Esercitazione sull'uso di `scrapy`

Questa esercitazione guidata riguarda l'uso della libreria Pyhton `scrapy`. Per iniziare, suggeriamo di creare un ambiente isolato che contenga le librerie che servono, ad esempio con:

```
mkdir pippo      # La directory in cui si intende lavorare
cd pippo
pipenv install   # Crea l'ambiente virtuale (solo la prima volta)
pip shell        # Entra nell'ambiente virtuale
pip install jupyter scrapy   # Installa le librerie
jupyter notebook # Avvia jupyter (se serve)
....
exit             # Esce dall'ambiente virtuale
```

Se necessario, consultate la documentazione di scrapy.

1. Obiettivo del progetto

L'obiettivo del progetto è di fare delle statistiche sul numero di volte che i miei studenti di Programmazione per Informatica hanno sostenuto l'esame. Iniziate guardando la mia pagina web: dal menu di sinistra guardate una delle pagine riguardanti il corso di Programmazione per Informatica. Vedrete che nella pagina c'è sempre una tabella contenente gli Appelli, che contiene una colonna con la stringa "Esiti". Cliccando su questa parola si va in una pagina dove c'è una tabella contenente i nomi, i cognomi e i numeri di matricola degli studenti iscritti a quel particolare appello.

2. Scraping

Create un nuovo progetto scrapy, con il comando

```
scrapy startproject nomeprogetto
```

Quindi andate nella directory (come indicato dopo il comando) e generate un template di spider con

```
scrapy genspider nomespider sito
```

Ora modificate manualmente lo spider (che si trova nella directory apposita), cambiando intanto l'elenco dei seed. Mettete come seed gli URL dei corsi di programmazione per tutti gli anni in cui l'ho insegnata.

A questo punto limitatevi a isolare l'URL (attributo href dell'elemento a) associato alla parola "Esiti". Fate un yield che si limiti a restituire un dizionario con una sola chiave associata a quell'URL.

Verificate che lo scraper funzioni come vi aspettate.

3. Scraping

Ora aggiungete una funzione per fare parsing della pagina corrispondente all'URL in modo da estrarre i nomi, i cognomi e le matricole dalla tabella.

Fate restituire un dizionario contenente nomi, cognomi, matricole e url. Fate funzionare lo scraper salvando i risultati in formato JSON.

Guardate il risultato:

- ci sono anche cose che riguardano i compitini; come potete eliminarle? [Suggerimento: fate il parsing di una pagina solo se l'URL non contiene la sottostringa...]
- ci sono caratteri spuri nei cognomi? se sì, eliminateli usando il metodo strip() di str [Attenzione, dovrete convertire esplicitamente nome, cognome ecc. in str]

4. Analisi

Quando avrete un file JSON corretto, aprite un notebook jupyter e caricate il file JSON. Vi conviene creare una mappa che mappi i numeri di matricola in coppie (nome, cognome).

Esiste in Python, nel pacchetto collections una classe Counter che serve per contare quante volte ogni elemento compare in una lista. Un Counter restituisce una lista di coppie, in cui il primo elemento è l'oggetto che si ripete e il secondo è il numero di ripetizioni

Create una lista con i numeri di matricola dei record JSON e usate un Counter per contare le matricole che compaiono più spesso.

Trovate ora un modo (un metodo di Counter) per avere solo gli elementi che si ripetono più spesso (diciamo, per fissare le idee, i 10 più frequenti).

Usando questo metodo trovate le matricole che hanno ripetuto l'esame più volte. Usando ora la mappa precedentemente create stampate i nomi e i cognomi degli studenti corrispondenti, e quante volte hanno ripetuto l'esame.