Hackathon Kit #dataTXT

Prima di tutto questo è un brogliaccio, che si può commentare ed usare come facilitatore. Per partire con il piede giusto, ecco il quadro delle cose che si possono fare con i dati: (preso da vecchi contest sul tema)

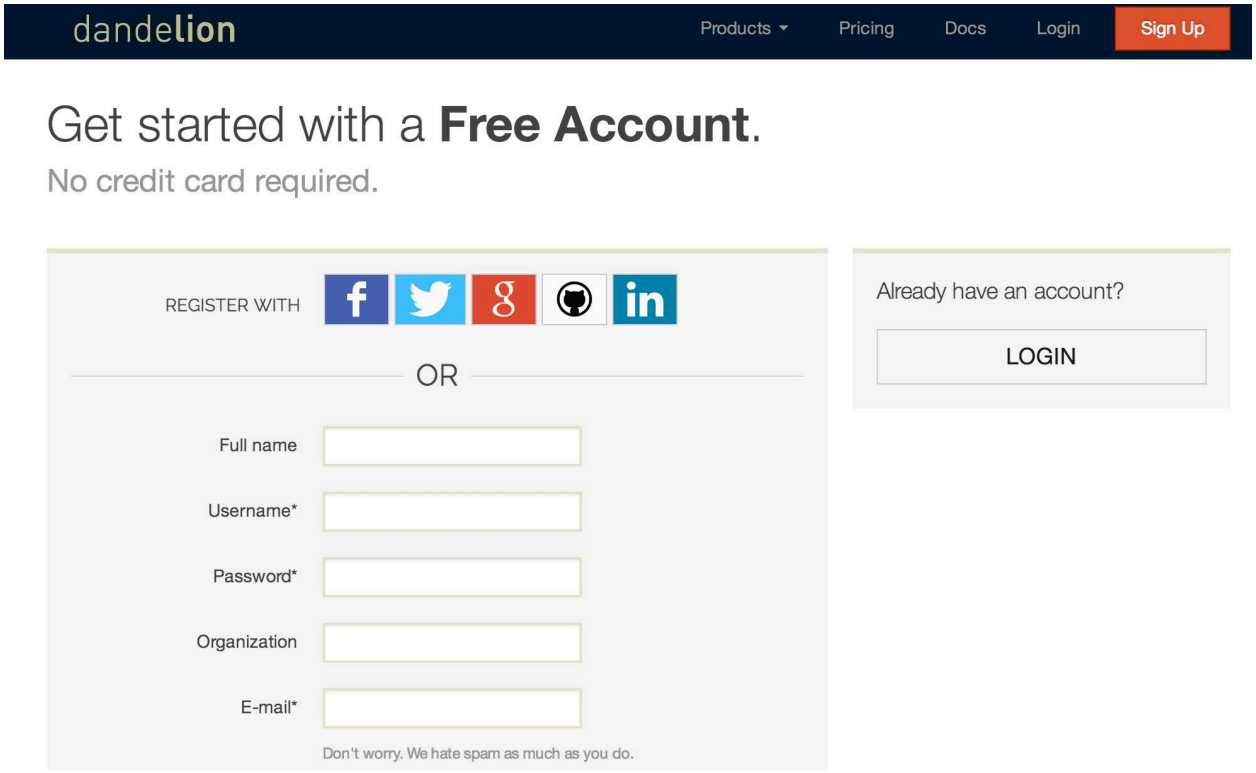
Processo	Descrizione
Data → Fatto Ricerca, Naviga, Estrai	Un dataset viene utilizzato direttamente per identificare un preciso fatto di interesse. Es. Far emergere la storia ed il percorso di voto di un determinato decreto legge...
Data → Informazione Manipolare, analizzare statisticamente, visualizzare, contestualizzare, creare report	Il contenuto di un dataset viene espresso attraverso un'infografica, che è una singola interpretazione di quei dati, ed un'unica rappresentazione. Es. Le infografiche sui dati di dati.piemonte.it realizzate da Visup.it
Data → Interfaccia Pulire, combinare, selezionare, configurare un'interfaccia, scrivere codice personalizzato per la logica, e fornire l'UI	Si fornisce un'interfaccia che permette una rappresentazione interattiva del dataset, sulla base delle azioni dell'utente. Es. mashup su mappa interattiva e ricercabile di dati anche geografici significativi ed integrati con altri
Data → Data Convertire il formato, filtrare i dati, combinare e migliorare il risultato, fornire delle API, ed il dataset per il download	Un dataset derivato, che viene fornito per il download, o per l'accesso via API o similari. Es. dal dataset in XML di Open Camera, al dataset Linked Data con API e SPARQL Endopoint su linkedopendata.it
Data → Servizio Integrare i dati in un prodotto/servizio già esistente, crearne uno nuovo	Si fornisce un servizio basato sugli open data, che non è detto sia pensato per l'utente finale

Cosa dovresti fare, per non perdere tempo dopo

Per usare le API di dataTXT ti servono due cose:
-> app_key e app_id

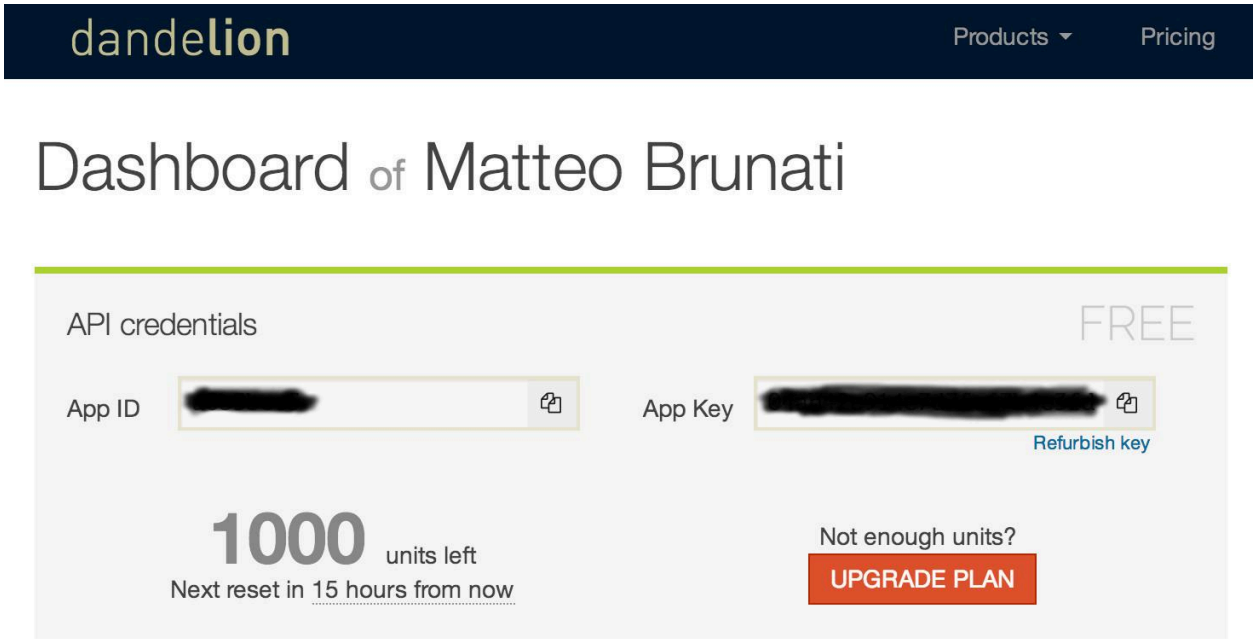
Che poi si usino le REST API, o si usino le integrazioni a Open Refine, o la libreria Python dedicata, poco importa: queste sono necessarie.

2 minuti e le avrai: https://dandelion.eu/accounts/register/



Una volta iscritto, loggati su dandelion.eu, e in alto a destra troverai il tuo username, con un menù a tendina, da dove potrai raggiungere la dashboard:

La tua dashboard avrà i valori dell' app_key e dell'app_id che potrai usare:


dandelion
Products
Pricing
Dashboard of Matteo Brunati
API credentials
FREE
App ID
App Key
Refurbish key
1000 units left
Next reset in 15 hours from now
Not enough units?
UPGRADE PLAN

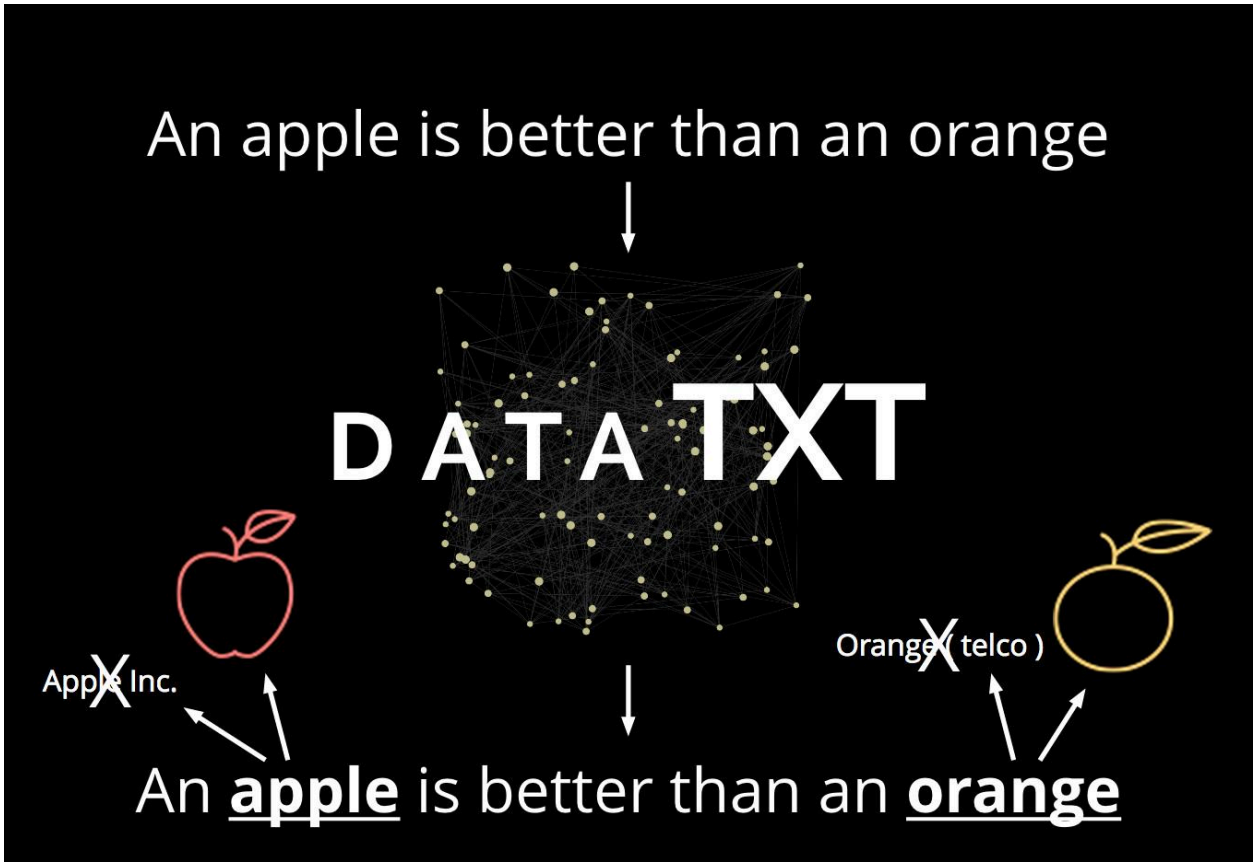
Cosa a fa dataTXT (parte intro, se vuoi il codice, vai avanti)

La famiglia di dataTXT - https://dandelion.eu/products/datatxt/ - fa diverse cose:

- **dataTXT-NEX** - estrae, disambigua e linka alle risorse su Wikipedia
  - demo https://dandelion.eu/products/datatxt/nex/demo/
- **dataTXT-SIM** - calcola la similarità tra due testi, ottimizzata per le frasi brevi
  - demo https://dandelion.eu/products/datatxt/sim/demo/?exec=true

Dettaglio su dataTXT-NEX

Dato in input un testo (una frase o un testo molto lungo), trova le entità presenti nel testo (luoghi, persone, organizzazioni, etc.) e le collega alla corrispondente pagina Wikipedia / DBpedia.



DBpedia è una conversione di Wikipedia in forma di database interrogabile. Si tratta di un progetto di Open Community Data.

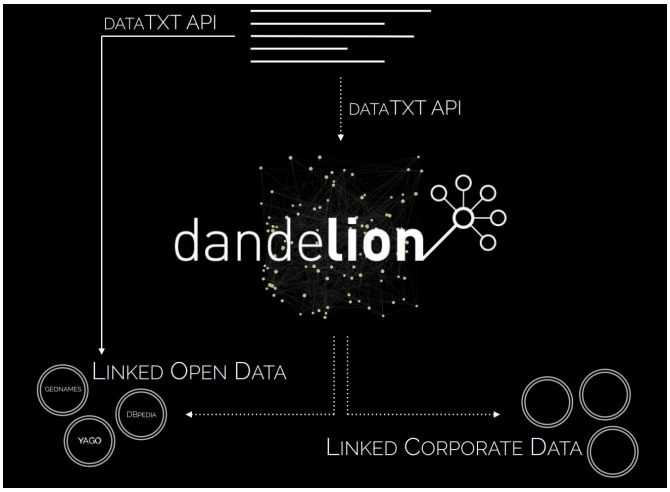
Maggiori informazioni: http://it.dbpedia.org

Un esempio di "entità" su DBpedia: http://it.dbpedia.org/page/Venezia

L'accoppiata delle due tecnologie (DBpedia usa il mondo del Linked Open Data) quindi permette:

- con dataTXT trovi le entità presenti in un testo (un documento, un post su un forum, un tweet, un articolo di giornale, l'imput dell'utente, una colonna di descrizione di una delibera da una fonte Open Data in csv, un URL di un blog post online… etc.) con un link a DBpedia / Wikipedia, ovvero non solo capisci che è un nome di persona, oppure il nome del luogo, ma capisce quale persona e quale luogo, linkandolo *(in gergo sono estratte quindi non con un servizio di Entity Extraction, ma di Named Entity Extraction)*
- Con una query su DBpedia si ottengono maggiori info sull'entità in questione, rendendo più "intelligenti" le applicazioni che consumano questi dati. Ad esempio i luoghi su DBpedia hanno già le coordinate geografiche
- puoi usare l'URI/URL dell'entità come chiave per disambiguare e usarla come identificativo univoco…

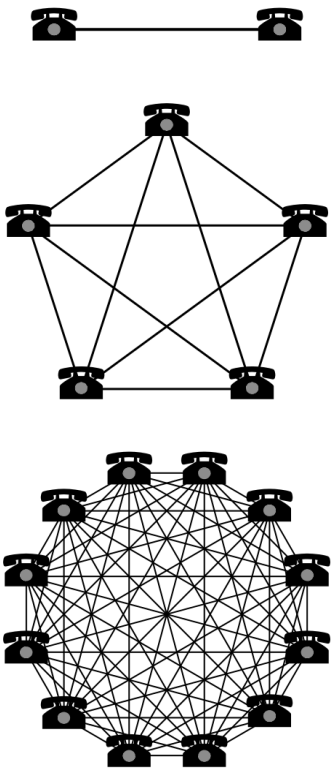
Tutto questo funziona perchè dataTXT si appoggia al grafo sottostante di dandelion.eu, che integra il grafo di Wikipedia, OpenStreetMap ed altre fonti, anche proprietarie:



Perché è importante il contesto?

Perchè Facebook ha così tanto valore oggi? O what's up?
Perchè ha utenti, tanti utenti. E perchè c'è una legge, la legge di MetCalfe che dice:

>> L'utilità e il valore di una rete sono pari ad n^2 - n dove n è il numero degli utenti



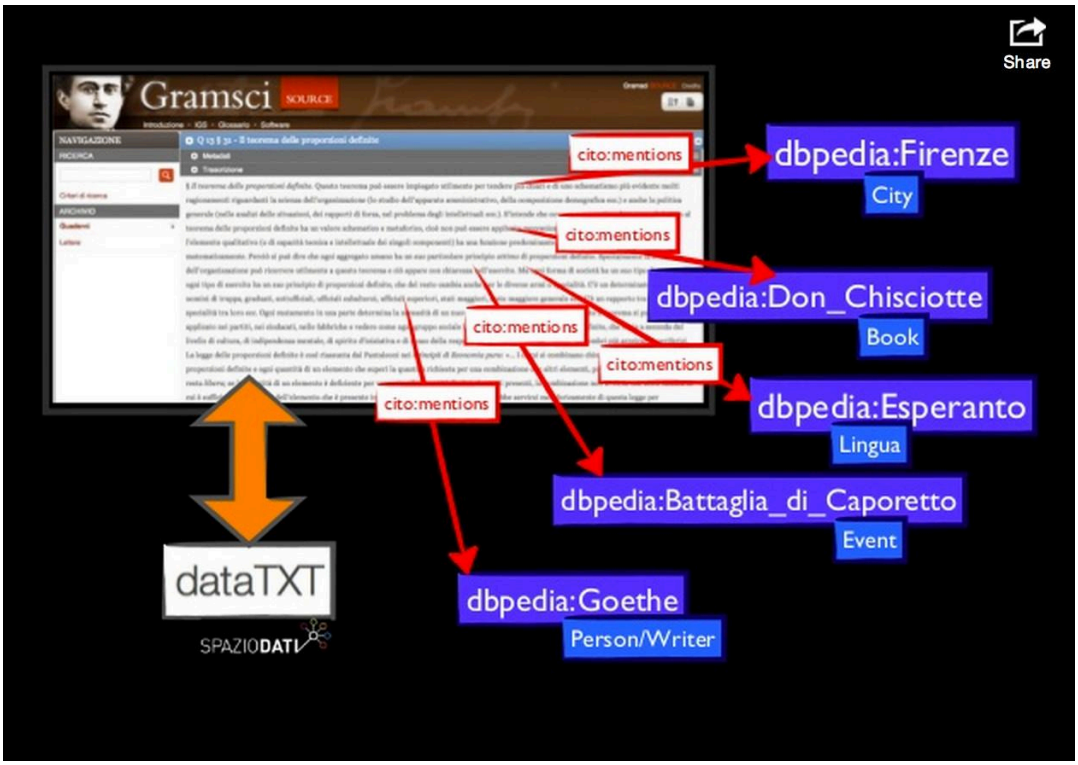
Ed entrambi quei servizi sono social network, ovvero sono delle reti sociali abilitate e costruite attraverso la tecnologia, in un certo senso.

Se lo applichiamo al mondo dei dati, allora significa che:
-> maggior contesto e quindi dati correlati, maggior sarà l'utilità percepita

DATI + LINK + INFO CORRELATE

Ovvero:

testo semplice >> dataTXT >> luoghi
persone
concetti e tag + link a Wikipedia / DBpedia
opere



esempio di utilizzo -> http://www.slideshare.net/christianmorbidoni/gramsci-sourcelod2014-roma

Ah, ovviamente su dandelion.eu esiste una demo a partire dall'input dell'utente o URL, per farvi vedere alcune delle potenzialità:
-> https://dandelion.eu/products/datatxt/nex/demo/?exec=true#results

Occhio che è una semplice interfaccia utente al di sopra dell'API di base.

Cosa fa dataTXT? (codice)

Partiamo dalle basi della doc:
https://dandelion.eu/docs/api/datatxt/nex/getting-started/

ovvero se ho una richiesta API di questo tipo:

```
https://api.dandelion.eu/datatxt/nex/v1/?lang=en&text=The%20doctor%20says%20an%20apple%20is%20better%20than%20an%20orange&include=types%2Cabstract%2Ccategories%2Clod&$app_id=YOUR_APP_ID&$app_key=YOUR_APP_KEY
```

avrò una risposta di questo tipo:
-> https://dandelion.eu/docs/api/datatxt/nex/v1/#example

Cosa ci potrai fare è ora riflessione collettiva :)

Come puoi usare dataTXT: tools

1. Primo livello: usare direttamente le REST dataTXT-NEX API, magari attraverso UniRest - http://unirest.io/ con il linguaggio che preferisci.
2. Secondo livello: usare la libreria nativa in Python
3. Terzo livello: usare Open Refine per pulizia, normalizzazione dei dati e link verso wikipedia da colonne di testo di descrizioni o simili (ha meno senso usarlo per colonne di dati/testi già puliti e di un solo tipo di entità) - un esempio su uno dei dataset presente su dati.veneto.it si trova verso la fine del documento

Dandelion-eu : la libreria ufficiale se usi Python

Come installarla via PyPI, il Python Package Index.

```
pip install dandelion-eu
```

Oppure da GitHub - https://github.com/SpazioDati/python-dandelion-eu
Documentazione di riferimento: http://python-dandelion-eu.readthedocs.org/en/latest/

Impostare le proprie credenziali che vi siete salvati:
http://python-dandelion-eu.readthedocs.org/en/latest/base.html#authentication

```
>>> from dandelion import DataTXT
>>> datatxt = DataTXT(app_id='YOUR_APP_ID', app_key='YOUR_APP_KEY')
>>> response = datatxt.nex('The doctor says an apple is better than an
orange')
>>> for annotation in response.annotations:
      print annotation
```

e poi via, pronto all'uso!

SDK per gli altri linguaggi

Un consiglio spassionato? Usare UniRest - http://unirest.io/ + documentazione ufficiale sulle nostre API di dataTXT.

Casi e scenari di utilizzo per dataTXT

- tutto quello che implica dover gestire testi semplici e renderli “actionable” ovvero collegarli con altri dati, partendo dal contesto delle informazioni correlate tratte da Wikipedia per abilitare filtri, navigazioni a faccette dei dati, o cose del genere.
    - Esempio: integrazione tra DocumentCloud (http://www.documentcloud.org/home) e dataTXT per navigare nei documenti PDF della commissione parlamentare sulla P2.
      -> http://fontitaliarepubblicana.it/ - entrando nel sito si accede alla ricerca

  Quando si clicca sulla voce “vedi le entità” la navigazione e l’estrazione di quelle entità avviene tramite le API di dataTXT.

Scenari

- trovare contenuti correlati, specie a partire da testi brevi provenienti dai social media, per dare maggior contesto ai dati estratti dal mondo Open Government Data, per linkarli e renderli connessi con informazioni aggiornate dal Real-Time Web (Twitter, Facebook status, e simili…)
- abilitare ricerche full text e granulari per tipo di entità presente nel contenuto dei dati da trattare, magari già collegati tra di loro per tema o contesto, per andare oltre al motore tematico offerto da hack4med (che ricerca solo nei metadati associati ai singoli datasets)
- fare data dissemination: sfruttare la diffusione di Wordpress per fare un plugin che permetta la navigazione trasversale di alcuni dei dati pubblicati -> trasparenza, senso civico (e qui ci sono gli atti raccontati attraverso descrizioni che sono filtrabili usando dataTXT, ad esempio…)
- ambito turistico - replicare il widget prototipo di TINDES usando dataTXT per aggregare fonti dal mondo social, ed aumentare l'attualità del valore de dato turistico, o culturale, con dati aggiornati provenienti dagli utenti - http://www.slideshare.net/barbz79it/tindes-esempio-di-riuso-degli-opendata-in-trentino
- preparare dataset puliti e correlati nativamente ad altre fonti: se troviamo una fonte dati molto sporca, oppure che debba essere esplorata, l'uso di Open Refine è fondamentale. Quando poi si ha:
    - una colonna di descrizione, magari di delibere o di descrizioni di luoghi o semplicemente di titoli di libri ad esempio
      -> posso ottenere una selezione di parole chiave da usare nella app o per creare faccette/filtri di navigazione in maniera automatica con l'estensione "Named Entity Extraction" installabile su Open Refine

...

Dove scaricare Open Refine

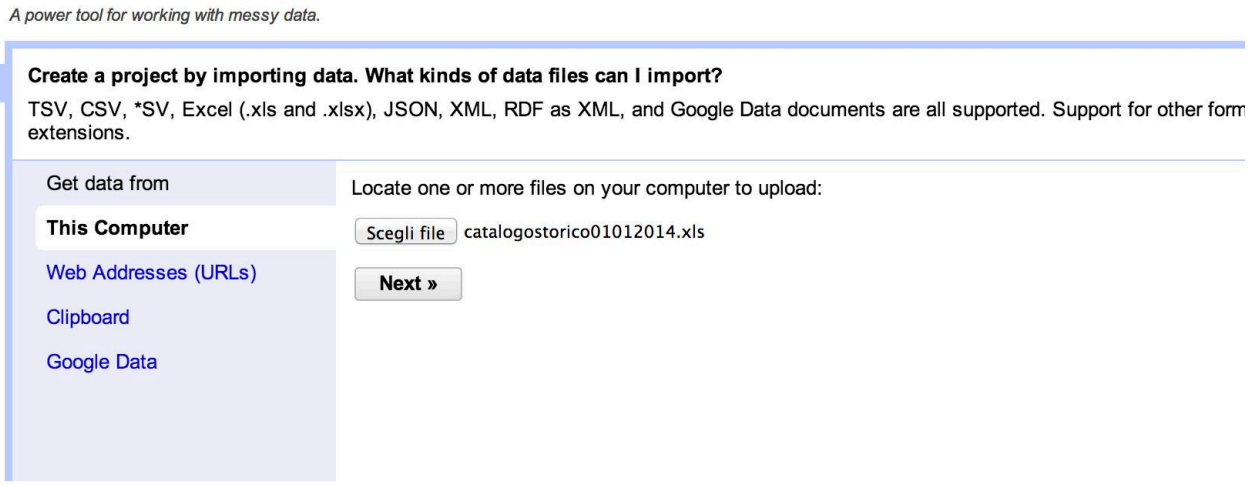
- Open Refine - versione ufficiale - meglio scaricare la stable
  http://openrefine.org/download.html
  (occhio alla versione su macosx, potrebbe avere dei problemi legati alle impostazioni di sicurezza, nel caso:
    - Open System Preferences
    - Open Security & Privacy
    - Go to the General Tab
    - "Allow applications downloaded from:" setting to "Anywhere")

-> va installata questa estensione http://freeyourmetadata.org/named-entity-extraction/

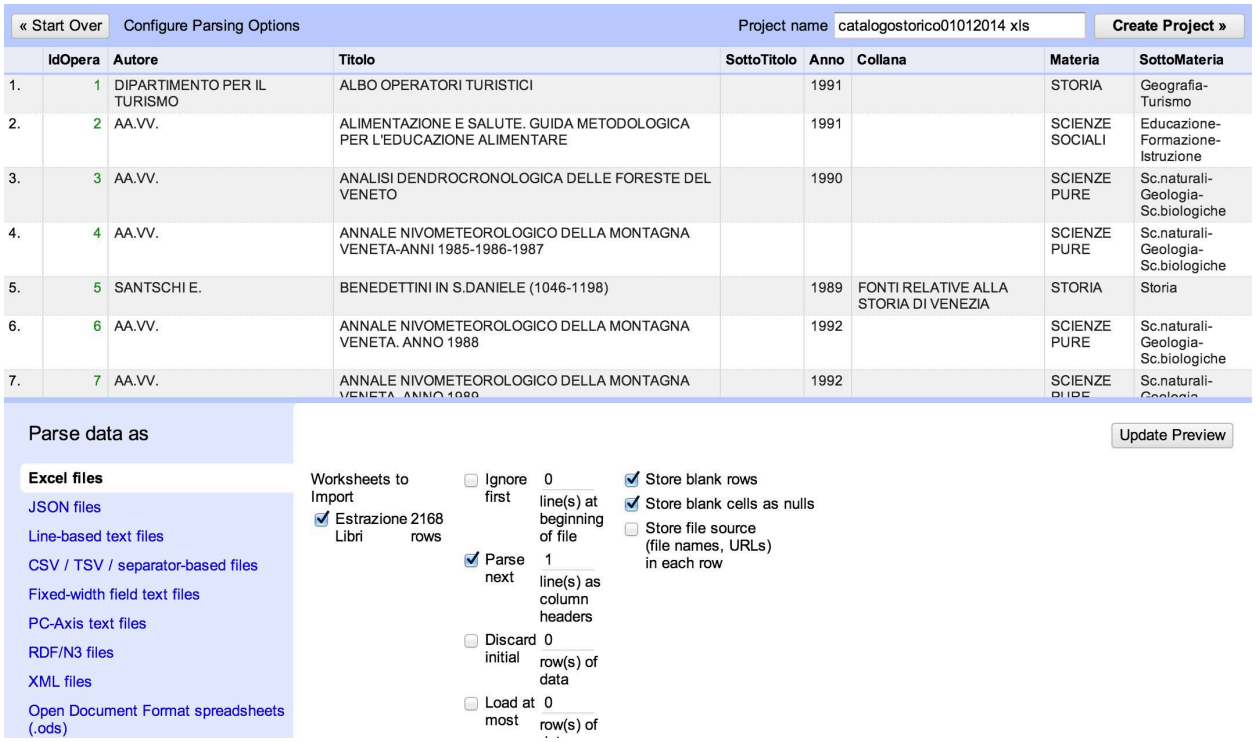
Cosa fa l'estensione "Named Entity Extraction" configurata con dataTXT

Partiamo da esempio operativo su un set di dati con una colonna di testo descrittiva: http://dati.veneto.it/dataset/libri-editi-o-promossi-dalla-regione-del-veneto

1. Scarichiamo il dataset in formato XLS.
2. Apriamo Open Refine e creiamo un nuovo progetto, importando il file:

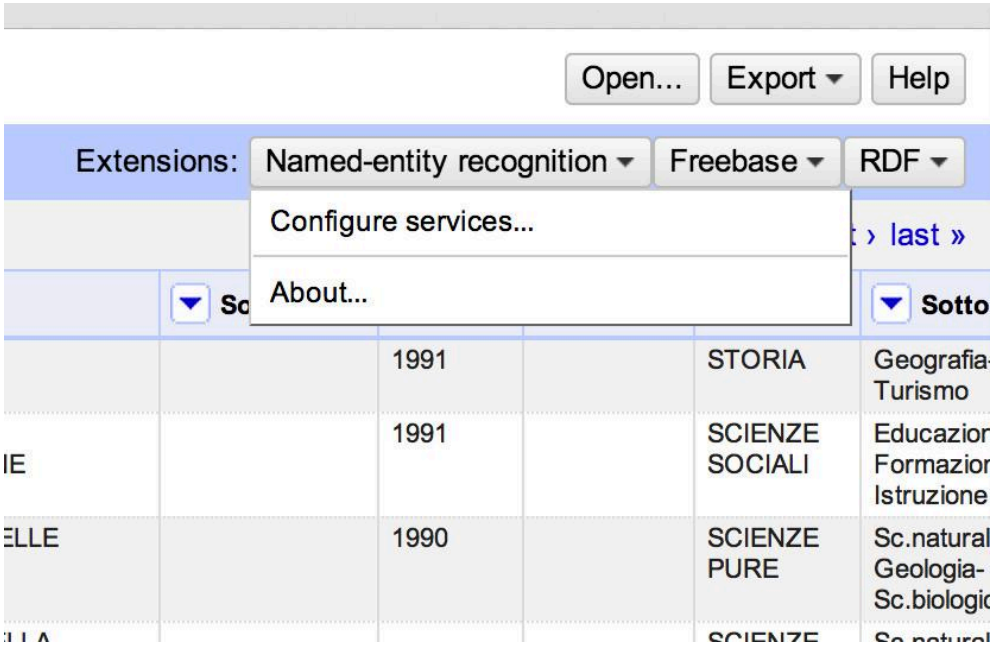


Al passaggio successivo, avremo un'anteprima dei dati presenti nel file:

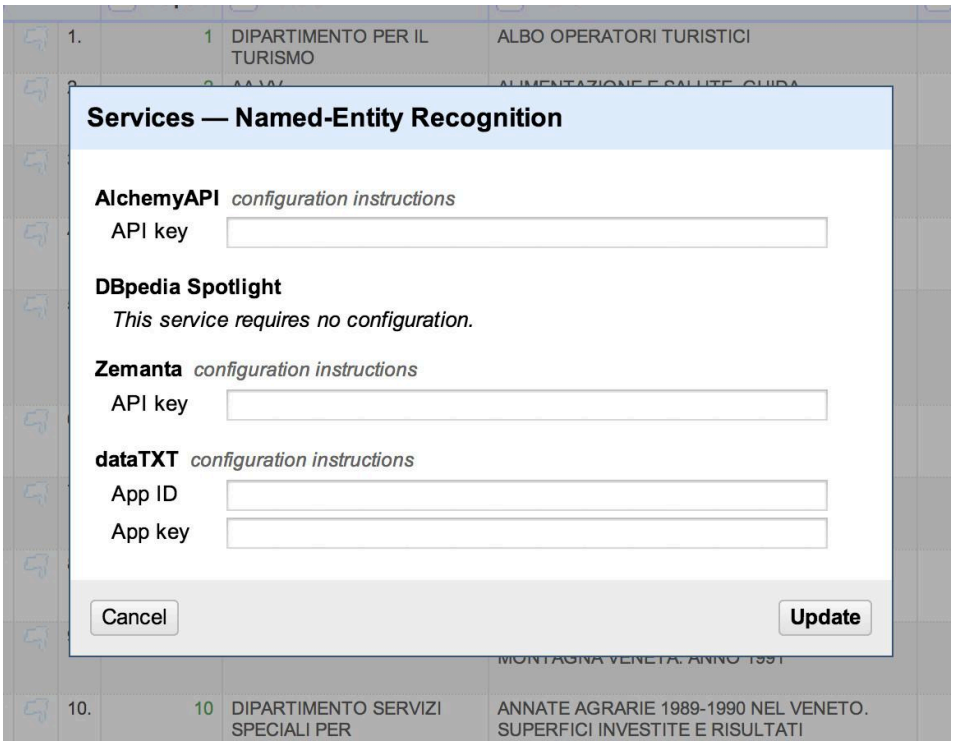


	IdOpera	Autore	Titolo	SottoTitolo	Anno	Collana	Materia	SottoMateria
1.	1	DIPARTIMENTO PER IL TURISMO	ALBO OPERATORI TURISTICI		1991		STORIA	Geografia-Turismo
2.	2	AA.VV.	ALIMENTAZIONE E SALUTE. GUIDA METODOLOGICA PER L'EDUCAZIONE ALIMENTARE		1991		SCIENZE SOCIALI	Educazione-Formazione-Istruzione
3.	3	AA.VV.	ANALISI DENDROCRONOLOGICA DELLE FORESTE DEL VENETO		1990		SCIENZE PURE	Sc.naturali-Geologia-Sc.biologiche
4.	4	AA.VV.	ANNALE NIVOMETEOROLOGICO DELLA MONTAGNA VENETA-ANNI 1985-1986-1987				SCIENZE PURE	Sc.naturali-Geologia-Sc.biologiche
5.	5	SANTSCHI E.	BENEDETTINI IN S.DANIELE (1046-1198)		1989	FONTI RELATIVE ALLA STORIA DI VENEZIA	STORIA	Storia
6.	6	AA.VV.	ANNALE NIVOMETEOROLOGICO DELLA MONTAGNA VENETA. ANNO 1988		1992		SCIENZE PURE	Sc.naturali-Geologia-Sc.biologiche
7.	7	AA.VV.	ANNALE NIVOMETEOROLOGICO DELLA MONTAGNA		1992		SCIENZE	Sc.naturali-

Una volta aperto Open Refine la prima volta con l'estensione installata, dovremo configurare l'estensione con le proprie credenziali (app_key e app_id).
In alto a destra c'è la voce "Extensions: Named-entity recognition":

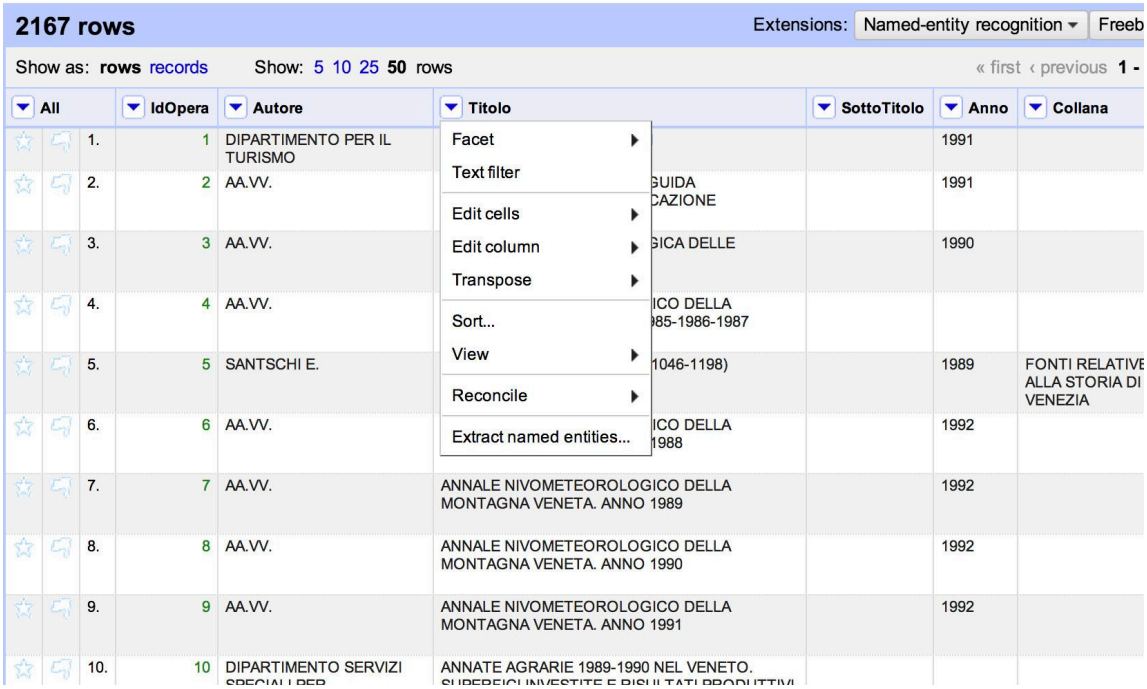


Alla voce "Configure services" arriveremo a questa finestra :

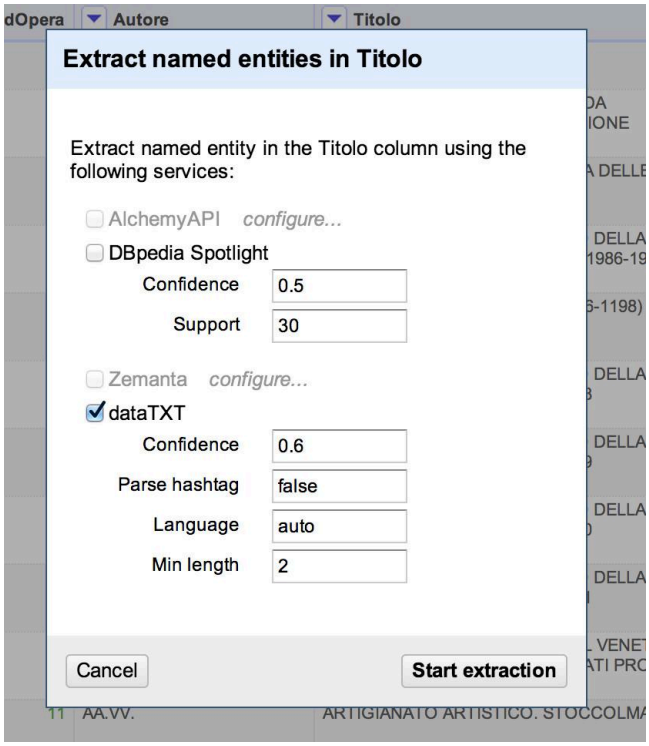


Dovrai quindi inserire le tue “app ID” e “”App key” per abilitare l’uso di dataTXT: se non ricordi dove sono, le trovi su dandelion.eu -> dnome utente in alto a destra -> dashboard.
(come descritto nella sezione iniziale di questo documento)

Una volta quindi configurato il plugin, sarà possibile selezionare quella colonna contenente i testi da filtrare grazie a dataTXT. La colonna del titolo è ottimale: ha testi corti, e con temi specifici. Cliccando sul nome della colonna potremo accedere a questo menù:



Cliccando alla voce “Extract Named entities” apparirà una finestra di questo tipo:

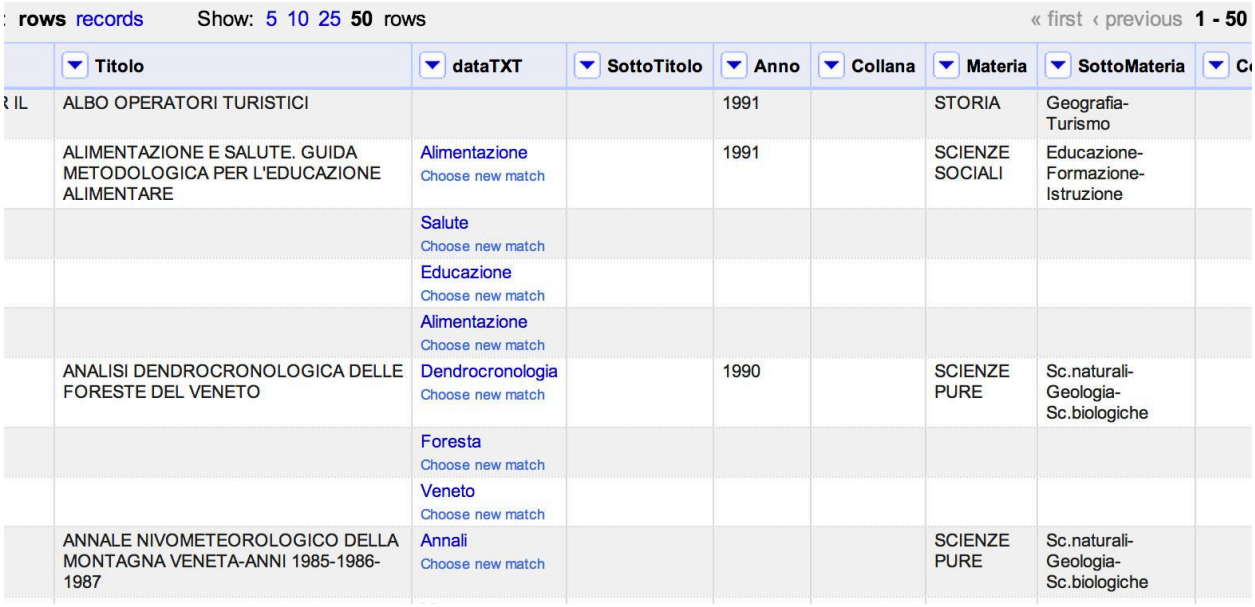


I parametri da configurare sono quelli che si trovano nella doc ufficiale:
-> https://dandelion.eu/docs/api/datatxt/nex/v1/#parameters

Quello più importante è quello della confidence, che rappresenta il valore di qualità del matching con l'entità.
Una volta impostati i valori, cliccando su "Start extraction" otterremo:

- una nuova area che avvisa del lavoro in corso batch sulla colonna "Titolo": in base al numero di righe ed alla complessità/lunghezza del testo da analizzare, sarà un processo più o meno lungo

Una volta terminata l'elaborazione otterremo una nuova colonna titolata "dataTXT":



rows records Show: 5 10 25 **50** rows « first ‹ previous **1 - 50**

	Titolo	dataTXT	SottoTitolo	Anno	Collana	Materia	SottoMateria	C
R IL	ALBO OPERATORI TURISTICI			1991		STORIA	Geografia-Turismo	
	ALIMENTAZIONE E SALUTE. GUIDA METODOLOGICA PER L'EDUCAZIONE ALIMENTARE	Alimentazione Choose new match		1991		SCIENZE SOCIALI	Educazione-Formazione-Istruzione	
		Salute Choose new match						
		Educazione Choose new match						
		Alimentazione Choose new match						
	ANALISI DENDROCRONOLOGICA DELLE FORESTE DEL VENETO	Dendrocronologia Choose new match		1990		SCIENZE PURE	Sc.naturali-Geologia-Sc.biologiche	
		Foresta Choose new match						
		Veneto Choose new match						
	ANNALE NIVOMETEOROLOGICO DELLA MONTAGNA VENETA-ANNI 1985-1986-1987	Annali Choose new match				SCIENZE PURE	Sc.naturali-Geologia-Sc.biologiche	

La cosa interessante è che:

- sono state aggiunte n righe per le n entità trovate all'interno del testo del titolo. A esempio per "Alimentazione e salute. Guida metodologica per l'educazione alimentare", dataTXT ha estratto
    - salute
    - educazione
    - alimentazione (2 volte, per alimentazione e alimentare)
- ogni entità è collegata alla relativa pagina di Wikipedia

Se confrontata con le colonne "Materia" e "Sottomateria" fa riflettere sulle potenzialità di un tagging del genere automatizzato. Senza contare che ci sono i link a Wikipedia che possono avere diversi potenziali utilizzi.

Tutorials ulteriori correlati su Open Refine

Dove si usa e si citano Zemanta, o OpenCalais, si può usare dataTXT-NEX attraverso l'estensione "Named Entity Extraction" appena citata.
Le tecnologie sottostanti utilizzate sono diverse, e producono quindi risultati diversi.

- http://www.slideshare.net/brandwatchsocial/bdb-bethgranter
- http://schoolofdata.org/2013/04/25/social-network-analysis-for-journalists-using-the-twitter-api/ - pensiamo di analizzare il testo di ogni singolo tweet, ad esempio, oltre che gli hashtag

- http://www.slideshare.net/lod2project/lod2-webinarzemanta20120129-16603204
- http://schoolofdata.org/handbook/recipes/cleaning-data-with-refine/