JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

JLIS.it

Text mining e network science per analizzare la complessità della lettura. Principi, metodi, esperienze di applicazione.

Chiara Faggiolani [a] **Lorenzo Verna**[b]
Maurizio Vivarelli[c]

a) University La Sapienza, Roma, Italy, http://orcid.org/0000-0003-2999-1883 b) Tykli srl, Italy
c) University of Turin, Torino, Italy, http://orcid.org/0000-0002-9328-094X

Contact: Chiara Faggiolani, chiara.faggiolani@uniroma1.it.
Received: 03 June 2017; **Accepted:** 21 July 2017; **First Published:** 15 September 2017

ABSTRACT

This paper proposes some reflections concerning the practice of reading, its conceptual structure and its transformations, the blurred profile of the information ecology in which it is inserted. At the same time illustrates some outcomes of a research project conducted with tools of text mining and network science on the social reading platform aNobii. The paper presents these main topics: a) general overview of reading's context; b) short discussion reading as a complex system; c) presentation of some central concepts of network science and of its applications; d) introduction to text mining with some results of analysis of aNobii's reviews; e) conclusions and prospectives.

KEYWORDS

Reading; Social Reading; Complex System; Text mining; Network Science; aNobii.

CITATION

Faggiolani, C., L. Verna, and M. Vivarelli. "Text mining e network science per analizzare la complessità della lettura. Principi, metodi, esperienze di applicazione". *JLIS.it* 8, 3 (September 2017): 115-136. doi: 10.4403/jlis.it-12414.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Elementi di scenario

La lettura è un oggetto di studio di straordinaria complessità.[1]

Ciò dipende da molti fattori, dei quali uno dei più rilevanti è che i suoi elementi costitutivi (il testo, il libro, il lettore, la ricezione) sono parti fondamentali della tradizionale culturale dell'Occidente europeo. Come ha scritto Michel Foucault il libro solo in apparenza è un oggetto semplice, annidato nella sua rassicurante e familiare forma. In realtà ciò che viene denotato e connotato dalla parola 'libro' si situa in campi del discorso articolati e variegati, radicati nelle diverse contingenze storiche, e tuttavia tenuti insieme da alcuni invarianti metastoriche, riconducibili alle attività ed alle azioni nucleari che alla lettura possono essere riferite. Il libro, scrive Foucault, già a partire dalla sua "individualizzazione materiale", solo in apparenza definisce "i limiti del suo inizio e della sua fine": basta interrogarlo in modo più fine ed analitico e quei limiti divengono indefiniti. Per questo "i confini di un libro non sono mai netti né rigorosamente delimitati: al di là del titolo, delle prime righe e del punto finale, al di là della sua configurazione interna e della forma che lo rende autonomo, esso si trova preso in una rete di rimandi ad altri libri, ad altri testi, ad altre frasi: il nodo di un reticolo". La rassicurante omogeneità della forma del libro si rivela dunque "relativa e variabile", e per questo "è inutile che il libro sia dia come oggetto che si ha sotto mano; e inutile che si rannicchi in quel piccolo parallelepipedo che lo racchiude"; questa unità, auspicata e desiderata, si manifesta invece nella forma della scheggia, del puzzle e del frammento (Foucault 2009, 31-32).

Questa fondativa complessità, in una combinatoria che già mostra i tratti delle vertigini borgesiane, aumenta ancora se situiamo le pragmatiche del libro nelle indefinite e sterminate tracce delle pratiche di lettura osservate diacronicamente e sincronicamente, variabili a seconda delle persone, delle condizioni, dei luoghi e dei contesti, ed in cui si riflettono le relazioni estetiche e cognitive che erano state stabilite tra testi, libri, informazioni e lettori; e queste tracce possono qualificarsi come il fondamento di una prospettiva di indagine fenomenologica, centrata direttamente sull'esperienza dell'atto del leggere, e sulla perlustrazione del suo costituirsi e del suo manifestarsi. L'obiettivo è intravedere il profilo della e delle identità eteronome della lettura in quei fondali dei "territori dell'anima" da cui, come ha scritto Luca Ferrieri, traggono origine le tracce dei discorsi distribuiti nella storia collettiva ed individuale della lettura, che divengono i peculiari "documenti" su cui una storia ed una teoria della lettura possano fondarsi (Ferrieri 2013, 22 e *passim*).

Le linee generali della prospettiva di ricerca che in questo contributo viene proposta si pongono in primo luogo l'obiettivo di "salvare i fenomeni" della lettura, cioè di interpretarli, consapevoli che proprio nella delicata fragilità dei fenomeni, e dei documenti che recano le loro tracce, la lettura si situa, e solo lì può essere, appunto, rintracciata.[2] Tra questi fenomeni, un rilievo particolare va attribuito a

[1] Gli autori condividono i contenuti del contributo nel suo insieme. Si precisa che vanno attribuiti a Chiara Faggiolani i paragrafi *La centralità delle parole: introduzione al text mining per lo studio dei comportamenti di lettura* e *Conclusioni: riflessioni di carattere metodologico*; a Lorenzo Verna il paragrafo *Un approccio olistico all'analisi dei dati: introduzione alla network science*; a Maurizio Vivarelli i paragrafi *Elementi di scenario* e *La lettura come sistema complesso*. Data di ultima consultazione dei siti web: 3 giugno 2017.

[2] L'espressione "salvare i fenomeni" fa riferimento ad un noto problema epistemologico, che può essere ricondotto alla esigenza di definire 'modelli' della realtà, per l'approfondimento del quale si rimanda a Losee (2001). Su questi temi un interessante contributo sulle applicazioni del concetto di 'modello' nella Library Science è costituito da Salarelli (2009).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

quelli che danno conto della riconfigurazione in atto della lettura, nel suo passare dal circuito testo/libro/lettore gutenberghiano, tipico della modernità, a quello immerso nel paradigma digitale, ed il cui luogo dunque è uno spazio "altro", il web, i cui segni frusciano tra macchine e menti delle persone, nascosti nella loro invisibile forma digitale, in una rete di reti della lettura embricata nella ragnatela tecno-cognitiva del web; e in questi ambienti continuano a sedimentarsi dati, strutturati e non strutturati, originati dalle azioni dei lettori e dalla elaborazione autonoma delle macchine; dati di cui prendere atto a partire dalla loro immediata consistenza etimologica di ciò che si dà, ciò che si offre agli strumenti della nostra interpretazione. Dai dati, come si diceva, trae origine uno sterminato labirinto di tracce, che si configurano secondo la forma, metaforica e concettuale, della rete. Sul concetto di 'rete' insiste una letteratura smisurata, che trova una sua integrazione significativa in autori come Gregory Bateson, Fritjof Capra, Albert-László Barabási, Edgar Morin, (Bateson 1976; Barabási 2004; Capra 2006; Morin 2012); nel suo fondamento si colloca il pensiero sistemico, in cui l'attenzione si sposta "dalle parti al tutto", ed in cui "L'universo materiale è visto come una trama dinamica di eventi interdipendenti", in cui "la coerenza globale delle relazioni reciproche determina la struttura dell'intera trama" (Capra 2006, 51). Si tratta allora di produrre mappe di questa complessità, mettendo insieme le tessere di questo sterminato puzzle, condividendo l'affermazione di Barabási secondo cui "la rivoluzione in atto nel campo delle reti ci ha fornito le mappe più importanti", per cui è possibile immaginare e pensare, "continente dopo continente, la forma di un mondo nuovo" (Barabási 2004, 237).

La lettura come sistema complesso

Della lettura ci si può occupare a partire da molte prospettive disciplinari, ed i tentativi di definirne le proprietà, sotto il profilo storico ed epistemologico, dovranno continuare a muoversi in campi argomentativi strutturalmente aporetici, come già hanno scritto Roland Barthes ed Antoine Compagnon quando si chiedevano appunto quale fosse il punto di vista da adottare per un pratica designata da una parola con troppi usi (Barthes e Compagnon 1979, 176).

A conclusioni non dissimili, passando dall'ambito delle scienze umane a quello delle scienze sociali, si giunge anche prendendo in esame le rassegne sistematiche che dei dati e sulle statistiche sulla lettura ha approfonditamente elaborato e discusso, nel corso degli anni, Giovanni Solimine (Solimine 2010 e 2014).[3] Disponiamo di una grande quantità di dati ed informazioni sulla lettura, e tuttavia non solo rimane senza risposta la domanda che si ponevano Barthes e Compagnon ("che cos'è la lettura"), ma, su di un terreno più concretamente radicato nelle dinamiche politiche, culturali, sociali, rimangono inevase anche le molte domande che potremmo porci, ad esempio, su quali siano le strategie migliori per avviare progetti di promozione i cui esiti possano essere razionalmente valutati (Greenwood e Davies 2004).

Le attività di misurazione e valutazione più diffuse – e con vario grado di granularità – finalizzate all'ottenimento di informazioni a supporto di decisioni organizzative, possono avere per oggetto: la descrizione statistica del fenomeno, a livello nazionale o locale; l'andamento del mercato editoriale; l'efficacia dell'azione dei diversi soggetti organizzatori di attività di promozione; l'impatto economico,

[3] Non è possibile in questa sede dar conto della amplissima letterarura relativa alle statistiche internazionali sulla lettura. Per un primo orientamento si rimanda alla sezione *Literacy and Reading* del sito web dell'IFLA (https://www.ifla.org/taxonomy/term/496).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

o socio-economico, delle attività programmate; la misura del benessere e degli effetti cognitivi ed emotivi suscitati dall'esperienza della lettura; la valutazione dei livelli di apprendimento conseguiti.

A partire da questa varietà di argomentazioni possiamo senz'altro convenire, dunque, che il campo della lettura si può qualificare come un sistema complesso, caratterizzato da questi elementi (Preskill e Gopal 2014, 5):

- non è statico, e la sua evoluzione è imprevedibile;
- gli elementi del sistema sono connessi, e si influenzano vicendevolmente;
- i dati e le informazioni alimentano le funzionalità del sistema;
- il contesto influenza il funzionamento del sistema;
- le situazioni che si verificano sono spesso uniche e singolari. Per questo, più che a "buone pratiche", serve il ricorso a "buoni principi"[4];
- le relazioni tra le entità del sistema sono importanti quanto le entità stesse;
- le cause e gli effetti non sono sequenziali, ma reticolari;
- molte entità del sistema agiscono sulla base di fattori autonomi.

In questo crocevia, collocato nella intersezione tra tradizione interpretativa delle scienze umane e delle scienze sociali si collocano le prospettive di studio che in questo contributo sono descritte. I dati, analizzati in base a metriche del linguaggio utilizzato per realizzarli, e dei modelli astratti che ne delineano la configurazione, si qualificano sia come documenti per una storia della lettura prossima ventura sia come strumenti a supporto di decisioni, quali quelle che si definiscono in ambito editoriale, ad esempio, o nello sfumato panorama delle istituzioni pubbliche che di lettura e della sua promozione si occupano.

I metodi descritti nei paragrafi che seguono, radicandosi nella struttura etimologica della parola, vanno in cerca di un ὁδός, cioè di una strada da percorrere; e per orientarci abbiamo bisogno di tracce da seguire (Vocabolario online Treccani, 2017).

Un approccio olistico all'analisi dei dati: introduzione alla network science

aNobii[5] è una piattaforma software che realizza ed espone le proprie funzionalità ai suoi utenti finali. Le funzionalità che la caratterizzano sono la possibilità di aggiungere un libro alla propria libreria, valutarlo, commentarlo e partecipare alla rete sociale costituita dagli altri utenti. La realizzazione e la

[4] L'analisi delle "buone pratiche", o *best practices*, come è noto, è riconducibile al campo del *benchmarking*, una metodologia a matrice economica attraverso la quale le aziende, e più in generale le organizzazioni, valutano la propria attività comparandosi con quelle che ottengono le migliori prestazioni; per un inquadramento generale si veda Fiondella (2010). Il ricorso ai «buoni principi» è reso necessario dalla adozione del paradigma della complessità, entro il quale debbono essere valorizzate le specificità informative di ciascun sistema, nella loro peculiare dimensione reticolare e contestuale.

[5] Nata nel 2006 a Hong Kong da un'idea di Greg Sung, aNobii è un social network dedicato ai libri che vanta più di un milione di utenti nel mondo e che trova la sua base più consistente proprio in Italia con 300.000 lettori. Il Gruppo Mondadori nel 2014 ha acquisito da aNobii Ltd. il marchio e gli *asset* di aNobii (http://www.aNobii.com).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

gestione di queste funzionalità produce un interessante insieme di dati e metadati che nel loro insieme rappresentano le tracce che definiscono il comportamento dei lettori e dei libri all'interno della piattaforma.

Nel volume *Le reti della lettura* (Faggiolani e Vivarelli, 2016) abbiamo dato conto dell'applicazione di un approccio olistico all'analisi dei dati che ci ha consentito di prendere in considerazione tutti i dati e metadati, anonimizzati. L'obiettivo è la sperimentazione di tecniche di analisi che consentano di integrare i vari frammenti di informazione strutturata e non strutturata disponibile per derivare un modello di indagine che permetta di analizzare e comprendere un fenomeno così complesso.

Abbiamo ritenuto utile adottare il formalismo delle reti, ovvero della teoria dei grafi, come modello per rappresentare le informazioni disponibili, analizzarne le proprietà ed i fenomeni emergenti, ritenendo che il modello delle reti sia appropriato sia per le caratteristiche specifiche del dato da analizzare, sia per la flessibilità di adozione, sia per la capacità di descrivere sistemi complessi. Le reti basano le loro proprietà matematiche e formali sulla teoria dei grafi. Il primo testo che prende in considerazione i grafi come entità matematiche è *Mechanica, siue Motus scientia analytice exposita* (stampato a San Pietroburgo nel 1736) del matematico e fisico svizzero Leonhard Euler (1707-1783),[6] noto in Italia come Eulero, in cui la teoria dei grafi viene utilizzata per risolvere il problema dei sette ponti di Königsberg, che consisteva nell'individuare una soluzione che consentisse di attraversare tutti i ponti percorrendoli una volta soltanto.

I grafi sono oggetti discreti che permettono di schematizzare una grande varietà di fenomeni e di processi, e di consentirne l'analisi quantitativa e lo studio attraverso algoritmi. La definizione di grafo in matematica è molto semplice: un *grafo* G è dato dalla coppia ordinata di due insiemi:

V = insieme di tutti i nodi (o vertici)

E = insieme di tutti gli archi che uniscono coppie di vertici appartenenti a V

Due vertici *u* e *v* appartenenti a V connessi da un arco *e* prendono il nome di *estremi dell'arco*. L'arco *e* viene identificato con la coppia formata dai suoi estremi (*u*, *v*). Due nodi *u* e *v* si definiscono *adiacenti* se e solo se esiste nell'insieme E l'arco {*u*,*v*}. In quel caso i due nodi *u* e *v* si chiamano *vicini*. L'ordine di un grafo G è dato dal numero dei suoi vertici |V| mentre la dimensione è data dal numero dei suoi archi |E|.

Da queste semplici definizioni derivano molte proprietà che forniscono gli strumenti per conoscere un grafo, ed analizzare il fenomeno che esso descrive. Ad esempio, il numero di archi incidenti in un vertice *v* definisce il *degree* del vertice *v*. La misura del *degree* di ciascun nodo del grafo rappresenta uno strumento efficace per conoscere la struttura della rete osservata. In un grafo potremo avere ad esempio pochi nodi con *degree* molto alto e molti nodi con *degree* basso, o equivalente a 1. La distribuzione dei *degree* fornisce una prima indicazione di alcune caratteristiche strutturali del grafo.

[6] https://it.wikipedia.org/wiki/Leonhard_Euler.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

La teoria dei grafi definisce e indaga numerose altre proprietà quali ad esempio la densità, la completezza, la modularità, fornendo strumenti via via più complessi per descrivere il grafo e comprenderne le caratteristiche (Trudeau, 1993).[7]

Sulla base della teoria dei grafi, la recente disciplina della *network science* o *scienza delle reti*[8] studia le rappresentazioni a rete di fenomeni fisici, biologici e sociali (National Research Council 2005). Le reti sono uno strumento adatto a descrivere sistemi complessi derivando regole e proprietà che consentono di modellare le proprietà di fenomeni e strutture in cui intervengono numerosi elementi che agiscono, seguendo regole non coordinate centralmente (Caldarelli e Catanzaro 2007).

Lo scopo della scienza delle reti è rappresentare i fenomeni oggetto di indagine come insiemi di nodi e archi e quindi applicare le tecniche di analisi proprie di questo modello per interpretare le proprietà del fenomeno analizzato. Lo studio della *network science* è orientato ad individuare queste caratteristiche della rete analizzata:

- la dimensione della rete e la sua densità;
- la distribuzione dei *degree* dei nodi;
- la lunghezza media dei percorsi (la media della lunghezza di tutti i percorsi minimi tra ogni coppia di nodi nella rete) e il diametro della rete (il più lungo di tutti i percorsi minimi calcolati in una rete);
- il coefficiente di *clustering*, ovvero la misura della proprietà "gli amici dei miei amici sono miei amici";
- diverse misure di centralità dei nodi;
- la struttura delle community.[9]

Nel lavoro con i dati raccolti abbiamo utilizzato le reti per descrivere i comportamenti di lettura (Verna 2016; Faggiolani e Verna 2016). Non ci siamo occupati della rete sociale esplicitamente espressa dalle relazioni di "amicizia" tra i membri della piattaforma. Abbiamo utilizzato le reti per consentire a ogni frammento di informazione disponibile di relazionarsi agli altri in base a come è stato utilizzato.

Abbiamo quindi generato una prima rete omnicomprensiva – che abbiamo denominato *rete plain*, 'piatta', in cui non esiste ancora una gerarchia di relazioni – i cui nodi sono di diverso tipo: libro, autore, commentatore, commento, testo, concetto, parola, ecc. Questa prima rete rappresenta i frammenti e gli atomi del dato sorgente che si organizzano in un sistema complesso, i cui nodi sono costituiti dalle azioni che compiono gli utenti. Tutti i frammenti di informazione definiti dai dati che corrispondono a

[7] Appunti di teoria dei grafi, aggiornati al 7 agosto 2008, di Alberto Amato, Mario Gionfriddo, Giorgio Ragusa, Dipartimento di Matematica e Informatica, Università di Catania, online su http://www.spazioblog.it/uploads/g/giorgioragusa/212322.pdf.

[8] La scienza delle reti è un'area di ricerca che negli ultimi decenni ha attratto l'interesse interdisciplinare di fisici, matematici, informatici, economisti, sociologi, ecc. Una delle figure di riferimento è il fisico di origine ungherese Albert-László Barabási autore di *Network Science*, http://barabasi.com/networksciencebook/.

[9] Si intende la possibilità di raggruppare i nodi della rete in modo che siano densamente connessi al loro interno. Così la rete può essere naturalmente divisa in gruppi con connessioni molto dense all'interno del singolo gruppo e piuttosto rade verso gli altri gruppi (Porter-Onnela-Mucha 2009).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

ciascun commento formano una rete molto estesa che al crescere del numero di oggetti che la alimentano andrà ad assumere una propria struttura i cui nodi, ovvero gli elementi del sistema, avranno ruoli e dinamiche proprie. Applicando algoritmi di *network analysis* calcoliamo per ogni nodo 'libro' della *rete plain* il "peso" (importanza calcolata) delle relazioni verso ciascun altro nodo 'libro' presente sulla rete. Nella Fig. 1 vediamo una esemplificazione degli elementi che contribuiscono al calcolo della forza della relazione tra il nodo libro A e il nodo libro B.

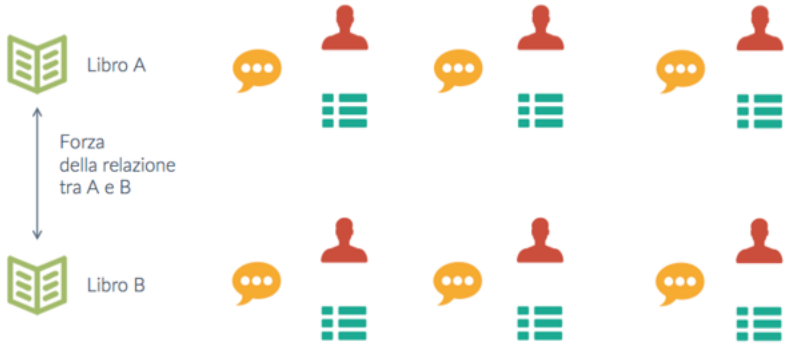

Figura 1. Visualizzazione delle relazioni tra libro A e libro B.

Ciascun nodo 'libro' sulla *rete plain* è collegato a utenti che lo hanno letto o che lo hanno commentato, a commenti, parole chiave estratte dalle recensioni, ecc. Il confronto tra due nodi 'libro' sulla *rete plain* porta a valutare come ciascun nodo si comporta sulla rete in termini di relazione pesata con gli altri nodi. Da ciò si può calcolare una misura della forza della relazione tra ciascuna coppia di nodi. Abbiamo quindi generato una nuova rete estraendo dalla *rete plain* i nodi di tipo 'libro' e le relazioni calcolate secondo le modalità appena descritte, configurando e visualizzando una rete di libri, le cui relazioni sono calcolate sulla base del come i libri sono commentati dagli utenti.

Ad esempio i primi 10 libri connessi a *Guida galattica per autostoppisti* di Douglas Adams sono in ordine di forza della relazione calcolata:

1. *Ristorante al termine dell'universo* (Douglas Adams)
2. *La vita, l'Universo e tutto quanto* (Douglas Adams)
3. *Addio, e grazie per tutto il pesce* (Douglas Adams)
4. *Il bar sotto il mare* (Stefano Benni)
5. *Fight Club* (Chuck Palahniuk)
6. *Se una notte d'inverno un viaggiatore* (Italo Calvino)
7. *Mattatoio n. 5* (Kurt Vonnegut)
8. *Il signore delle mosche* (William Golding)
9. *Survivor* (Chuck Palahniuk)
10. *Bar sport* (Stefano Benni)

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Considerando la nuova rete dove i nodi rappresentano i libri e gli archi pesati rappresentano la misura della relazione tra ciascuna coppia di libri, abbiamo applicato alcuni filtri per rimuovere i nodi e gli archi con scarsa significatività – ad esempio, libri con meno di 5 commenti e archi con peso inferiore a 0.01 – e abbiamo ottenuto una rete di 18.023 nodi (libri) e di circa 175 mila archi.

Sulla rete dei libri così generata possiamo applicare le diverse analisi classiche della *network science*. In particolare ci è parso interessante l'uso degli algoritmi di *community detection*, tecniche che cercano di individuare la "struttura a community" dei nodi di una rete, presente quando i nodi della rete possono essere raggruppati in insiemi tali che ciascun insieme sia densamente connesso al suo interno e con connessioni più rade verso gli altri gruppi. Abbiamo individuato 15 classi, ovvero comunità di libri densamente connesse tra loro. Le classi mostrano 15 insiemi di libri fortemente relazionati tra loro sulla base degli effetti delle azioni effettuate dagli utenti; azioni che, nella rete, si trasformano in relazioni.[10]

La rappresentazione della rete basata sulla struttura dei grafi è sufficiente ad applicare le tecniche di analisi e il calcolo delle proprietà della rete. Spesso gli archi sono rappresentati attraverso matrici, che sono a loro volta oggetti matematici di notevole interesse con le loro proprietà disciplinate dall'algebra lineare; rappresentare gli archi di un grafo come una matrice consente di sfruttare le proprietà matematiche delle matrici e semplificare alcuni calcoli.

Per una interpretazione umana delle reti né insiemi di coppie che identificano gli archi né gigantesche matrici sono di grande aiuto, e infatti molto spesso i grafi e le reti vengono visualizzati utilizzando tecniche di raffigurazione (Kaufmann e Wagner 2001; Di Battista-Eades-Tamassia -Tollis, 1999). La raffigurazione dei grafi è un'area di studio interdisciplinare che coinvolge la matematica, l'informatica, la teoria dei grafi, la geometria e le tecniche di visualizzazione delle informazioni.

Consideriamo il semplice grafo G dato dalla coppia di insiemi:

V={a,b,c,d}
E = {(a,b);(a,d);(b,d);(c,d)}

La rappresentazione a matrice del grafo G è rappresentata nella Fig. 2. La rappresentazione grafica più comune di una rete è il diagramma nodi-archi, dove ogni nodo è mostrato come un punto, un cerchio, un poligono o qualche altro oggetto grafico di piccole dimensioni, e ciascun arco come un segmento di linea o di curva che collega due nodi. Utilizzando questa tecnica otteniamo la visualizzazione del grafo G, come nella Fig. 3.

[10] In Faggiolani-Verna (2016) si possono trovare i riferimenti ai nodi-libro più rappresentativi per ciascuna classe.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

\	a	b	c	d
a	0	1	0	1
b	1	0	0	1
c	0	0	0	1
d	1	1	1	0

Figura 2. Rappresentazione a matrice del grafo *G*.

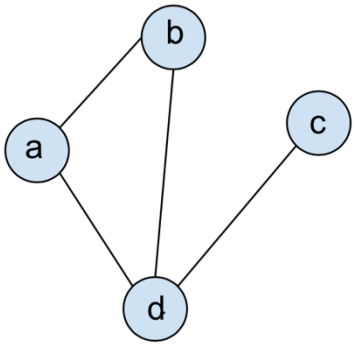

Figura 3. Una possibile raffigurazione del grafo *G*.

La stessa convenzione si può utilizzare anche per reti che modellano fenomeni fisici e sociali complessi, come ad esempio il comportamento dei lettori su aNobii. La visualizzazione di reti complesse deve affrontare problemi di geometria e topologia non banali. La rappresentazione della rete sul piano o su tre dimensioni è affidata ad algoritmi di layout che cercano di risolvere il problema di calcolare la posizione nello spazio per ciascun nodo e ciascun arco quando questi sono migliaia o milioni (Kaufmann e Wagner 2001; Di Battista, Eades, Tamassia e Tollis 1999).

Nell'affrontare i calcoli che consentono di trovare la posizione di ciascun nodo nello spazio di raffigurazione, le tecniche di *network visualization* devono soddisfare alcuni principi, tra i quali: rendere evidenti i nodi e gli archi più importanti; evidenziare le proprietà strutturali della rete; manifestare la forza delle relazioni; evidenziare i gruppi coesi e le communities (Ognyanova 2016).

Considerando le mappe, i fattori che determinano il risultato finale sono diversi. Tra i principali ricordiamo: il colore dei nodi e il colore degli archi; la posizione dei nodi nel piano; la dimensione dei cerchi che rappresentano i nodi o dello spessore degli archi; la forma dei cerchi e degli archi; le etichette che possono aggiungere contenuto informativo ai nodi e agli archi.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Le moderne tecniche di generazione dei layout di rete sono ottimizzate per la velocità di esecuzione e la gradevolezza del risultato, e in particolare cercano di soddisfare alcune regole estetiche: minimizzare gli incroci degli archi; uniformare la lunghezza degli archi e la simmetria; evitare la sovrapposizione dei nodi.

Gli algoritmi che generano i layout di rete cercano le disposizioni dei nodi ottimali per soddisfare questi principi. Il loro compito non è semplice quando il numero di nodi e archi cresce. Non è di fatto possibile trovare una soluzione che soddisfi tutti i criteri per ciascun nodo della rete da visualizzare. Una approccio molto diffuso per affrontare questo problema è basato su tecniche dette di *force-based layout* (McGuffin 2012; Herman, Melancon e Marshall 2000; Kobourov 2012). Si immaginano i nodi come delle particelle fisiche che sono inizializzate con una posizione casuale e che gradualmente si spostano per l'effetto di diverse forze fino a giungere alla posizione finale. Semplificando possiamo immaginare due forze: una forza repulsiva tra ciascuna coppia di nodi e una forza di attrazione tra le coppie di nodi *adiacenti*. Queste due forze contrapposte agiscono su ciascun nodo fino a un punto di equilibrio che rappresenta la configurazione finale.

Questa breve digressione ci fornisce alcuni elementi necessari per confrontarci con la lettura della rappresentazione visiva di una rete di migliaia di nodi. In particolare la posizione di un nodo sul piano ha sicuramente una valenza oggettiva ma non assoluta, in quanto è il risultato di uno o più calcoli che hanno come obiettivo la convergenza di diverse esigenze spesso antagoniste tra loro.

Nella fig. 4 possiamo vedere una possibile raffigurazione della rete dei libri di aNobii, generata utilizzando Gephi,[11] un software di visualizzazione free e open-source. La disposizione dei nodi sul piano è stata ottenuta a seguito di numerose iterazioni e configurazioni dell'algoritmo di network layout ForceAtlas2 (Jacomy, Venturini, Heymann e Bastian 2014).

La dimensione dei nodi è determinata in funzione dell'importanza del libro nella comunità di aNobii che il nodo rappresenta; in particolare la dimensione del nodo è in funzione del numero di commenti che il libro ha ricevuto. Il colore dei nodi identifica la community a cui il nodo appartiene. L'algoritmo di disposizione dei nodi nel piano ha rappresentato questo aspetto strutturale della rete, ed infatti troviamo con alto grado di probabilità i nodi appartenenti alla stessa community prossimi tra loro. Questa visualizzazione fornisce un buon colpo d'occhio e riflette in modo abbastanza coerente il comportamento dei libri nella rete. Troviamo nella nebulosa centrale i libri di maggiore rilevanza e trasversalità, mentre allontanandosi man mano dal centro si configurano gruppi più ristretti di libri con un minor numero di lettori ma legami più coesi.

[11] https://gephi.org/.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Figura 4. Una visualizzazione della rete dei libri di aNobii.

Sono possibili molte altre rappresentazioni della rete, ciascuna in grado di cogliere certi aspetti o certi criteri geometrici. È bene considerare che la visualizzazione di oggetti complessi come le reti con decine di migliaia di nodi non rappresenta l'esito finale dello studio e delle analisi delle reti stesse. Piuttosto può fornire una forma d'insieme, un'immagine macroscopica che sintetizza il contributo di ciascun elemento permettendo all'osservatore di cogliere alcune prime proprietà strutturali. L'interazione e l'esplorazione dei nodi e delle relazioni avvengono attraverso il supporto di strumenti diversi, che consentono di interrogare la rete e di restituire informazioni analitiche.

La centralità delle parole: introduzione al *text mining* per lo studio dei comportamenti di lettura

La percezione del lettore – la *significazione* – sfugge completamente agli strumenti tradizionalmente utilizzati per l'analisi dei comportamenti di lettura (Faggiolani 2016). Questo è un dato di fatto ed è il più grande limite delle indagini attualmente disponibili. Non si tratta di un limite metodologico: le indagini statistiche – ad esempio quelle di Istat che ogni anno ci fornisce una fotografia dell'andamento della lettura di libri in Italia[12] – non vogliono indagare i *significati* ma i *comportamenti* producendo una *generalizzazione*, non un *approfondimento*. I dati sui comportamenti di lettura, in quest'ottica, sono *modalità* (solitamente espresse in numeri) di *variabili* e l'esercizio delle indagini è individuarne la relazione, ovvero cogliere l'influenza di certe variabili su altre. È grazie a queste che sappiamo che circa 4 milioni di persone hanno smesso di leggere libri negli ultimi sei anni. Se vogliamo capire il perché di

[12] L'indagine annuale *Aspetti della vita quotidiana* rileva dal 1993 annualmente il numero di lettori di libri e la loro profilazione rispetto alle classiche variabili socio-demografiche. Per una panoramica sull'evoluzione delle indagini Istat dalle prime indagini dedicate alla lettura di libri del 1965 e del 1973 si rimanda a Savioli (2009). Per una riflessione articolata sulla lettura in Italia si veda Solimine (2010).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

questa perdita, o di questo travaso – immaginiamo che abbiano cessato di leggere per fare altro – se vogliamo individuare sistemi per promuovere la lettura, è necessario entrare nel contesto della lettura e nei suoi significati.

La *significazione* – il significato attribuito alle cose – passa per le parole (Foucault, 2016). Nel caso dello studio dei comportamenti di lettura, è evidente che le parole assumono un ruolo ancora più incisivo perché esse sono l'oggetto di quel comportamento che intendiamo studiare che si manifesta attraverso l'incontro del lettore con il testo scritto. La parola scritta, il cuore del lavoro editoriale, è al centro della comunicazione sui social media, è la materia degli *users-generated content* (UGC), ovvero i contenuti prodotti dagli utenti del/nel web, per esempio le conversazioni di cui rimane una traccia (digitale) nei social media generalisti e – aspetto che ci interessa particolarmente in questa sede – nelle piattaforme di social reading[13] come aNobii. Spazi in cui la lettura di libri oggi si sta profondamente riconfigurando.

Abituati a essere immersi nella comunicazione multimediale tendiamo a non dare la dovuta importanza al fatto che il web (sociale) è prima di tutto un mondo fatto di parole (Giuliano, 2013), e che queste oggi rappresentano la materia di cui è fatta una parte consistente dei cosiddetti *big data* che rappresentano una opportunità straordinaria di conoscenza e al contempo una sfida complessa per la metodologia della ricerca nelle scienze sociali.

È proprio per questa centralità delle parole che prendiamo in esame il *text mining*, inteso come processo di estrazione di informazione e conoscenza interessante da testi non strutturati. Si tratta di un complesso ambito di studi al quale sono ascrivibili le tecniche di estrazione delle informazioni da materiali espressi in linguaggio naturale – *Information Retrieval* (IR) e *Information Extraction* (IE) – utili per avere accesso alla conoscenza nascosta dentro le tracce digitali lasciate dagli utenti, per estrarre e visualizzare informazioni rilevanti.

Il trattamento automatico dei testi secondo un approccio di tipo metrico (analisi automatica del testo - AAT), effettuata attraverso software dedicati,[14] con l'obiettivo di rappresentare il contenuto dei testi oggetto di analisi e di estrarre informazioni di interesse attraverso misure quantitative, è l'approccio necessario quando si ha a disposizione una imponente molte di dati testuali per i quali non è possibile applicare analisi del contenuto di tipo interpretativo.

Le nozioni fondamentali per impadronirsi di questo approccio e il suo lessico specifico necessitano di una trattazione ben più ampia di quello possibile in questo spazio e peraltro già ampiamente presente in letteratura[15]. Qui il tentativo è quello di riprendere alcuni aspetti essenziali di questo approccio, secondo la particolare prospettiva dell'obiettivo che ha guidato la nostra ricerca sui comportamenti di lettura. Per fare questo è utile descrivere brevemente il set di dati con il quale ci siamo confrontati.

[13] Per una tassonomia si veda la proposta di Bob Stein dell'Institute for the Future of the Book http://futureofthebook.org/socialreading/.

[14] Tra i software di maggior rilievo possiamo segnalare TaLTaC2, Alceste, T-LAB, IRaMuTeQ, Lexico3: Cfr. Giuliano (2013). Le esemplificazioni che seguono sono frutto di elaborazioni condotte con IRaMuTeQ (http://www.iramuteq.org).

[15] Qui facciamo riferimento alle linee tracciate da Sergio Bolasco, uno dei primi studiosi italiani che si è occupato di analisi statistica di dati testuali (Bolasco 2013). Si fa riferimento anche a Lebart-Salem (1988; 1994).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Al momento dell'estrazione (giugno 2016) aNobii era frequentata da 1.202.909 utenti, di cui 353.663 italiani, con 8.020.066 di libri presenti, dei quali 1.203.007 nella nostra lingua.

Se applicassimo le tradizionali analisi statistiche considerando gli utenti di aNobii come collettivo statistico, potremmo ad esempio rilevare che le donne sono più numerose degli uomini e la fascia d'età più presente è quella 25-34 anni (Faggiolani e Verna 2016, 237). Nella logica del *text mining* non è questo il dato che ci interessa, se non come meta-dato[16], e neanche le relazioni tra gli attori della rete, di cui si è detto nel paragrafo precedente: sono le 2.552.955 recensioni, di cui 1.740.394 in italiano, per un totale di 80 milioni di parole circa a costituire l'oggetto d'analisi.

Associando le entità fondamentali di questo approccio a quelle statistiche tradizionali (Bolasco 2013, 206) osserviamo che l'insieme di tutte le recensioni è assimilabile al concetto di collettivo statistico. Queste costituiscono il *corpus* di dati, da intendersi come una raccolta di testi omogenea sotto qualche punto di vista: nel nostro caso, ad esempio, la condizione di enunciazione all'interno di aNobii. La parola che chiameremo 'forma grafica'[17] – ovvero una sequenza di caratteri delimitata da due separatori – è l'unita elementare del testo (*type*) e può essere considerata l'unità statistica sulla quale vengono operate le analisi. Il numero di volte in cui il *type* appare nel *corpus* determina le sue *occorrenze* (*tokens*)[18].

Per intraprendere l'analisi, il primo passo è quello di esaminare il *corpus* osservando la relazione tra *type* e *token*: si definiscono cioè le occorrenze di tutte le forme grafiche nel *corpus*. La distribuzione statistica è rintracciabile nella lista di frequenza dei *type*. Questo sarà il *vocabolario* del *corpus* e la sua ricchezza è data dal rapporto tra numero di *tokens* e numero di *types*.[19]

Immaginiamo ad esempio di voler conoscere quali sono le parole più utilizzate dagli utenti della piattaforma aNobii nel parlare dei libri letti e di osservare se esiste una differenza rispetto alla variabile sesso.[20] Creiamo così due diversi *corpora* testuali descritti di seguito dagli abstract, a partire dall'estrazione di un campione casuale di recensioni (il 10% del totale).[21]

[16] Nella logica dell'AAT i meta-dati rappresentano "quelle *informazioni* che in forma di annotazioni di vario genere arricchiscono i dati testuali e consentono una loro gestione in processi di estrazione di informazione (*data mining* e *text mining*) […] i meta-dati sono oggetti virtuali e "stratificabili", ossia sfruttabili nel trattamento automatico del testo attraverso "chiamata" del corrispondente strato (inteso come livello di analisi)" (Bolasco 2013, 83) [corsivo nel testo].

[17] "L'approccio all'analisi testuale sulla base della forma grafica delle parole introduce un elemento di novità sostanziale perché si tratta di un approccio formale indipendente dalla lingua, in cui l'analisi (automatica) è basata sul significante (la forma scritta della parola) e l'obiettivo è di arrivare al senso degli enunciati concreti, cioè a un insieme di significati esplicitati dal contesto". (Giuliano e La Rocca 2008, 154).

[18] I testi vengono sottoposti normalmente ad una serie di operazioni preliminari, sulla base degli obiettivi che guidano la ricerca: a) *normalizzazione*: standardizzazione del testo operata sulle parole, sulle frasi ecc.; b) *POS-tagging:* attribuzione di ogni forma alla sua categoria grammaticale; c) *lemmatizzazione*: trasformazione della forma nel lemma corrispondente. Il *lemma* è la forma corrispondente all'entrata del termine nel dizionario e rappresenta tutte le flessioni con cui quell'unità lessicale può presentarsi. Ad esempio, le occorrenze <leggerai> e <leggevo> sono due forme grafiche distinte appartenenti allo stesso *lemma*: la forma verbale <leggere>.

[19] È importante ricordare che generalmente il numero di *tokens* è maggiore del numero di *types*. Tuttavia in casi eccezionali possono essere uguali.

[20] Questa domanda di ricerca è del tutto esemplificativa. Essa si inserisce in una riflessione sulla specificità del linguaggio maschile e femminile rispetto alla lettura di libri di cui si è dato conto in Bandera, Caruso, Faggiolani e Ricci (2016).

[21] La scelta di lavorare su un campione deriva dalla necessità di rendere più agevole l'analisi dei dati.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Abstract Corpus Recensioni Uomini (campione 10%)

- Numero di testi (recensioni): 36.372
- Numero di occorrenze (*token*): 3.630.529
- Numero di forme grafiche (*type*): 84.082
- Numero di forme grafiche che occorrono una sola volta (*hapax*): 42.307

Abstract Corpus Recensioni Donne (campione 10%)

- Numero di testi (recensioni): 52.319
- Numero di occorrenze (*token*): 4.761.334
- Numero di forme grafiche (*type*): 85.515
- Numero di forme grafiche che occorrono una sola volta (*hapax*): 43.096

Ogni analisi basata su criteri statistici assegna alla frequenza delle parole un ruolo estremamente importante anche se non sempre questo costituisce un criterio decisivo di estrazione di conoscenza. Anche le parole incontrate poche volte o una sola (*hapax*) – perfino le parole assenti, talvolta – possono avere un valore rilevante.

Preliminare all'analisi del vocabolario è la distinzione tra *parole piene* e *parole vuote*: le prime fanno riferimento ai termini che hanno un senso in sé e comprendono le forme verbali, i nomi, gli aggettivi e gli avverbi; le seconde sono tra le parole più frequenti (secondo la legge di Zipf) e non sono portatrici di significato autonomo. Tra queste gli articoli, le parole finalizzate a funzioni grammaticali ecc.

A partire dai vocabolari dei due *corpora*, osserviamo nella Tab. 1 le prime 20 parole piene in ordine decrescente (per numero di occorrenze): alcune sono scontate e presenti in entrambe le tabelle: 'libro', 'leggere', 'romanzo' ecc. Possiamo considerare queste forme "parole tema"; altre si distinguono per essere presenti in una sola delle due; alcune, infine, presentano caratteristiche sulle quali varrebbe sicuramente la pena aprire una riflessione: per esempio la presenza del modale 'dovere'.[22]

[22] Si rimanda alla riflessione sul lemma 'dovere' nella percezione di *Gomorrra* di Roberto Saviano in Brugnatelli-Faggiolani (2016).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

	Donne		Uomini	
	Type	Token	Type	Token
1	ESSERE	125.960	ESSERE	86.994
2	LIBRO	34.919	LIBRO	23.823
3	FARE	29.761	FARE	20.330
4	STORIA	19.133	POTERE	12.680
5	POTERE	18.087	STORIA	12.321
6	LEGGERE	13.198	PRIMO	10.212
7	ROMANZO	11.681	ROMANZO	9.540
8	VITA	11.629	LEGGERE	7.903
9	PERSONAGGIO	10.857	VITA	7.339
10	VOLERE	10.355	PERSONAGGIO	7.061
11	PROPRIO	9.521	ANNO	6.817
12	TROVARE	9.307	AUTORE	6.649
13	DOVERE	8.918	PROPRIO	6.278
14	BELLO	8.684	SCRIVERE	6.275
15	PIACERE	8.642	DOVERE	5.827
16	ANNO	8.279	TROVARE	5.566
17	SCRIVERE	7.767	PAGINA	5.507
18	AUTORE	7.503	MONDO	5.434
19	PROTAGONISTA	7.486	RACCONTO	5.377
20	RIUSCIRE	7.423	LETTURA	5.233

Tabella 1. Le prime 20 forme per numero di occorrenze. Recensioni di uomini e donne a confronto.

Dopo una prima esplorazione del vocabolario, i *corpora* possono essere interrogati rispetto a diverse domande di ricerca:

- individuare le parole tema attraverso la visualizzazione rapida e intuitiva del contenuto delle nuvole di parole (*word cloud*);[23]
- visualizzare l'interconnessione delle parole interessanti con le altre parole maggiormente co-occorrenti;
- analizzare le concordanze, ovvero la presentazione delle parole nel testo e il contesto linguistico precedente e successivo (*cotesto*);
- misurare la specificità delle parole di uomini e donne a confronto (creando un unico *corpus*): dunque, non considerare quelle più ricorrenti ma quelle più specifiche in relazione alla variabile sesso rispetto ad una media attesa. Ad esempio nel nostro caso è interessante notare come le forme 'fantascienza', 'fumetto', 'saggio', 'economia', 'militare', 'sociale' siano molto specifiche nel linguaggio maschile mentre 'amore', 'storia', 'romance', 'sentimento', 'emozione', commuovere' lo siano per il linguaggio femminile, confermando uno stereotipo abbastanza diffuso.

[23] Queste ci aiutano a riconoscere le "parole tema", ovvero quelle scritte con una font più grande.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Molto interessante rispetto all'obiettivo di estrarre conoscenza da importanti moli di dati è la possibilità di segmentare i testi, ovvero di individuare cluster di testi caratterizzati da una forte omogeneità interna, rispetto alle parole e ai frammenti di testo. Specularmente a quanto osservato nel paragrafo precedente in cui le metriche di rete hanno evidenziato delle communities di libri, è possibile evidenziare all'interno del testo communities di contenuti. Si delineano così i "mondi lessicali" soggiacenti, ovvero classi lessicali in cui ricorrono, con maggiore frequenza, alcune espressioni che sono, quindi, individuate come tipiche delle porzioni di testo stesse[24]. I dendrogrammi riportati nelle Figg. 5 e 6 rappresentano i cluster estratti dai *corpora* delle recensioni delle donne e degli uomini a confronto.[25] I *cluster* sono contraddistinti ciascuno da un colore diverso, che ne permette l'individuazione anche nel grafico dei piani fattoriali (Figg. 7 e 8). Nel leggere i dendrogrammi è importante ricordare che essi permettono di ricavare l'organizzazione delle classi ottenute in termini di affinità di contenuto e che le ramificazioni ci dicono quanto e come i cluster sono connessi tra loro. Più le scomposizioni riguardano la fine dei rami più le classi si somigliano tra loro. Come si può notare osservando i grafici, il *corpus* "donne" è coperto da 3 classi – la classe 1 copre il 48,8%; la classe 2 il 26,6%; la classe 3 il 25% – per gli uomini le classi sono ben 5, che coprono dal 11,7% (classe 3) al 26,1% (classe 5).

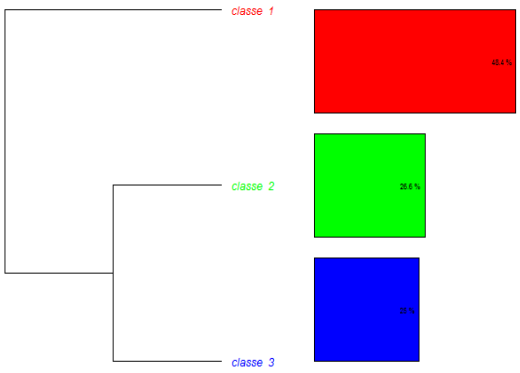

Figura 5. Clusterizzazione del *corpus* "donne".

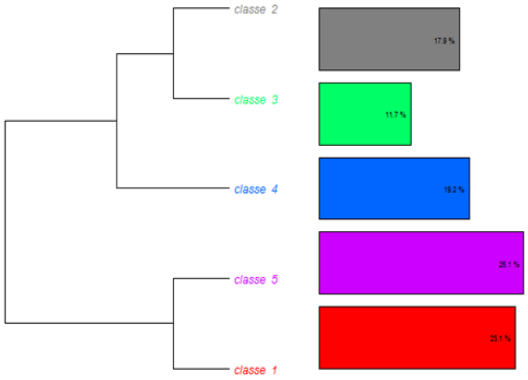

Figura 6. Clusterizzazione del *corpus* "uomini".

[24] "Noi chiamiamo "mondi lessicali" le impronte lessicali di questi luoghi nell'enunciazione, mondi che sono visualizzati tecnicamente, dal vocabolario specifico delle classi" (Reinert 1998, 292).

[25] A questo scopo IRaMuTeQ utilizza il metodo ALCESTE– *Analyse des Lexemes Cooccurrents dans les Enoncé s Simplifié s d'uni Texte* – che si basa sulla logica della ricerca delle similitudini, rintracciando nel testo la presenza co-occorrente delle stesse forme grafiche (parole o lessemi) (Reinert 1990).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

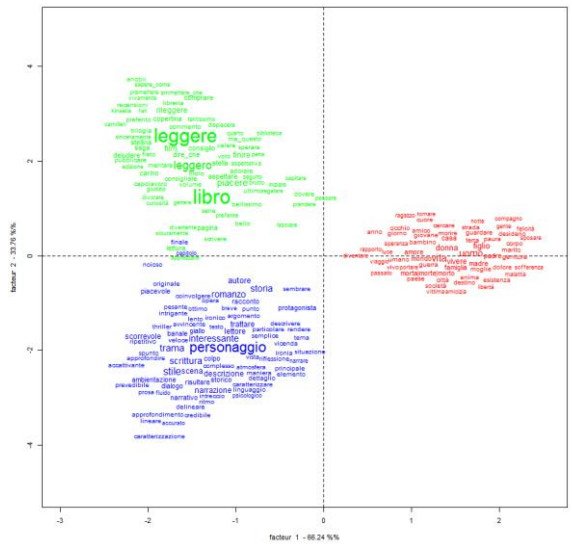

Figura 7. Piano fattoriale *corpus* "donne".

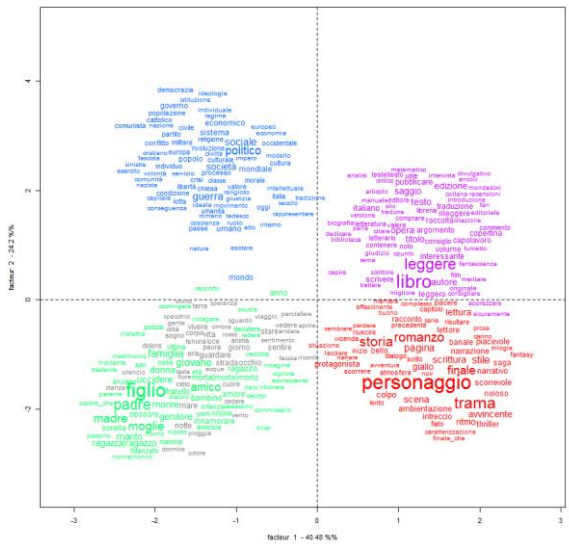

Figura 8. Piano fattoriale *corpus* "uomini".

La lettura dei profili ci permette di identificare le aree semantiche fondamentali. Di seguito si presentano sinteticamente le classi emerse per ciascun *corpus*.

CORPUS DONNE

Classe 1 (48,4%): area semantica con parole chiave: 'vita', 'uomo', 'donna', 'amore', 'dolore'.

Classe 2 (26,6%): area semantica con parole chiave 'libro', 'leggere', 'piacere', 'finire', 'copertina', 'consiglio'.

Classe 3 (25%): area semantica con parole chiave 'personaggio', 'stile', 'trama', 'interessante', 'romanzo'.

CORPUS UOMINI

Classe 1 (25,1%): area semantica con parole chiave 'personaggio', 'trama', 'storia', 'finale', 'scena', 'avvincente'.

Classe 2 (17,9%): area semantica con parole chiave 'vita', 'vivere', 'uomo', 'amore', 'notte'.

Classe 3 (11,7%): area semantica con parole chiave 'figlio', 'padre', 'madre', 'moglie', 'amico'.

Classe 4 (19,2%): area semantica con parole chiave 'politico', 'guerra', 'sociale', 'economico', 'governo'.

Classe 5 (26,14%): area semantica con parole chiave 'leggere', 'libro', 'piacere', 'capolavoro', 'film', 'copertina'.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Conclusioni: riflessioni di carattere metodologico

L'analisi della lettura è una attività complessa, perché la lettura è un sistema complesso.

Su questa assunzione si basano le attività e la vocazione inter-disciplinare del nostro gruppo di ricerca. L'esperienza che abbiamo maturato attraverso l'analisi dei dati estratti da aNobii ci ha consentito di mettere in evidenza empiricamente lo straordinario valore informativo dei dati che giacciono all'interno delle piattaforme di social reading, intesi come tracce delle azioni, delle scelte cognitive, della *significazione* dei lettori di libri. Tuttavia siamo consapevoli di essere solo agli inizi. Abbiamo cercato di descrivere, di circoscrivere, di mappare e di addomesticare un territorio ancora in larga misura inesplorato rispetto al quale è necessario porsi con molta cautela. Per questa ragione, in conclusione, abbiamo selezionato tra i tanti 4 aspetti che a nostro avviso meritano di essere oggetto di una riflessione condivisa, senza alcuna pretesa di esaustività ma con il solo obiettivo di delineare strade di lavoro possibili.[26]

1) La fonte dei dati

Le fonti di dati per lo studio dei comportamenti di lettura, in riferimento agli UGC, possono essere diverse e ciascuna presenterà diverse peculiarità: si pensi ad esempio alla differenza tra i social network generalisti (Facebook, Twitter, Instagram ecc.) e le piattaforme di social reading. Dal punto di vista dell'analista, c'è un elemento che accomuna queste fonti: a differenza dei dati raccolti attraverso indagini tradizionali (qualitative o quantitative), gli UGC non nascono con lo scopo di essere 'dati' ma vengono prodotti dagli utenti spontaneamente. Essi non derivano dalla definizione di alcun disegno di ricerca, non vengono raccolti dal ricercatore attraverso le consuete tecniche di raccolta dati della metodologia della ricerca sociale. I dati esistono e compito del ricercatore è "farli parlare". Questo aspetto costituisce un grande valore aggiunto, in termini di conoscenza induttiva ma al contempo si configura anche come un rischio, facilmente sintetizzabile attraverso la logica del *garbage in garbage out*. La quantità di dati a disposizione non implica necessariamente valore in termini di conoscenza estraibile e la pulizia dei dati (*data cleaning*) diventa un passaggio centrale,[27] così come una riflessione attentissima sul campionamento. I campioni in questo senso si auto-selezionano e il concetto di rappresentatività statistica e, dunque, di generalizzazione assume un significato completamente diverso.

2) La morfologia dei dati

La materia dei dati – non strutturati – è fluida e ambigua per definizione: le relazioni sono in divenire, il flusso di parole inarrestabile. Sono necessarie strutture cognitive ma anche infra-strutture per l'analisi completamente diverse che sempre più richiederanno la convergenza di competenze diverse e il confronto su un terreno caratterizzato dall'essere intrinsecamente trans-disciplinare.

[26] Sulle questioni metodologiche relative al paradigma *big data* si rimanda al numero dedicato di "Sociologia e Ricerca Sociale", *Sulle tracce dei big data. Questione di metodo e percorsi di ricerca*, 37, 2016, 109. I saggi sono il risultato della rielaborazione di relazioni svolte al convegno "Sulle tracce dei *big data*. Questioni di metodo e percorsi di ricerca", tenutosi a Roma il 26 settembre 2015 presso il Dipartimento di Comunicazione e ricerca sociale della Sapienza Università di Roma.

[27] La riflessione su questi temi è molto ampia e chiama in causa diversi aspetti: per esempio un tema centrale è la 'veridicità' dei contenuti: le persone manifestano la loro identità o costruiscono una identità ambita e desiderata?

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

3) La triangolazione metodologica

Le fonti di dati sopra descritti devono essere quanto più possibili integrate. La lettura come sistema complesso ha bisogno di questo. Ciò implica la necessità di applicare un approccio misto. Rimanendo dentro i confini di questo articolo, i due approcci sinteticamente descritti – la *network science* e il *text mining* – hanno messo in evidenza modalità di accostarsi all'analisi dei dati che, pur traendo origine da principi e metodi completamente diversi, risultano particolarmente efficaci in un'ottica di integrazione. Attenzione, dunque, a non considerarli alternativi, essi possono essere assolutamente complementari. Un esempio interessante è la costruzione dell'*ego-network*[28] di una parola (o forma grafica) che si ritiene particolarmente interessante per poterne analizzare le relazioni semantiche all'interno della rete.

4) Stimolare i dati

Se è vero che i dati esistono come effetto collaterale delle tracce lasciate dagli utenti in rete, è anche vero che essi producono conoscenza se sollecitati attraverso stimoli precisi: le domande di ricerca, per utilizzare un gergo tradizionale. Il valore dei dati in termini conoscitivi dipende, dunque, non soltanto dalla loro veridicità/pulizia (punto 1) ma anche dal progetto e dal processo di analisi cui vengono sottoposti.

Non solo la scelta delle domande, come ovvio, è discriminante rispetto all'avanzamento della conoscenza su certi temi e non su altri ma, determinando anche l'accesso a certi dati e non ad altri, le domande stesse permettono al ricercatore di confrontarsi con questioni importanti anche dal punto di vista metodologico. Per quanto riguarda la *network science*, per esempio, la scelta di lavorare sul *total network* o sull'*ego-network* determina il disegno di una rete dai confini completamente diversi.

La segmentazione del pubblico per gusti/comportamenti di lettura, la clusterizzazione dei libri secondo le reazioni che ne scaturiscono, la percezione della lettura in ambiente digitale, i mondi lessicali di uomini e donne che leggono sono soltanto alcuni dei temi che possono essere approfonditi. Solo dall'incrocio delle diverse conoscenze emerse e dall'applicazione di diverse metodologie potranno derivare significativi elementi novità che consentiranno di capire meglio il ruolo di alcune delle principali variabili connesse alla lettura come sistema complesso.

Bibliografia

Bandera, Stefano, Giovanni Caruso, Chiara Faggiolani, Andrea Ricci, 2016. Qualcosa di nuovo sulla lettura. Nuove prospettive di conoscenza con i big data, *Biblioteche oggi Trends*, 2, 1, 84-95.

Barabási, Albert-László. 2016. *Network Science*. Cambridge: Cambridge University Press. http://barabasi.com/book/network-science.

------. 2004. *Link. La nuova scienza delle reti*. Trad. di Benedetta Antonielli D'Oulx. Torino: Einaudi.

Barthes, Roland, e Antoine Compagnon. 1979. "Lettura", in *Enciclopedia*, vol. 8. Torino: Einaudi, 176-199.

Bateson, Gregory. 1976. *Verso un'ecologia della mente*. Trad. di Giuseppe Longo. Milano: Adelphi.

[28] Si intende lo studio della rete di un singolo attore: che sia un lettore, un libro, una parola. Si veda Mattioli, Anzera e Toschi (2014).

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Bolasco, Sergio. 2013. *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.

Brugnatelli, Edoardo e Chiara Faggiolani. 2016. "Gomorra: 10 anni di conversazioni su aNobii". In *Le reti della lettura. Tracce, modelli, pratiche del social reading*, a cura di Chiara Faggiolani e Maurizio Vivarelli, 261-303. Milano: Editrice Bibliografica.

Caldarelli Guido, Michele Catanzaro. 2007. *A Very Short Introduction to Networks*. Oxford: Oxford University Press.

Capra, Fritjof. 2006. *La rete della vita*. Trad. di C. Capararo. Milano: BUR.

Di Battista, Giuseppe, Peter Eades, Roberto Tamassia e Ioannis G. Tollis. 1999. *Graph Drawing: Algorithms for the Visualization of Graphs*. Upper Saddle River, NJ: Prentice-Hall.

Euler, Leonhard. 1736. *Mechanica, siue Motus scientia analytice exposita*. Petropoli: ex Typographia Academiae scientiarum, 1736.

Faggiolani, Chiara. 2016. "Morfologia dei dati sulla lettura (di libri)". In *I percorsi della conoscenza. Dialogando con Giovanni Solimine su biblioteche, lettura e società*, a cura di Giovanni Di Domenico, Giovanni Paoloni, Alberto Petrucciani, 169-183. Milano: Editrice Bibliografica.

Faggiolani, Chiara e Lorenzo Verna. 2016. "La lettura sul lettino: primi tentativi di data analysis". In *Le reti della lettura. Tracce, modelli, pratiche del social reading*, a cura di Chiara Faggiolani e Maurizio Vivarelli, 231-259. Milano: Editrice Bibliografica.

Faggiolani, Chiara e Maurizio Vivarelli (a cura di). 2016. *Le reti della lettura. Tracce, modelli, pratiche del social reading*. Milano: Editrice Bibliografica.

Ferrieri, Luca. 2013. *Fra l'ultimo libro letto e il primo nuovo da aprire: letture e passioni che abitiamo*. Firenze: Olschki.

Fiondella Clelia. 2010. *Il benchmarking dei processi aziendali: profili teorici ed operativi*, Torino, Giappichelli.

Foucault, Michel. 2009. *L'archeologia del sapere. Una metodologia per la storia della cultura*. Trad. di Giovanni Bogliolo. Milano: BUR.

------. 2016. *Le parole e le cose – un'archeologia delle scienze umane*. Trad. di Emilio Panaitescu. Milano: BUR.

Giuliano, Luca. 2013. *Il valore delle parole. L'analisi automatica dei testi in Web 2.0*. Roma: Dipartimento di Scienze statistiche.

Giuliano, Luca e Gevisa La Rocca. 2008. *L'analisi automatica e semi-automatica dei dati testuali*. Milano: Led.

Greenwood, Helen e James Eric Davies. 2004. "Designing tools to fill the void: a case study in developing evaluation for reading promotion projects", *Performance Measure and Metrics*, 5, 3, 106-111.

Herman, Ivan, Guy Melancon e Scott M. Marshall. 2000. "Graph Visualization and Navigation in Information Visualization: A Survey". *IEEE Transactions on Visualization and Computer Graphics* (TVCG), 6, 1, 24-43.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann e Mathieu Bastian. 2014. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software", *Plos One*, June 10, doi: 10.1371/journal.pone.0098679.

Kaufmann, Michael e Dorothea Wagner (a cura di). 2001. *Drawing Graphs: Methods and Models*. Berlin: Springer.

Kobourov, Stephen G. 2012. "Spring Embedders and Force Directed Graph Drawing Algorithms", https://arxiv.org/pdf/1201.3011.pdf.

Lebart, Ludovic e André Salem. 1988. *Analyse statistique de données textuelles*. Paris: Dunod.

------. 1994. *Statistique textuelle*. Paris: Dunod.

Losee, John. 2001. *Filosofia della scienza. Un'introduzione*. Trad. di Piero Budinich. Milano: Il saggiatore

Mattioli, Francesco, Giuseppe Anzera e Luca Toschi. 2014. *Teoria e ricerca nell'analisi delle reti sociali*. Roma: Euroma.

Mcguffin, Michael J. 2012. "Simple Algorithms for Network Visualization: A Tutorial", *Tsinghua Science and Technology*, 17, 4, p. 1-16.

Morin, Edgar. 2012. *La via. Un avvenire per l'umanità*. Trad. di S. Lazzari. Milano: Raffaello Cortina

National Research Council (U.S.). Committee On Network Science For Future Army Applications. 2005. *Network science* / Committee on Network Science for Future Army Applications, Board on Army Science and Technology, Division on Engineering and Physical Sciences, National Research Council of the National Academies. Washington, D.C. : National Academies Press.

Ognyanova, Katherine. 2016. *Network Visualization with* R, POLNET 2016 Workshop, St. Louis, MO, Rutgers University, http://kateto.net/polnet2016.

Porter, Mason A., Jukka Pekka Onnela e Peter J. Mucha . 2009. "Communities in Networks", *Notices of the AMS*, 56, 9, 1082–1097. http://www.ams.org/notices/200909/rtx090901082p.pdf.

Preskill, Hallie e Srik Gopal. 2014. "Evaluating Complexity. Propositions for Improving Practice", November 11, 2014, in *IssueLab*, http://www.fsg.org/publications/evaluating-complexity.

Reinert, Max. 1990. "ALCESTE: Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval," *Bulletin de méthodologie sociologique*, 26, 1, 24-54.

------. 1998. "Mondes lexicaus et topoi dans l'approche Alceste". In *Mots chiffrés et déchiffrés: mélanges offerts à Étienne Brunet*, textes rassemblés par Sylvie Mellet e Marc Vuillaume, 289-303. Paris: Honoré Champion.

Salarelli, Alberto. 2016. "Towards a Critique of the Concept of Model in Library Science". In *The Identity of the Contemporary Public Library. Principles and Methods of Analysis, Evaluation, Interpretation*, edited by Margarita Pérez Pulido and Maurizio Vivarelli, 153-168. Milano: Ledizioni.

Savioli, Miria. 2009. "Il lettore di libri questo (s)conosciuto", *Libri e riviste d'Italia*, 5, 7-31.

Solimine, Giovanni. 2014. *Senza sapere: il costo dell'ignoranza in Italia*. Roma-Bari: Laterza.

JLIS.it 8, 3 (September 2017)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12414

------. 2010. *L'Italia che legge*. Roma-Bari: Laterza.

Trudeau, Richard J. 1993. *Introduction to Graph Theory (Corrected, enlarged republication)*. New York: Dover Pub.

Verna, Lorenzo. 2016. "Prospettive di analisi dei dati". In *Le reti della lettura. Tracce, modelli, pratiche del social reading*, a cura di Chiara Faggiolani e Maurizio Vivarelli, 219-229. Milano: Editrice Bibliografica.

Vocabolario online Treccani. 2017. s.v. 'metodo', http://www.treccani.it/vocabolario/metodo/.