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITS now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



[Find HITS Now](#)

or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITS - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITS completed in minutes
- Pay only when you're satisfied with the results



[Get Started](#)

Συμβουλές

- Καθορίστε ακριβώς την μέθοδο καθαρισμού των δεδομένων.
- Δεν είναι σίγουρο ότι οι συμμετέχοντες χρήστες θα βρουν τις μη σωστές καταχωρήσεις.
- Μπορεί να χρειαστεί να καθαρίσετε ξανά αυτά που επεξεργάστηκαν οι χρήστες.
- Ο Πληθοπορισμός είναι ενδιαφέρουσα τεχνική αλλά δεν μπορεί να είναι η στάνταρ μέθοδός σας για καθαρισμό δεδομένων.



Η περίπτωση των Μεγάλων Συνόλων Δεδομένων (Big Data Sets)

- Ο τρόπος που τα αντιμετωπίζουμε είναι διαφορετικός από την περίπτωση των μικρών σετ δεδομένων.
- Χωρισμός των δεδομένων σε πολλούς συνδεδεμένους πίνακες.
- Διευκολύνει το καθαρισμό των δεδομένων.

