CONFRONTO IMMATRICOLATI-LAUREATI NELLE UNIVERSITA' DI SPAGNA E ITALIA

Come da titolo, l'obbiettivo di questo progetto è confrontare i dati riguardo gli iscritti ed i laureati in Spagna con i dati riguardo gli iscritti ed i laureati in Italia. In particolare, abbiamo trovato dei dataset contenenti informazioni riguardo i corsi di laurea (con la specifica università) e la suddivisione per genere, molto simili quelli spagnoli, la differenza è che non sono riportate le università specifiche, i dati sono stati raccolti in base alle regioni e abbiamo una differenziazione tra università pubbliche e private.

Abbiamo scelto questo argomento per avere un'idea generale di come funziona il sistema universitario di altri paesi, per confrontarlo con il nostro in modo da migliorarlo affinché la formazione ricevuta dall'università sia di pari livello con quella degli altri paesi se non superiore.

DATASETS

Università italiane:

- http://dati.ustat.miur.it/dataset/immatricolati/resource/75da19ea-bc6d-4d0f-8892-3628736f02aa (Immatricolati)
- http://dati.ustat.miur.it/dataset/laureati/resource/21a6312b-29c0-45e0-aa0a-8904653ecc65 (Laureati)

Università spagnole:

- https://datos.gob.es/en/catalogo/a07002862-matriculados-por-ramas-en-universidades-publicas (Immatricolati)
- https://datos.gob.es/en/catalogo/a07002862-evolucion-de-egresados-en-las-universidades (Laureati)

LICENZE

Per quanto riguarda i dataset italiani, osservando i metadati possiamo dire che hanno una licenza open, in particolare: Italian Open Data License v2.0 {*è un contratto di licenza che ha lo scopo di consentire agli utenti di condividere, modificare, usare e riusare liberamente la banca di dati, i dati e le informazioni con essa rilasciati, garantendo al contempo la stessa libertà per altri*}

Per quanto riguarda i dataset spagnoli, abbiamo la licenza CCBY4.0 {*consente di condividere, adattare, trasformare ecc.. il materiale per qualsiasi scopo, anche commerciale. Alle seguenti condizioni: è necessario indicare il proprietario, fornire un collegamento alla licenza e indicare se sono state apportate modifiche. Inoltre, non è possibile applicare termini legali o misure tecnologiche che impediscano legalmente di fare ciò che la licenza consente*}

PULIZIA DEI DATASETS

Iniziamo con i dati delle università italiane: osservando i dataset, non abbiamo ritenuto necessario effettuare pulizie. Abbiamo rimosso dei dati inerenti al vecchio ordinamento e alle lauree del tipo dm 509/99, perché non vi sono corrispondenze con gli immatricolati degli ultimi anni.

Per i dati spagnoli: abbiamo trovato dei dataset in un formato diverso dal csv e con una codifica diversa dal UTF-8. Abbiamo cambiato la codifica, in modo da ottenere un file codificato in UTF-8 e poi abbiamo scritto uno script per effettuare il parsing dei file trovati in modo da costruire dei file in formato csv.

PIPELINE DI ELABORAZIONE

Abbiamo visualizzato i grafici degli immatricolati spagnoli (per singoli corsi) e abbiamo notato che il numero di donne ha un andamento leggermente crescente (circa 4,5%), mentre il numero di uomini presenta una leggera decrescita (circa 2,2%). Stessa operazione per quanto riguarda gli immatricolati italiani, ovviamente con andamenti diversi.

Per quanto riguarda i dataset spagnoli abbiamo eliminato alcune colonne e creato una colonna in cui segnalavamo per ogni regione il tipo di università (pubblica o privata). Infine, abbiamo inserito una colonna (ISCED_f_3) in cui indichiamo il codice del corso, in modo da poter effettuare i confronti con i dati italiani.

Per quanto riguarda i dataset italiani: nel dataset dei laureati abbiamo rimosso le colonne “Lau_M” e “Lau_F” e inserito due colonne: una che indica il genere (maschio o femmina) e una che indica il numero di laureati (per una questione di uniformità con gli altri datasets). Nel dataset degli immatricolati, abbiamo inserito una colonna contenente i codici ISCED.

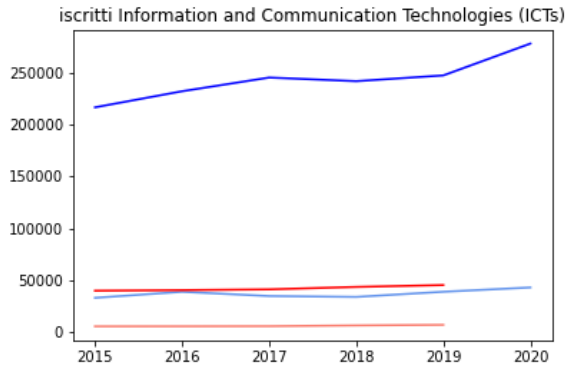
Abbiamo effettuato il join tra i datasets italiani, in modo da ottenere un unico dataset. Stessa cosa per quanto riguarda i datasets spagnoli. In questo modo, possiamo procedere al confronto dei dati dei due paesi.

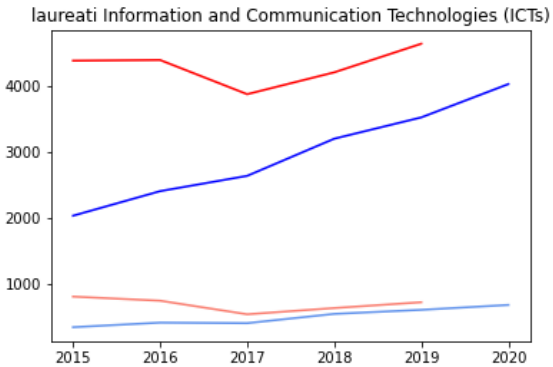
Infine, abbiamo concatenato il dataset finale degli italiani a quello degli spagnoli ottenendo un unico dataset.

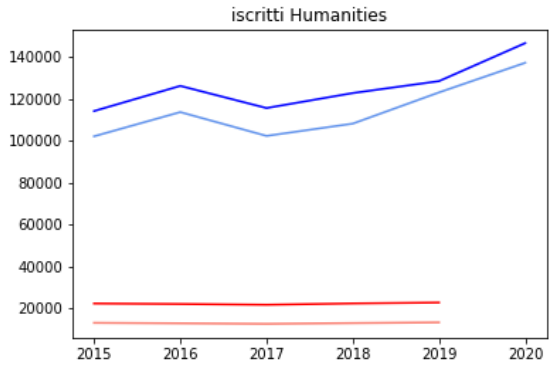
GRAFICI

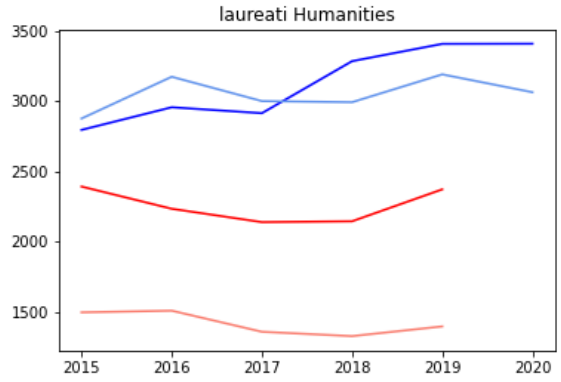
I seguenti grafici sono stati realizzati con la libreria matplotlib di python. Vogliamo confrontare il numero di iscritti/laureati italiani con il numero di iscritti/laureati spagnoli negli anni 2015 – 2020 relativi a 3 corsi di laurea: "Information and Communication Technologies (ICTs)", "Humanities", "Engineering, manufacturing and construction". I trend sembrano simili, il numero di iscritti in Italia è maggiore di quello in Spagna; tuttavia, il rapporto laureati/iscritti spagnoli è superiore, questo è visibile soprattutto nei corsi di ICTs dove i laureati spagnoli sono maggiori in numero assoluto, per entrambi i generi.

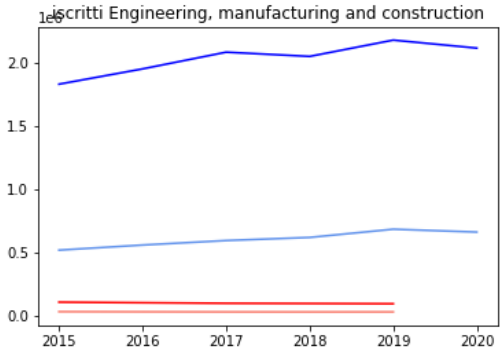
Legenda:

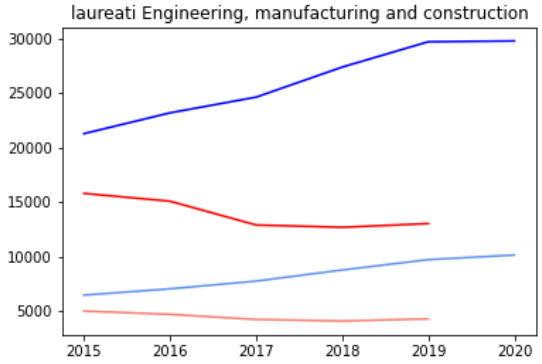












RDF

Abbiamo utilizzato l'ontologia e le risorse di dbpedia per definire i predicati e gli oggetti. Inoltre, abbiamo ritenuto necessario definire (inventare) una URI per il codice ISCED che identifica i settori dei corsi di laurea. Le triple create sono costituite da soggetto, predicato e oggetto. In particolare, il nostro soggetto è il codice ISCED. Vi sono le seguenti triple:

- Una tripla definisce il tipo per il nostro soggetto
- Una tripla indica il numero di iscritti per codice isced
- Una tripla indica il numero di laureati per codice isced
- Una tripla indica il numero di iscritti per anno*
- Una tripla indica il numero di laureati per anno*

Il file di output è un file turtle, questo rappresenta le informazioni in modo che le macchine possano leggerle, in quanto sono strutturate secondo degli standard.

*in queste triple abbiamo deciso di non utilizzare la funzione bind() per far vedere la URI completa relativa al soggetto.

Relazione progetto di open data a cura di: Castagna Andrea, Maniscalco Antonino e Mosca Laura.