**RIASSUNTO INCONTRO VENERDI’ 23 LUGLIO**

Nella prima parte dell’incontro con il professor Pianzola abbiamo visto ed analizzato insieme il codice HTML delle pagine relative ai siti coreani NAVER e YES24 al fine di poter iniziare ad approcciare il problema dello scraping delle reviews coreane. Questa parte non verrà approfondita in questo report dal momento che il professor Pianzola condividerà degli appunti dettagliati relativi a quanto fatto e visto a livello tecnico. Mi limito tuttavia a riportare alcuni dubbi sorti e proposte relative alla fase di scraping:

* I dati delle reviews di Amazon e Goodreads fornitici dal professore Pianzola all’interno del suo Planning arrivano fino al dicembre 2018 e al dicembre 2019 rispettivamente. Dunque, si sono aperte due strade possibili: integrare quei dati con uno scraping mirato sulle reviews di Amazon e Goodreads per gli anni 2019,2020 e 2021 pur sapendo che questa operazione comporterà ulteriore tempo da impiegare, oppure limitare la analisi fino all’anno 2019 (il che ci sembra essere un po’ limitante). Includere anche i dati del 20-21 aprirebbe inoltre alla possibilità di analizzare anche differenze, di lessico e di tono ad esempio, nelle reviews prima e dopo il periodo covid. Lei cosa ne pensa?
* Prima di implementare un codice di scraping che sia definitivo (al quale lavorerò la settimana prossima, prima di partire il 30 Luglio) occorre definire la lista dei generi letterari di analisi, i titoli dei libri bestellers che si vogliono analizzare ed il quantitativo degli stessi. Di questo lavoro si occuperà il professor Pianzola la prossima settimana. L’idea era quella di selezionare una 50ina circa di bestsellers per lingua d’origine italiana, coreana e inglese (totale 150 circa) in modo tale da analizzare le differenze di lessico, attitudini, atteggiamenti e quant’altro per le reviews di ogni lingua. In aggiunta si pensava di inserire anche una lista comune di bestellers (una 50ina anche in questo caso) tra le 3 lingue, così da poter fare confronti sullo stesso materiale.
* Problematiche relative alle distribuzioni dei rating delle reviews sui siti coreani. Di solito viene fornita una media del rating fornito tramite le reviews per un determinato libro, senza accennare alle distribuzioni. Questo non permette di rispondere alla seguente domanda di ricerca proposta dal professor Pianzola: *determine the distribution of stars for all ratings and averaged by book, for each genre and language.*

Dunque, a questo punto le priorità potrebbero essere:

* Selezionare le liste di bestsellers da analizzare
* Implementare del codice per lo scraping dai siti coreani (e successivamente da Amazon e Goodreads qualora decidessimo di intraprendere quella strada)
* Definire la questione pre-registrazione (della quale non si è però parlato nell’incontro, se decideremo di intraprendere quella via)