Secondo progetto di Social Computing 2020-2021

Francesco Bombassei De Bona (144665)
Andrea Cantarutti (141808)
Alessandro Zanatta (143154)

17 gennaio 2021

1 Introduzione

Il seguente elaborato espone il processo di analisi dei dati ottenuti in seguito al dispiegamento di un task di Crowdsourcing, ottenuti attraverso le metodologie offerte dal framework *Crowd Frame*. In particolare, il lavoro è stato suddiviso nelle seguenti fasi:

- Ottenimento dei dati serializzati su un bucket Amazon S3
- Analisi dei dati e produzioni di grafici esplicativi

Si noti che è possibile visualizzare una **versione interattiva** di ogni grafico presente nel seguito dell'elaborato cliccandoci sopra.

2 Recupero dei dati

I dati relativi allo svolgimento del task da parte di ogni singolo worker sono stati serializzati direttamente da *Crowdframe* in diversi documenti JSON all'interno del bucket Amazon S3 predisposto. I dati raccolti sono stati scaricati tramite gli script descritti nell'elaborato precedente.

Il task è stato svolto da un totale di **54 worker**, dei quali 6 appartenenti rispettivamente a due gruppi coinvolti nello svolgimento del medesimo task e 48 reclutati al fine di incrementare la numerosità dei dati acquisiti.

Per ogni worker, è stata scaricata una directory contenente i dati di completamente del task (denominata in base all'id del worker). L'insieme dei dati raccolti è situato all'interno della directory `Data`.

3 Analisi dei dati raccolti

L'analisi dei dati è contenuta all'interno della directory `pyAnalysis` ed è stata svolta all'interno di un apposito Jupyter Notebook.

3.1 Osservazione dei dati ottenuti dal questionario

Al fine di rendere possibile una visualizzazione d'insieme in merito alle compilazione del questionario introduttivo, è stato prodotto il seguente diagramma "sunburst".

Inoltre, sono stati prodotti istogrammi e diagrammi a barre relativi alla frequenza relativa di ogni variabile campionata per mezzo del questionario.

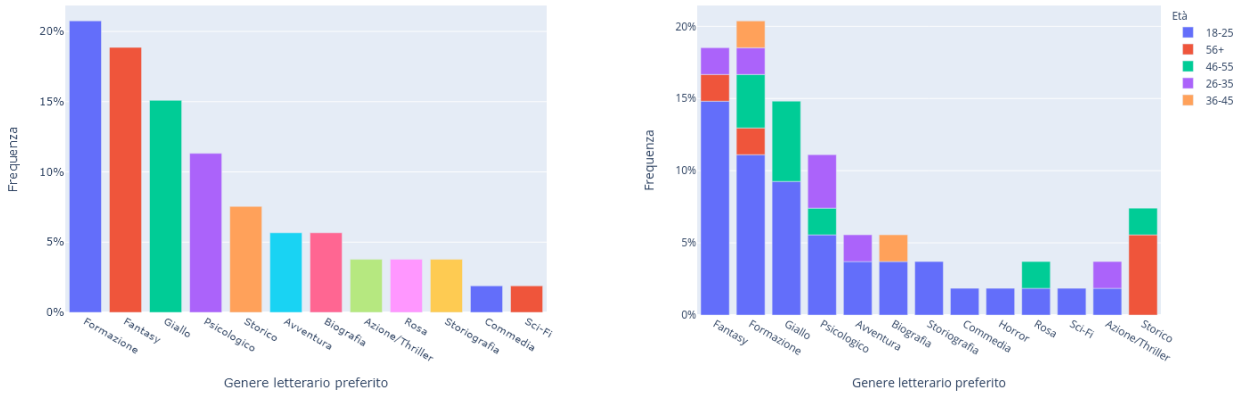

Figura 1: Barplot relativo al genere letterario preferito, anche in relazione all'età

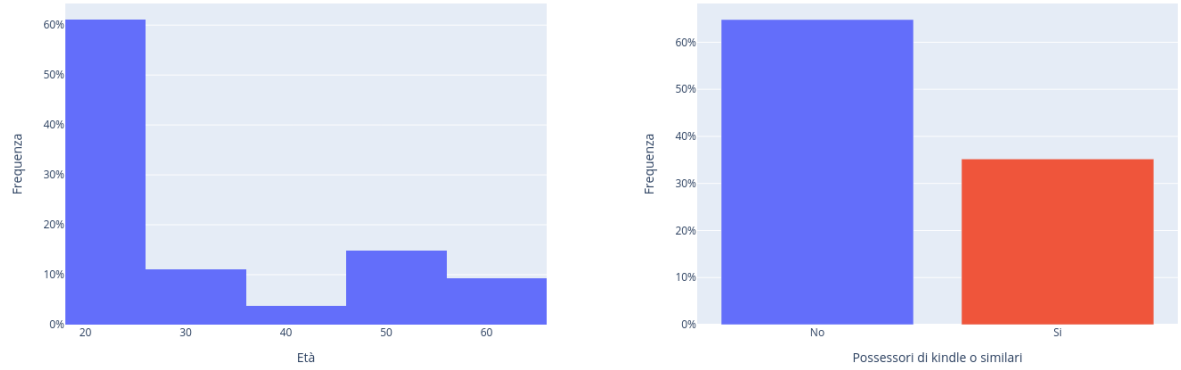

Figura 2: Istogramma in relazione all'età e barplot relativo ai possessori di *eBook Reader*

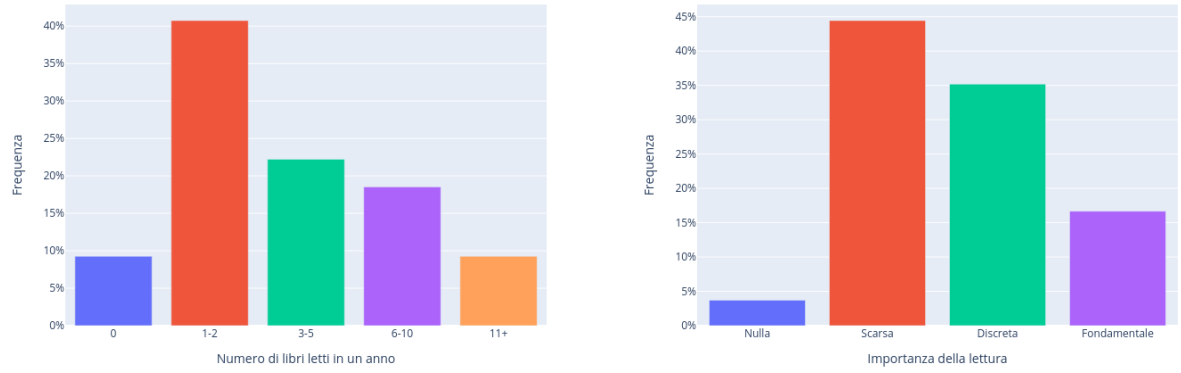

Figura 3: Barplot relativi alle preferenze in relazione alla lettura

Si è osservato come i worker, prevalentemente studenti di età compresa fra i 18 e i 26 anni, abbiano dimostrato un complessivo interesse verso la lettura. I tre generi letterari preferiti sono risultati essere rispettivamente Formazione, Fantasy (principalmente interessato dai worker giovani) e Giallo. Più del 60% dei worker ha, inoltre, dichiarato di **non possedere strumenti per la lettura digitale**.

3.2 Osservazione dei dati relativi alle dimensioni proposte

3.2.1 Grado medio di adeguatezza del prezzo

Durante lo svolgimento di ogni singolo HIT, ai worker è stato richiesto di esprimere un giudizio sull'adeguatezza del prezzo dei libri in analisi. In particolare, sono state predisposte le seguenti dimensioni:

- *Il prezzo ti sembra adeguato?* (sì/no)
- *Indica quanto il prezzo ti sembra adeguato* (slider con valori compresi fra 0 e 5, con 0 completamente inadeguato e 5 del tutto adeguato).

Sulla base dei dati ottenuti dai worker, sono state calcolate opportune statistiche riassuntive al fine di osservare il grado (massimo, minimo e medio) di adeguatezza del prezzo in relazione ad ogni *singola edizione* e, successivamente, ad *ognuno dei titoli proposti*.

Si osserva, nel seguente schema riassuntivo, la **media** calcolata in relazione all'adeguatezza del prezzo.

```
\begin{table}[H]
\centering
\def\arraystretch{1.3}
\begin{tabular}{|r|c|c|}
\hline
\multicolumn{1}{|c|}{\textbf{Edizione}} & \textbf{Il prezzo è adeguato?} & \textbf{Quanto è adeguato?} \\
Le cronache di Narnia (italiano, cartaceo)          & 0.944444 & 3.944444 \\ \hline
Le cronache di Narnia (italiano, eBook)             & 0.777778 & 3.777778 \\ \hline
The Chronicles of Narnia (inglese, cartaceo)        & 0.833333 & 3.222222 \\ \hline
Assassinio sull'Orient Express (italiano, cartaceo) & 0.833333 & 3.666667 \\ \hline
Assassinio sull'Orient Express (italiano, eBook)    & 0.722222 & 3.444444 \\ \hline
Assassinio sull'Orient Express(inglese, cartaceo)   & 0.833333 & 3.888889 \\ \hline
Così parlò Zarathustra (italiano, cartaceo)         & 0.833333 & 3.666667 \\ \hline
Così parlò Zarathustra (italiano, eBook)            & 0.722222 & 3.666667 \\ \hline
Così parlò Zarathustra (inglese, cartaceo)          & 0.777778 & 3.611111 \\ \hline
\end{tabular}
\end{table}
```

Di seguito si osserva, invece, uno studio relativo alla **mediana** del livello di adeguatezza del prezzo.

```
\begin{table}[H]
\centering
\def\arraystretch{1.3}
\begin{tabular}{|r|c|c|}
\hline
\multicolumn{1}{|c|}{\textbf{Edizione}} & \multicolumn{1}{c|}{\textbf{Il prezzo è adeguato?}} & \multicolu
Le cronache di Narnia (italiano, cartaceo)          & 1.0 & 4.0 \\ \hline
Le cronache di Narnia (italiano, eBook)             & 1.0 & 4.0 \\ \hline
Le cronache di Narnia (inglese, cartaceo)           & 1.0 & 3.0 \\ \hline
Assassinio sull'Orient Express (italiano, cartaceo) & 1.0 & 4.0 \\ \hline
Assassinio sull'Orient Express (italiano, eBook)    & 1.0 & 4.0 \\ \hline
Assassinio sull'Orient Express(inglese, cartaceo)   & 1.0 & 4.0 \\ \hline
Così parlò Zarathustra (italiano, cartaceo)         & 1.0 & 3.5 \\ \hline
Così parlò Zarathustra (italiano, eBook)            & 1.0 & 4.0 \\ \hline
Così parlò Zarathustra (inglese, cartaceo)          & 1.0 & 3.5 \\ \hline
\end{tabular}
\end{table}
```

Per ogni edizione presa in analisi sono stati, inoltre, calcolati indicatori quali **massimo** e **minimo** in relazione all'espressione di adeguatezza permessa dalle due dimensioni.

```
\begin{table}[H]
\centering
\renewcommand\extrarowheight{6pt}
\begin{tabular}{c|r|c|c|c|c|}

\cline{3-6}
    \multicolumn{2}{l|}{} &
    \multicolumn{2}{c|}{\textbf{Il prezzo è adeguato?}} &
    \multicolumn{2}{c|}{\textbf{Quanto è adeguato?}}

\\ \cline{2-6}

    \multicolumn{1}{l|}{} &
    \multicolumn{1}{c|}{\textbf{Libro}} &
    \textbf{Tipo} &
    \textbf{Valore} &
    \textbf{Tipo} &
    \multicolumn{1}{c|}{\textbf{Valore}}

\\ \hline

    \multicolumn{1}{|c|}{\multirow{3}{*}{\rotatebox[origin=c]{90}{\textbf{Massimo}}}} &
    Le cronache di Narnia &
    Italiano, cartaceo &
    0.944444 &
    Italiano, cartaceo &
    3.944444

\\ \cline{2-6}

    \multicolumn{1}{|c|}{} &
    Assassinio sull'Orient Express &
    Italiano, cartaceo &
    0.833333 &
    Inglese, cartaceo &
    3.888889

\\ \cline{2-6}

    \multicolumn{1}{|c|}{} &
    Così parlò Zarathustra &
    Italiano, cartaceo &
    0.833333 &
    Italiano, cartaceo &
    3.666667

\\ \hline

    \multicolumn{1}{|c|}{\multirow{3}{*}{\rotatebox[origin=c]{90}{\textbf{Minimo}}}} &
    Le cronache di Narnia &
    Italiano, eBook &
    0.777778 &
    Inglese, cartaceo &
    3.222222
```

```
\\ \cline{2-6}

    \multicolumn{1}{|c|}{} &
    Assassinio sull'Orient Express &
    Italiano, eBook &
    0.722222 &
    Italiano, eBook &
    3.444444

\\ \cline{2-6}

    \multicolumn{1}{|c|}{} &
    Così parlò Zarathustra &
    Italiano, cartaceo &
    0.722222 &
    Italiano, eBook &
    3.611111

\\ \hline

\end{tabular}
\end{table}
```

3.2.2 Grado medio, mediana e deviazione standard di adeguatezza del prezzo (per libro)

- Il prezzo è adeguato?

```
\begin{table}[H]
\caption{Aggregazione rispetto alla domanda \textit{Il prezzo è adeguato?}}
\centering
\def\arraystretch{1.3}
\begin{tabular}{r|c|c|c|}
\cline{2-4}
\multicolumn{1}{c|}{} & \textbf{Le cronache di Narnia} & \textbf{Assassinio sull'Orient Express} & \textbf
\multicolumn{1}{|r|}{\textbf{Media}}      & 0.851852 & 0.796296 & 0.777778 \\ \hline
\multicolumn{1}{|r|}{\textbf{Mediana}}    & 1.000000 & 1.000000 & 1.000000 \\ \hline
\multicolumn{1}{|r|}{\textbf{Deviazione}} & 0.358583 & 0.406533 & 0.419643 \\ \hline
\end{table}
```

Il libro con il grado medio di adeguatezza del prezzo è, secondo la metrica sopracitata, *Le cronache di Narnia*.

- Quanto adeguato è il prezzo?

```
\begin{table}[H]
\caption{Aggregazione rispetto alla domanda \textit{Quando adeguato è il prezzo?}}
\centering
\def\arraystretch{1.3}
\begin{tabular}{r|c|c|c|}
\cline{2-4}
\multicolumn{1}{c|}{} & \textbf{Le cronache di Narnia} & \textbf{Assassinio sull'Orient Express} & \textbf
\multicolumn{1}{|r|}{\textbf{Media}}      & 3.648148 & 3.666667 & 3.648148 \\ \hline
\multicolumn{1}{|r|}{\textbf{Mediana}}    & 4.000000 & 4.000000 & 4.000000 \\ \hline
\multicolumn{1}{|r|}{\textbf{Deviazione}} & 1.066778 & 1.243853 & 1.519153 \\ \hline
\end{tabular}
\end{table}
```

Il libro con il grado medio di adeguatezza del prezzo è, secondo la dimensione di cui sopra, *Assassinio sull'Orient Express*.

3.2.3 Giustificazioni dei worker

3.2.4 Analisi delle giustificazioni fornite dai worker

Sono state estratte le giustificazioni scritte fornite dai worker in relazione ad ogni HIT e sono state determinate:

- La **lunghezza media** delle giustificazioni fornite
- La giustificazione con **lunghezza massima**
- La giustificazione con **lunghezza minima**

Si riportano di seguito la giustificazione più lunga e più corta:

```
\begin{table}[H]
\centering
\renewcommand\extrarowheight{4pt}
\begin{tabular}{c|c|l|}
\cline{2-3}
\multicolumn{1}{l|}{} &
  \multicolumn{1}{c|}{Lunghezza} &
  \multicolumn{1}{c|}{Giustificazione} \\ \hline
\multicolumn{1}{|c|}{Più lunga} &
  70 &
  \begin{tabular}[c]{@{}l@{}}Viene spesso detto di non giudicare il libro dalla copertina ma la realtà \\
  secondo me è diversa, molto spesso in una libreria i libri che ci attirano si piú \\
  sono quelli con una copertina accattivante, elaborata e particolare. \\
  La storia narrata è sicuramente di mio gradimento e su questo non \\ ho nulla da dire. Avendolo già lett
```

```
\multicolumn{1}{|c|}{Più corta} &
  11 &
  Trovo il prezzo un po' elevato trattandosi di una versione digitale \\ \hline
\end{tabular}
\end{table}
```

3.2.5 Analisi delle giustificazioni fornite dai worker

In seguito, sono state estratte le giustificazioni scritte fornite dai worker in relazione ad ogni HIT e sono state determinate:

- La **lunghezza media** delle giustificazioni fornite
- La giustificazione con **lunghezza massima**
- La giustificazione con **lunghezza minima**

In particolare, il conteggio è stato basato sul numero di parole contenute in una singola giustificazione piuttosto che sul numero di caratteri, al fine di ottenere risultati strettamente correlati alla **verbosità** del testo.

3.2.5.1 Lunghezza media La lunghezza media delle giustificazioni è risultata essere pari a **17.29 parole**, più alta della lunghezza minima richiesta durante lo svolgimento del task. Si è osservato, di conseguenza, un'interesse e un'interazione da parte dei worker che hanno svolto il task.

3.2.5.2 Lunghezza massima La giustificazione con lunghezza massima ha presentato un totale di **70 parole**. Il suo contenuto è di seguito riportato:

> "Viene spesso detto di non giudicare il libro dalla copertina ma la realtà secondo me è diversa, molto spesso in una libreria i libri che ci attirano si piú sono quelli con una copertina accattivante, elaborata e particolare. La storia narrata è sicuramente di mio gradimento e su questo non ho nulla da dire. Avendolo già letto lo consiglierei o acquisterei volentieri ma, per attrarre maggior clientela si dovrebbe migliorare la facciata."

3.2.5.3 Lunghezza minima La giustificazione con lunghezza minima ha, invece, presentato una lunghezza pari a **11 parole**. Il contenuto è il seguente:

> "Trovo il prezzo un po' elevato trattandosi di una versione digitale"

La lunghezza media delle giustificazioni è risultata essere pari a **17.29 parole**, più alta della lunghezza minima richiesta durante lo svolgimento del task. Si è osservato, di conseguenza, un'interesse e un'interazione da parte dei worker nei confronti del task svolto.

3.2.6 Analisi di correlazione

È stato, successivamente, analizzato il livello di correlazione fra le diverse dimensioni proposte ai worker. Sulla base dei dati ottenuti, è stata realizzata la seguente *heatmap*.

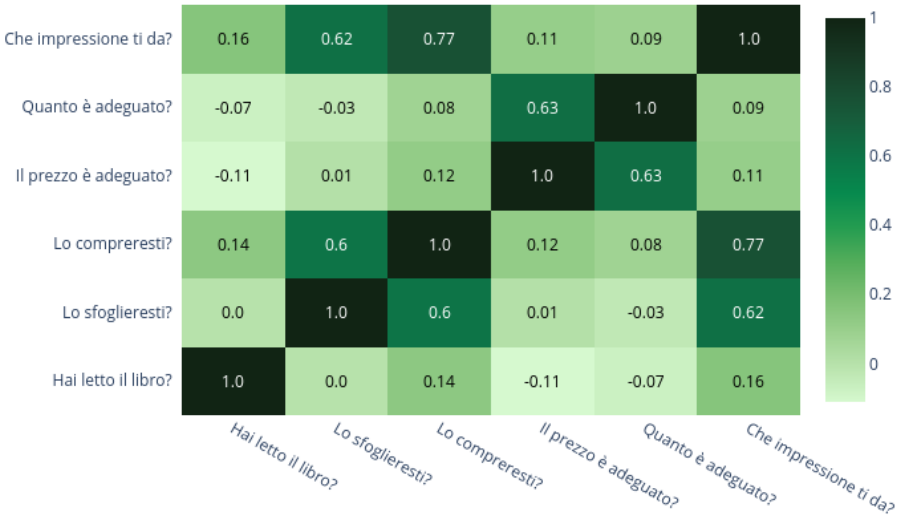

Figura 4: Heatmap illustrante la correlazione fra dimensioni

Si osserva, in particolare, la presenza di un buon grado di correlazione fra dimensioni quali l'**impressione data dal libro** e il **desiderio di sfogliarlo** o **acquistarlo**, oppure l'**adeguatezza** e il **grado di adeguatezza**. I dati ottenuti dimostrano uno svolgimento coerente del task da parte dei worker.

Nello specifico, sono stati studiati due aspetti rilevanti:

- Il legame fra la lingua di un libro e il desiderio di acquistarlo
- Il legame fra l'interesse verso un'edizione digitale e il possesso di un eBook Reader

Nel primo caso, si è osservato, nonostante tutti i worker fossero di nazionalità italiana, un **basso grado di correlazione positiva** (pari a 0.026) che suggerisce una lieve tendenza ad un maggiore interesse per i libri in lingua inglese.

Nel secondo, è stata aggregata la risposta alla domanda *"Possiedi un eBook reader (Kindle, KoBo, . . .) o utilizzi un'applicazione per la lettura di libri digitali?"* presente nel questionario iniziale. Si sono osservate:

- Una minima correlazione negativa (-0.06) fra il possesso di un eBook Reader e il desiderio di acquistare un'edizione digitale
- Una bassa **correlazione positiva** (0.16) fra il possesso di un eBook Reader e il desiderio di acquistare un'edizione cartacea.

Si osserva, di conseguenza, come il possesso di uno strumento per la lettura di edizioni digitali non abbia avuto un'influenza decisiva nelle risposte ottenute dai worker.

4 Conclusioni

Il processo descritto nel primo elaborato, assieme al recupero e all'analisi dei dati raccolti, hanno permesso la conduzione di una completo esperimento di crowd-sourcing secondo le metolodie offerte da *Crowd Frame*. Grazie ai 54 svolgimenti ottenuti, inoltre, è stato possibile effettuare un'analisi efficace.