Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

Data Analytics
Sentiment Analysis and Network Analysis on Amazon

Cocca Umberto - 807191

Anno Accademico 2019 - 2020

Indice

Introduzione

Negli ulitmi anni si è visto un numero crescente di ricerche che hanno ampliato la comprensione del sentiment delle risorse testuali determinando l'avvento di servizi online che hanno cambiato il volto allo shopping.

Applicazioni di commercio online come Amazon concepiscono una quantità spropositata di dati per mezzo delle transizioni e degli utenti di questo servizio, infatti una parte consistente è data dai contenuti generati dagli utenti che valutano i prodotti acquistati e condividono la loro esperienza procedendo con valutazioni numeriche, seguite spesso da delle recensioni.

La Sentiment Analysis estrae dei dati strutturati da queste risorse testuali, permettendo un'analisi statistica sulle tendenze di comunità di acquirenti sotto diversi aspetti. Le aziende vogliono sempre trovare opinioni ed emozioni del pubblico o dei loro consumatori sui loro prodotti e servizi. Non solo, anche i potenziali clienti vogliono conoscere le opinioni e le emozioni degli utenti che hanno già usufruito di un certo servizio o acquistato un certo prodotto. Conoscere gli elementi più o meno apprezzati di un prodotto, secondo le diverse categorie di utenti, permette di condurre una migliore previsione del mercato e quindi attuare strategie aziendali favorevoli.

Le recensioni vengono recepite come fonti affidabili, rappresentando quindi uno strumento molto potente.

È possibile trovare il listato dei codici alla seguente **repository**.

Sentiment Analysis

Il Sentiment Analysis serve per interpretare il linguaggio naturale e identificare informazioni soggettive che denotano opinioni, emozioni e sentimenti, determinndo la polarità corrispondente (positiva, negativa o neutra) e comprendere il soggetto / oggetto target.

In questa fase, viene analizzata sistematicamente le parti testuali delle recensioni per estrarne un'opinione. Una parte preliminare pre-processing servirà per preparare il dataset. Vengono scartate recensioni troppo lunghe o troppe corte.

Infine, viene utilizzato ASUM (Aspet Sentiment Unification Model) per poter estrarre quelle che sono un insieme di topic che sono riferiti ai sentiment positivi e negativi. Usando ASUM si assume che il documento sia composto da frasi.

Dataset

Il dataset utilizzato è estratto da Amazon, in formato JSON, in cui sono presenti le recensioni rilasciate dagli utenti sul sito. Dal dataset si è specializzato per la categoria videogiochi.

Software utilizzati

Il flusso di lavoro si compone di due aree, una fase di manipolazione dei dati attraverso Python e una fase di processione dei dati attraverso il modello ASUM per il sentiment analysis.

Python

Per eseguire il preprocessing si è lavorato con Python, scelta dovuta alla grande quantità di strumenti e librerie open source disponibili per questo linguaggio. Le librerie utilizzate sono le seguenti:

- *Pandas*: per caricare e manipolare il dataset
- *NLTK*: per separare ogni review in una lista di frasi

- *re*: per eseguire una pulizia parziale sui dati, ad esempio eliminando quelle parole composte solo da numeri, o da caratteri inadeguati.

ASUM

Per mezzo di Python si è costruito l'input ad hoc per la versione Java di ASUM creata da Yohan Jo and Alice Oh, consultabile al seguente **link**.

L'input del programma è costituito da tre file, due obbligatori e uno opzionale:

- *BagOfSentences.txt* (obbligatorio)
  Questo file è una rappresentazione dell'elenco di parole dei documenti nel corpus. Per ogni documento, la prima riga è il numero di frasi. Dalla riga successiva viene visualizzato un elenco di indici che si riferiscono alla posizione relativa nel WordList;
- *WordList.txt* (obbligatorio)
  Questo file mappa le parole con indici di parole. Ogni parola è scritta in una riga. Si presume che la prima parola nel file abbia l'indice 0, la seconda parola abbia indice 1 e così via...;
- *SentiWords-0.txt, SentiWords-1.txt, ...* (opzionale)
  Questi file sono parole chiamati "semi sentimentali". Il numero del file dovrebbe iniziare da 0 e aumentare gradualmente. Nel modello ASUM è possibile aiutare il processo di campionamento facendo uso di queste informazioni a priori. Se sappiamo che una determinata parola è positiva perché appartiene al lessico dei positivi allora la sua probabilità di essere positiva la si conosce.

  Nello specifico per il progetto si sono usati due sentiment, uno positivo e uno negativo sfruttando due liste di parole italiane recuperate dal seguente **link** e manipolato leggermente aggiungendo alcune emoticon testuali.

Specializzazione del dataset

Vista la quantità considerabile dei dati per via del numero di utenti e del numero di prodotti (nell'ordine del milione) è complicato fare analisi esplorative approfondite su ogni singolo utente e su ogni singolo prodotto. Visto l'enorme quantità di dati si è selezionata una categoria ben specifica per il processamento dei dati: videogiochi.

Descrizione dataset

Il dataset è in formato JSON e viene caricato in memoria in formato DataFrame, con la libreria Pandas, in modo molto efficiente nonostante il dataset sia composto da circa due milioni di recensioni.

Infatti, il caricamento e il pre processamento del dataset sono le più impegnative computazionalmente, impiegando gran parte del tempo totale, ma comunque rimanendo in un tempo ragionevole. Durante lo sviluppo si è lavorato su un gruppo molto più piccolo di dati, per non essere troppo vincolati dal costo computazionale.

In generale, durante lo sviluppo si è lavorato con circa 100 reviews e per la fase di processo finale invece 10000 reviews.

```
REVIEWS:
_id: identificatore univoco della recensione

product: identificatore del prodotto a cui fa riferimento la recensione

title: titolo della recensione

author-id: identificatore univoco dell'autore della recensione

author-name: nome dell'autore della recensione

date: data in cui la recensione è stata scritta

rating: valutazione (in stelle, da 1 a 5) della recensione

helpful: numero di persone che hanno trovato utile la recensione

verified: indica se l'acquisto è stato verificato o meno

body: contenuto della recensione
```

Fig. 1: *reviews.json attributi*

Il dataset possiede gli attributi mostrati in tabella. Ogni record del dataset è la rappresentazione di una singola recensione svolta da parte di un utente per un certo prodotto nella data indicata. Per l'identificazione dell'utente abbiamo a disposizione il campo author-name e il campo author-id. Per quanto riguarda i campi relativi alla recensione, abbiamo a disposizione body per il contenuto della recensione. La quasi totalità dell'elaborazione dei dati avverrà da questo campo. Per identificare il prodotto abbiamo a disposizione solamente il campo asin, che è un codice univoco. Le recensioni sono classificate come verified se provengono da un acquisto su Amazon tale che l'autore della recensione ha effettivamente acquistato il prodotto senza ricevere sconti particolari. Sono state selezionate solo quelle verified.

```
PRODUCTS:
_id: identificatore univoco del prodotto

title: titolo del prodotto

category: categoria del prodotto

price: prezzo del prodotto

avg_rating: media dei voti delle recensioni (in stelle, da 1 a 5)

reviews_number: numero di recensioni

questions_number: numero di domande poste riguardo al prodotto

pictures: array di url delle immagini del prodotto

description: descrizione del prodotto

features: array di caratteristiche del prodotto

versions: array contente altre versioni dello stesso prodotto

bought_together: array di prodotti spesso acquistati insieme

also_bought: array di prodotti spesso comprati da gente che ha acquistato il prodotto

also_viewed: array di prodotti spesso visti da gente che ha visto il prodotto
```

Fig. 2: *products.json attributi*

E' stato eseguito il left join delle tabelle linkando quindi i prodotti alle rispettive reviews attraverso gli indici "_id" e "product".

Sentiment analysis

Il recupero di feedback dei clienti è un operazione estremamente importante per le aziende, soprattutto per le aziende leader che non badano a spese pagando miloni e milioni di euro per tale servizio. Per molto tempo si è fatto ciò manualmente (e tutt'ora per alcune aziende si procede in questo modo).

In questi anni però, vista l'ingente mole di informazioni, come ad esempio i feedback dei clienti, è fondamentale strutturare il raccoglimento dei dati e le loro analisi per il processo decisionale.

In particolare, per le opinioni su prodotti e servizi viene in aiuto la sentiment analysis, una discplina che può fornire risposte riguardo le questioni più importanti dal punto di vista dei clienti. Il processo di sentiment analysis permette, attraverso l'elaborazione del linguaggio naturale, di estrarre e analizzare in modo automatizzato opinioni soggettive espresse dall'utente, determinarne la polarità (positiva, neutrale, negativa) e, successivamente, riassumerle in maniera da poter essere di valore per l'azienda.

In questo modo, le decisioni possono essere prese sulla base di una quantità di dati significativa, piuttosto che da una semplice intuizione che non sempre si rivela corretta.

La sentiment analysis è importante perché le aziende vogliono che il loro marchio sia recepito positivamente. A tal proposito, ci si può concentrare su commenti positivi o negativi oltre che sul feedback del cliente, per valutare sia i punti di forza che i punti su cui migliorare.

Preprocessing

Prima di immergerci con l'aspect based sentiment analysis, è stato necessario una fase di preprocessing. Innanzitutto, si sono rimossi dal dataset i campi ritenuti superflui per l'analisi.

La manipolazione è avvenuta sequenzialmente e con step standard per analisi di questo tipo:

- *Divisione*: delle 10000 reviews si è andato a suddividere ciascuna in una lista di frasi attraverso la libreria di python NLTK;
- *Tokenization*: per ogni frase si è andati a tokenizzarle per creare una lista di parole;
- *Pulizia*: tutte le parole che fanno parte di punteggiatura, di simboli strani e di tutti quei elementi stopwords come articoli, congiunzioni che di fatto non aiutano a comprendere quello di cui si sta parlando;

  - Prima:
  ['Non', 'ho', 'mai', 'scritto', 'una', 'recensione', 'in', 'vita', 'mia', 'ma', 'per', 'questo', 'gioco', 'sento', 'proprio', 'il', 'bisogno', 'di', 'scriverci' , 'qualcosa', '.']

  - Dopo:
  ['mai', 'scritto', 'recensione', 'vita', 'gioco', 'sento', 'proprio', 'bisogno', 'scriverci', 'qualcosa']

- *Dizionario*: man mano che si scorrevano le parole si è costruito un dizionario necessario per poter eseguire ASUM;
- *Indicizzazione*: per ogni parola delle frasi si è dovuto mappare la sua corrispettiva parola nel dizionario per generare BagOfSentences.txt, obbligatorio per ASUM come input. Di seguito un esempio di costruzione documento:

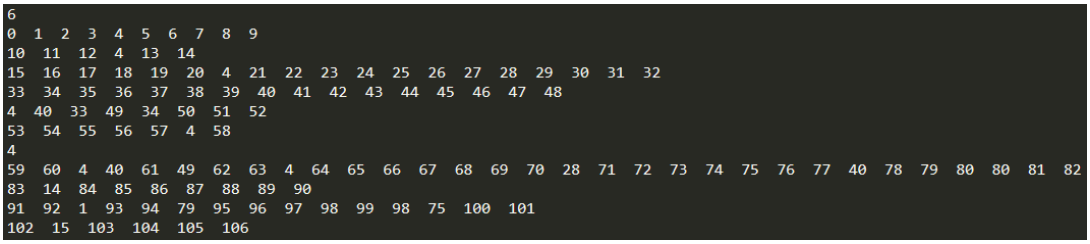

Fig. 3: *esempio input BagOfSentences.txt*

La prima riga indica il numero di frasi presenti nella review, di seguito infatti ci sono un numero di frasi in cui ogni indice si riferisce alla posizione della parola nella wordList;

- *Output*: infine si è creato l'output WordList.txt e BagOfSentences.txt per mezzo della funzione to_csv() della libreria Pandas.

Creazione di Bag Of Sentencess

Sono state rimosse dall'analisi le recensioni:

- *frasi lunghe*: con più di 300 parole;
- *frasi corte*: con meno di 5 parole .

Analisi dei risultati

L'output generato dal programma ASUM scritto in JAVA di Yohan Jo and Alice Oh restituisce diversi CSV, di cui a noi interessa quello relativo alle probabilità. Il nome dell'output dipende dalla stringa data in input. Infatti STO2-T[t]-S[s]([d])-A[a]-B[b]-G[g]-I[i]-[variable].csv dove:

- t: il numero di topics (aspetti);
- s: numero di sentimenti;
- d: numero di parole seed;
- a: valore alpha simmetrico;
- b: valore beta;
- g: valore gamma;
- i: numero di iterazioni di campionamento;
- *variable*: variabile da inferire;

	S0-T0	S0-T1	S0-T2	S0-T3	S0-T4	S1-T0	S1-T1	S1-T2	S1-T3	S1-T4
2	molto (0,016)	arrivato (0,037)	gioco (0,045)	gioco (0,038)	gioco (0,055)	gioco (0,037)	controller (0,016)	gioco (0,033)	gioco (0,010)	gioco (0,030)
3	prodotto (0,014)	gioco (0,036)	consiglio (0,016)	molto (0,015)	molto (0,038)	giocato (0,009)	ricarica (0,012)	italiano (0,025)	pedaliera (0,010)	molto (0,010)
4	ottimo (0,010)	prodotto (0,022)	saga (0,013)	grafica (0,012)	figlio (0,022)	tempo (0,009)	senza (0,011)	arrivato (0,019)	po (0,009)	giocare (0,008)
5	volante (0,009)	amazon (0,020)	giochi (0,013)	storia (0,008)	regalo (0,017)	molto (0,009)	gioco (0,010)	inglese (0,015)	volante (0,009)	giochi (0,006)
6	controller (0,008)	spedizione (0,020)	giocato (0,013)	giochi (0,006)	ottimo (0,015)	dire (0,008)	carica (0,009)	prodotto (0,012)	viti (0,008)	solo (0,006)
7	prezzo (0,008)	tempi (0,017)	serie (0,011)	sempre (0,005)	anni (0,015)	dopo (0,008)	due (0,009)	lingua (0,011)	troppo (0,008)	senza (0,006)
8	gioco (0,007)	prezzo (0,016)	genere (0,011)	ottimo (0,005)	prezzo (0,012)	l'ho (0,007)	problema (0,008)	solo (0,009)	molto (0,006)	po (0,005)
9	sempre (0,006)	ottimo (0,015)	molto (0,011)	giocare (0,005)	stato (0,011)	consiglio (0,007)	prodotto (0,008)	senza (0,008)	istruzioni (0,006)	fare (0,005)
10	base (0,006)	consegna (0,015)	mai (0,010)	bello (0,005)	bello (0,010)	mai (0,007)	sempre (0,008)	confezione (0,008)	fissaggio (0,006)	online (0,005)
11	ricarica (0,006)	sempre (0,014)	grafica (0,008)	davvero (0,004)	divertente (0,009)	po (0,006)	molto (0,008)	amazon (0,008)	forse (0,005)	storia (0,004)
12	bene (0,006)	perfetto (0,010)	ottimo (0,008)	titolo (0,004)	consiglio (0,009)	giochi (0,006)	joypad (0,008)	sottotitoli (0,008)	montaggio (0,005)	modalitÃ (0,004)
13	qualitÃ (0,006)	molto (0,010)	amanti (0,007)	personaggi (0,004)	acquistato (0,009)	altri (0,006)	base (0,008)	versione (0,007)	fori (0,005)	troppo (0,004)
14	solo (0,005)	condizioni (0,010)	gta (0,007)	solo (0,004)	prodotto (0,008)	senza (0,005)	quando (0,007)	copertina (0,007)	peccato (0,005)	consiglio (0,004)
15	due (0,005)	giorno (0,009)	bello (0,007)	fatto (0,004)	grafica (0,008)	sicuramente (0,005)	dopo (0,007)	italiana (0,006)	poco (0,005)	prima (0,004)
16	funziona (0,005)	perfettamente (0,008)	consigliato (0,007)	consiglio (0,004)	l'ho (0,008)	saga (0,005)	ore (0,006)	custodia (0,006)	pÃ² (0,005)	dopo (0,004)
17	batterie (0,005)	italiano (0,008)	bellissimo (0,007)	saga (0,004)	regalato (0,007)	poco (0,005)	batterie (0,006)	pacco (0,006)	solo (0,005)	missioni (0,004)
18	carica (0,004)	puntuale (0,007)	assolutamente (0,007)	online (0,004)	fatto (0,007)	ancora (0,004)	solo (0,006)	tempo (0,006)	quindi (0,005)	altri (0,004)
19	consiglio (0,004)	giorni (0,007)	giocare (0,007)	modalitÃ (0,004)	bambini (0,007)	aspettative (0,004)	fa (0,005)	giorno (0,006)	dovuto (0,004)	mai (0,003)
20	comodo (0,004)	confezione (0,007)	titolo (0,007)	divertente (0,004)	dire (0,007)	arrivato (0,004)	cavo (0,005)	stelle (0,006)	unica (0,004)	anni (0,003)
21	avere (0,004)	pacco (0,007)	capitolo (0,007)	giocato (0,004)	nipote (0,006)	prodotto (0,004)	po (0,005)	stato (0,005)	pecca (0,004)	tempo (0,003)
22	essere (0,004)	stato (0,007)	uncharted (0,007)	ben (0,004)	comprato (0,006)	anni (0,004)	led (0,005)	spedizione (0,005)	fissare (0,004)	giocato (0,003)
23	supporto (0,004)	anticipo (0,007)	dire (0,006)	versione (0,003)	natale (0,006)	titolo (0,004)	volta (0,004)	tempi (0,005)	piastra (0,004)	trama (0,003)
24	fa (0,004)	veloce (0,006)	fan (0,006)	prezzo (0,003)	arrivato (0,006)	sempre (0,004)	funziona (0,004)	problema (0,005)	vite (0,004)	prezzo (0,003)
25	pedaliera (0,004)	ben (0,006)	migliori (0,006)	veramente (0,003)	giocare (0,006)	figlio (0,004)	pad (0,004)	francese (0,005)	missioni (0,004)	puÃ² (0,003)
26	quando (0,003)	prima (0,006)	prezzo (0,006)	serie (0,003)	sempre (0,006)	bel (0,004)	caricare (0,004)	prima (0,004)	difetto (0,003)	cosa (0,003)
27	giochi (0,003)	descrizione (0,005)	sempre (0,006)	altri (0,003)	lego (0,006)	problemi (0,004)	poco (0,004)	giorni (0,004)	caso (0,003)	volta (0,003)
28	perfettamente (0,003)	perfette (0,005)	solo (0,005)	anni (0,003)	davvero (0,006)	acquistato (0,004)	problemi (0,004)	consegna (0,004)	tende (0,003)	quindi (0,003)
29	po (0,003)	nuovo (0,005)	veramente (0,005)	capitolo (0,003)	spedizione (0,005)	c'Ã¨ (0,004)	mai (0,004)	scritto (0,004)	pezzo (0,003)	l'ho (0,003)
30	joypad (0,003)	dopo (0,005)	storia (0,005)	prima (0,003)	bene (0,005)	trama (0,004)	ricaricare (0,004)	prezzo (0,004)	fare (0,003)	pecca (0,003)

Fig. 4: *output ASUM with -s 2 -t 5 -i 1000*

Di seguito vediamo l'output ottenuto con s = 2 e t = 5, cioè due sentimenti positivo e negativo e 5 aspetti di analisi. In ogni colonna visualizziamo le parole con una probabilità decrescente andando dall'alto verso il basso.

1	S0-T0	S0-T1	S0-T2	S0-T3	S0-T4
2	molto (0,016)	arrivato (0,037)	gioco (0,045)	gioco (0,038)	gioco (0,055)
3	prodotto (0,014)	gioco (0,036)	consiglio (0,016)	molto (0,015)	molto (0,038)
4	ottimo (0,010)	prodotto (0,022)	saga (0,013)	grafica (0,012)	figlio (0,022)
5	volante (0,009)	amazon (0,020)	giochi (0,013)	storia (0,008)	regalo (0,017)
6	controller (0,008)	spedizione (0,020)	giocato (0,013)	giochi (0,006)	ottimo (0,015)
7	prezzo (0,008)	tempi (0,017)	serie (0,011)	sempre (0,005)	anni (0,015)
8	gioco (0,007)	prezzo (0,016)	genere (0,011)	ottimo (0,005)	prezzo (0,012)
9	sempre (0,006)	ottimo (0,015)	molto (0,011)	giocare (0,005)	stato (0,011)
10	base (0,006)	consegna (0,015)	mai (0,010)	bello (0,005)	bello (0,010)

Fig. 5: *output ASUM s 0*

Focalizzandoci nell'area in cui abbiamo i valori positivi possiamo notare che compaiono parole che infatti hanno un valore positivo. Notiamo che dal momento che trattiamo dei videogiochi compare per 3 topic su 5 come prima parola "gioco" con una probabilità maggiore di 0.037.

S1-T0	S1-T1	S1-T2	S1-T3	S1-T4
gioco (0,037)	controller (0,016)	gioco (0,033)	gioco (0,010)	gioco (0,030)
giocato (0,009)	ricarica (0,012)	italiano (0,025)	pedaliera (0,010)	molto (0,010)
tempo (0,009)	senza (0,011)	arrivato (0,019)	po (0,009)	giocare (0,008)
molto (0,009)	gioco (0,010)	inglese (0,015)	volante (0,009)	giochi (0,006)
dire (0,008)	carica (0,009)	prodotto (0,012)	viti (0,008)	solo (0,006)
dopo (0,008)	due (0,009)	lingua (0,011)	troppo (0,008)	senza (0,006)
l'ho (0,007)	problema (0,008)	solo (0,009)	molto (0,006)	po (0,005)
consiglio (0,007)	prodotto (0,008)	senza (0,008)	istruzioni (0,006)	fare (0,005)
mai (0,007)	sempre (0,008)	confezione (0,008)	fissaggio (0,006)	online (0,005)

Fig. 6: *output ASUM s 1*

Focalizzandoci nell'area in cui abbiamo i valori negativi possiamo notare che compaiono parole che oscillano tra un valore positivo e valore negativo. Il problema principale deriva dalla lingua utilizzata. Notiamo che la parola gioco è predominante ancora con una percentuale mediamente più bassa. Notiamo che nonostante sia stato fatto con regex l'eliminazione di parole con l'apostrofo una è riuscita a passare lo stesso "l'ho", probabilmente un apostrofo con una formattazione diversa

Conclusioni

L'esplorazione delle recensioni di prodotti Amazon ha permesso di visualizzare un buon numero di informazioni che possono essere estratte da opinioni degli acquirenti con lo

scopo di stilare statistiche e valutazioni e poter quindi prendere decisioni in ambito aziendale per migliorare i servizi offerti o centrare meglio la propria clientela. Tali tecniche sono più affinate per il linguaggio inglese, pertanto bisogna intervenire con un miglior preprocessing e con un numero di SentiWords più idoneo e di qualità.