

Informatica per bibliotecarie?



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

software e dati, pandemic edition!

Piero Grandesso

4 e 11 maggio 2021

ABIS – Unità di processo Gestione e sviluppo della biblioteca
digitale di Ateneo – AlmaDL



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Indice



mar 4, 09:30–12:30

09:30–10:00	Introduzione e panoramica <i>Library Carpentry</i>
10:00–10:25	Gestione dei dati, buone pratiche
10:25–10:50	Espressioni regolari – I
10:50–11:05	<i>pausa</i> ☕
11:05–11:20	Espressioni regolari – II
11:20–12:30	OpenRefine



mar 11, 09:30–11:30

09:30–10:20 riepilogo e domande

10:20–10:30 *pausa* ☕

10:30–11:30 l'esercitazione! ☠



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Introduzione



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Disclaimer:
questo corso < *Library Carpentry*

<https://librarycarpentry.org/>



Regole del gioco

1. domande: siamo in ~~tropp...~~ tant[ei]
segnatevi le domande e mettetele nella chat solo quando ve lo chiedo
2. pausa → puntualità
c'è poco tempo
3. dialoghiamo, pur in modalità asincrona
l'incontro di restituzione serve per quello



Principi

- strumenti liberi (*free software*)
- documentazione accessibile (*open access*)
- condivisione di competenze (<https://stackoverflow.com/>)
- formati interoperabili



Principi

- strumenti liberi (*free software*)
- documentazione accessibile (*open access*)
- condivisione di competenze (<https://stackoverflow.com/>)
- formati interoperabili
- *Findable, Accessible, Interoperable, Reusable...*
suona già sentito?



«Che c'entro io con l'informatica?»

L'**informatica** è la scienza che si occupa del **trattamento dell'informazione** mediante procedure automatizzate.
[[wikipedia: Informatica](#)]

Una **biblioteca** è un servizio finalizzato a soddisfare bisogni **informativi** [...] realizzato sulla base di una **raccolta organizzata** di supporti delle **informazioni**, fisici [...] o digitali.
[[wikipedia: Biblioteca](#)]



Altri buoni motivi...

- automazione di processi ripetitivi
- riduzione del rischio di errori
- facilitare lo scambio di dati tra colleg[h]i
- collaborare con più efficacia con gli informatici ☹️

Is It Worth the Time?



HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?
(ACROSS FIVE YEARS)

	HOW OFTEN YOU DO THE TASK					
	50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
6 HOURS				2 MONTHS	2 WEEKS	1 DAY
1 DAY					8 WEEKS	5 DAYS

[<https://xkcd.com/1205/> by Randall Munroe, Creative Commons BY-NC 2.5]



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Uno sguardo nel mondo di Library Carpentry



La shell

- la *shell Unix*, è un potente strumento per manipolare i file
- presente su computer Linux e Mac ma anche installabile su Windows[*]
- fondamentale per grandi quantità di file, file molto pesanti, file di testo semplice
- comandi ripetitivi possono essere raggruppati in *script* ripetibili al bisogno, possono includere variabili, condizioni e altre cose per renderli più flessibili

[*]: ad esempio <https://gitforwindows.org/>



Git

Git è un famoso software per il controllo della versione dei file, serve per:

- tenere traccia delle modifiche (e backup) dei file di un progetto
- permettere di annullare modifiche (*rollback*) o creare più rami di un progetto (*branches*)
- collaborare tra più utenti nella modifica e scrittura
- rende trasparente lo sviluppo del software che potreste utilizzare



SQL

Structured Query Language è un linguaggio per interrogare e manipolare database relazionali.

Un database relazionale è un insieme di tabelle (con colonne di dati omogenei e righe contenenti i vari record) che contengono dati messi in relazione tra loro ed è la base dei dati standard per quasi tutti gli applicativi web



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Gestione dei dati, alcune buone pratiche

Prendiamo una tabella...



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Cibo	Colore
mela	gialla rossa verde
banana	gialla marrone
mozzarella	bianca
arancia	arancione rossa



Plain text – vantaggi

- minore complessità
- gestibili con una molteplicità di strumenti, lo saranno anche nel futuro
- minori distrazioni (puro contenuto)



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Editor di testo

Programmi pensati per visualizzare e modificare file di testo semplice:



Editor di testo

Programmi pensati per visualizzare e modificare file di testo semplice:

- [Notepad++](#) (windows)



Editor di testo

Programmi pensati per visualizzare e modificare file di testo semplice:

- [Notepad++](#) (windows)
- [Atom](#) (windows, mac, linux)



Editor di testo

Programmi pensati per visualizzare e modificare file di testo semplice:

- [Notepad++](#) (windows)
- [Atom](#) (windows, mac, linux)
- [TexEdit](#) (mac – ma occhio alla modalità)



Editor di testo

Programmi pensati per visualizzare e modificare file di testo semplice:

- [Notepad++](#) (windows)
- [Atom](#) (windows, mac, linux)
- [TexEdit](#) (mac – ma occhio alla modalità)
- **non** Word **né** il blocco note!



Editor di testo – feature

- contatore di righe e caratteri
- funzione di Trova/Sostituisci avanzata (regex)
- evidenziazione della sintassi (se scrivete XML o Markdown il testo si colora per facilitare la lettura)
- scorciatoie, estensioni, correttori automatici, suggeritori di parola etc



Formati a testo semplice

Puri: TXT, CSV (*comma-separated values*), TSV (*tab-separated values*)

Strutturati (o annotati): XML, HTML, Markdown, JSON, TeX



I caratteri

Esistono varie codifiche e una buona dose di complessità, anche nel testo semplice...

- ASCII

```
!"#$%&'()*+,-./0123456789:;<=>?  
@ABCDEFGHIJKLMNPOQRSTUVWXYZ[\]^_  
`abcdefghijklmnopqrstuvwxyz{|}~
```



I caratteri – set estesi

Si crea una babele di codifiche differenti, quasi mai compatibili tra loro...

[*]

[*]: se vi interessa l'argomento <https://www.youtube.com/watch?v=MijmeoH9LT4>



I caratteri – set estesi

Si crea una babele di codifiche differenti, quasi mai compatibili tra loro...

[*]

- vecchi: [ISO/IEC 8859](#) e derivati ([Windows-1252](#) per l'alfabeto latino)

[*]: se vi interessa l'argomento <https://www.youtube.com/watch?v=MijmeoH9LT4>



I caratteri – set estesi

Si crea una babele di codifiche differenti, quasi mai compatibili tra loro...

[*]

- vecchi: [ISO/IEC 8859](#) e derivati ([Windows-1252](#) per l'alfabeto latino)
- esotici: [Shift JIS](#), [GB 2312](#) ...

[*]: se vi interessa l'argomento <https://www.youtube.com/watch?v=MijmeoH9LT4>



I caratteri – set estesi

Si crea una babele di codifiche differenti, quasi mai compatibili tra loro...

[*]

- vecchi: [ISO/IEC 8859](#) e derivati ([Windows-1252](#) per l'alfabeto latino)
- esotici: [Shift JIS](#), [GB 2312](#) ...
- e nuovi: **Unicode (UTF-8)**

[*]: se vi interessa l'argomento <https://www.youtube.com/watch?v=MijmeoH9LT4>



Machine readable

Facilitiamo la vita alle macchine (ma pure a noi!) adottando gli standard esistenti: **en-US** o **CL** è molto più gestibile e neutro rispetto alla lingua usata che scrivere “inglese americano” o “Cile”.

Occhio agli standard esistenti, in particolare nelle sigle di lingue ([ISO 639](#)) e paesi ([ISO 3166](#)) e nelle date ([ISO 8601](#)).

Gli standard – le date



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013.II.27. $27\frac{1}{2}$ -13 2013.158904109
MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800
 $((3+3)\times(111+1)-1)\times 3/3-1/3^3$ ~~2013~~  Missss
10/11011/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 & \end{matrix}$

[<https://xkcd.com/1179/> by Randall Munroe, Creative Commons BY-NC 2.5]



Nomi dei file e percorsi

nomi parlanti ma brevi, evitare spazi all'interno dei nomi, date nei nomi dei file (inizio o fine?), attenzione all'ordine alfabetico!

percorsi (struttura delle cartelle): gerarchie multilivello vs orizzontalità



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Espressioni regolari



Definizione

Le espressioni regolari (o *regex*) sono sequenze di caratteri che consentono di definire una serie di stringhe (che combaciano con la sequenza: *match*). Definiscono un percorso di ricerca (*pattern*) all'interno di un testo, con molta più precisione rispetto alle *wildcard* spesso usate nei motori di ricerca (* ? %).

wildcard

student* può includere "studenti" e "studentesse", ma anche "studentato", "studentelli"...

regex

student(i|esse) è molto più preciso, non lascia spazio ad alternative.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:



Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:

- qualsiasi editor di testo semplice (degno del nome)



Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:

- qualsiasi editor di testo semplice (degno del nome)
- LibreOffice / Open Office



Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:

- qualsiasi editor di testo semplice (degno del nome)
- LibreOffice / Open Office
- Open Refine



Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:

- qualsiasi editor di testo semplice (degnò del nome)
- LibreOffice / Open Office
- Open Refine
- ma in realtà quasi ogni software avanzato per la gestione di dati



Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:

- qualsiasi editor di testo semplice (degnò del nome)
- LibreOffice / Open Office
- Open Refine
- ma in realtà quasi ogni software avanzato per la gestione di dati
- ah no, niente Excel! 🤔

Principali regole



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

elemento seq.	match
[abc]	Un carattere tra i seguenti: a, b, c
[^abc]	Un carattere eccetto: a, b, c
[a-z]	Un carattere nell'intervallo: a-z
[^a-z]	Un carattere fuori dall'intervallo: a-z
[a-zA-Z]	Un carattere negli intervalli: a-z, A-Z
.	Un qualsiasi carattere
\s	Un carattere di spaziatura (orizz+vert)
\S	Un carattere eccetto quelli di spaziatura



Principali regole – 2

el. seq.	match
<code>\d</code>	Una cifra = <code>[0-9]</code>
<code>\D</code>	Un carattere eccetto le cifre
<code>\w</code>	<i>word-character</i> = <code>[a-zA-Z0-9]</code>
<code>\W</code>	Un <i>non-word-character</i>
<code>(...)</code>	<i>Capture group</i>
<code>(ab cd)</code>	O la seq. “ab” oppure “cd”

Principali regole – 3



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

el. seq.	match
a?	Zero o un "a"
a*	Zero o più "a"
a+	Una o più "a"
a{3}	Esattamente 3 "a" = aaa
a{3, }	3 o più "a"
a{3, 6}	Non meno di 3 e non più di 6 "a"
^	Inizio della stringa
\$	Fine della stringa



regex – esempio

Cosa prende questa *regex*?

```
^199[5-7]$
```

1993

millenovecentonovantacinque

2995

199

1995

19957

1999

1996

19 97

1997



Limiti

Non esiste uno standard unico per le *regex*, la cui sintassi e le cui possibilità variano quindi a seconda degli strumenti e del contesto.

I testi possono presentare codifiche differenti e caratteri di vario tipo, che non sempre le nostre *regex* sapranno trattare.

Soprattutto: attenzione a non *matchare* più del voluto!



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Esercizi!

materiali → <https://bit.ly/2QRCn18>



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

OpenRefine



Flusso ideale

È uno strumento per pulire e ordinare dati sporchi, perché raccolti a mano, non standardizzati e magari provenienti da varie fonti.

Consente di importarli → pulirli e uniformarli → esportarli affinché possano essere usati.

Ogni dataset inserito in OpenRefine è un progetto le cui modifiche vengono tracciate.



Installazione

1. <https://openrefine.org/download.html> → Windows kit
serve Java, se non l'avete e non potete installarlo provate con
Windows kit with embedded Java
2. scompattare lo zip in una cartella a vostra scelta
3. doppio click su `openrefine.exe`
o se non funziona su `refine.bat`
4. si apre una finestra (nera) che potete ignorare,
si aprirà il browser su <http://127.0.0.1:3333/>



Scaletta

- Importare un progetto
- Interfaccia
- Facet/filtri
- Trasformazioni
- Crea colonna da progetto
- Clustering
- Reconciling
- Esportare un progetto



Importare un progetto

1. In “Crea progetto” varie opzioni:
 - carica da file (riconosce in automatico, potete aggiustare impostazioni, codifica etc)
 - appunti (copia/incolla)
 - link a Google Spreadsheet
 - ...
2. Aggiustare le impostazioni, scegliere un nome e cliccare sul bottone “Crea progetto” (alto a dx)

Interfaccia



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Operazioni su colonna

Facet / Filter

Undo / Redo 12 / 15

1001 rows

Extensions: Wikidata

Refresh

Reset All

Remove All

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 10 next > last »

Language

change

4 choices Sort by: name count

Cluster

EN 871

English 107

ES 7

FR 1

(blank) 15

Facet by choice counts

Filtri e faccette

Cronologia

colonna

URL

Date

Language

Subjects

17127853

<https://doaj.org/article/b75e8d5cca3f46cbb>

agriculture5041172

<https://doaj.org/article/0edc5af6672641c0b>

jms161226101

<https://doaj.org/article/d9fe469f75a044238>

norganics3040534

<https://doaj.org/article/95606ed39deb4f43b96f7e6308ad15d3>

01/11/2015

EN

lanthanide|cerium|scorpic non-innocent|Inorganic cl

Text facet

Numeric facet

Timeline facet

Scatterplot facet

Custom text facet...

Custom Numeric Facet...

Customized facets

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Esportare un progetto



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Varie opzioni di esportazione, qui export flessibile in TSV oppure fogli Excel.

Apri... **Esporta** ▼ Aiuto

	Archivia il progetto come file	idata ▼
	Valori separati da Tab	ultimo »
	Valori separati da Virgola	Subjects
aj.org/article cca3f46cbbd63e91be5b3...	Tabella HTML	sher inform
	Excel (.xls)	obabilities
	Excel 2007+ (.xlsx)	
	Foglio elettronico ODF	
aj.org/article 672641e0bd45608812a...	Esportazione tabulare personalizzata...	latoxins A
	Esportazione SQL...	arcinoma /
	Esportazione con modello...	
aj.org/article 75a0442382b84ba4f5000	Archiviando progetto di OpenRefine su	KS NRPS
		enomics u
		nvironmen



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Fine



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Grazie

Piero Grandesso

<perio.grandesso2@unibo.it>

ABIS – Unità di processo Gestione e sviluppo della biblioteca digitale di Ateneo – AlmaDL

www.unibo.it