Introduzione

Viviamo in un mondo sempre più pieno di assistenti digitali, che ci permettono di connetterci con altre persone e che possiedono vaste risorse informative. Parte del fascino creato da questi dispositivi intelligenti risiede nel fatto che non solo sono capaci di fornire informazioni, ma spesso anche di interpretarle, creando una sorta di interazione tra l'utente e il sistema, molto vicina a quella umana; basti pensare ad applicazioni come la traduzione automatica, i sistemi di domande e risposte, la scrittura di messaggi tramite comando vocale o ancora i *chatbot*, che sono sempre più parte integrante del nostro quotidiano.
Grazie alle recenti scoperte nel settore del *machine learning* e dell'elaborazione dei dati, nell'ultima decade sono nate nuove aree di studio e nuove figure professionali. Una di queste è il *data scientist*, in parte statistico, in parte *data analyst*; questo tipo di lavoro rappresenta l'unione di più forze e menti, che convogliano le conoscenze che per tanti anni sono state di dominio quasi esclusivamente accademico, in un prodotto commerciale perfettamente spendibile e ormai divenuto anche indispensabile. Non a caso, il lavoro di *data scientist* è stato classificato come uno dei lavori "più sexy" del Ventunesimo secolo, ma allo stesso tempo anche uno dei meno compresi.
Cercando su Internet, troviamo definizioni come la seguente:

> In generale, un "data scientist" è qualcuno che è in grado di estrarre e interpretare dati, cosicché siano necessarie competenze di statistica e machine learning. Il suo lavoro è fatto prevalentemente di raccolta e pulizia dati, e solo in minor parte estrapolazione di nuove informazioni.

Una definizione come la precedente non lascia molto spazio all'interpretazione: un *data scientist* lavora con i dati, rendendoli comprensibili e fruibili per altri scopi. Tutto ciò sembra distante anni luce da quello che invece è il lavoro di uno sviluppatore, che ha a che fare continuamente con i dati; questo perché, nell'immaginario collettivo, un programmatore è una figura mitologica che passa il suo tempo a scrivere geroglifici su una schermata nera, cercando di costruire sistemi infernali… Ironia a parte, il programmatore

è una figura professionale che si occupa quasi nella sua totalità del tempo di analizzare e progettare sistemi software in grado di trovare una soluzione a uno o più problemi. Un *data scientist*, al contrario, manipola i dati, ma non è detto che debba saper programmare; difatti può lavorare esclusivamente con strumenti forniti da software di gestione dei database, senza scrivere neppure una riga di codice.
E qui potrebbe sorgere spontanea una domanda: se un programmatore scrive codice per poi produrre qualcosa che sia visibile a un utente (come un sito web oppure un programma gestionale), che cosa produce un *data scientist*? In un contesto come quello del marketing, la *data science* è in grado di produrre modelli di analisi di mercato utilizzati per prendere coscienza di processi di cambiamento all'interno della vendita o della fruizione di un certo servizio o prodotto. Alcune delle attività che svolge tipicamente un *data scientist*, oltre a "pulire" i dati, è anche l'interpretazione degli stessi; azioni come l'analisi dei dati per identificare schemi e tendenze o la creazione di prototipi che permettano a un'azienda, o più in generale a un'entità, di adottare soluzioni *data-driven*.
La comunicazione umana è uno degli attributi più affascinanti dell'essere senzienti. Come persone, sappiamo come l'interazione con altri individui possa essere estremamente complessa. Spesso inviamo e riceviamo messaggi sbagliati oppure i nostri messaggi vengono mal interpretati da altri; ogni giorno diamo per scontata la nostra capacità di trasmettere del significato ai nostri colleghi e familiari. Comunichiamo in una grande varietà di modi, tra cui la parola e i simboli scritti. La comunicazione umana può essere semplice, attraverso uno sguardo in una stanza o in mezzo a decine di persone; può essere però anche scritta tramite mezzi di comunicazione tecnologici, come ormai avviene da più di mezzo secolo.
Secondo molte statistiche di mercato, il volume dei dati raddoppia ogni due anni, ma in futuro questo lasso di tempo potrebbe ridursi ulteriormente. La gran parte di questi dati (circa il 79%) è di natura testuale; l'elaborazione del linguaggio naturale, Natural Language Processing (abbreviato in NLP), è un ramo secondario del settore della *data science* che tenta di estrarre informazioni dal "testo".
Il futuro dell'elaborazione del linguaggio naturale sta soprattutto in questo concetto: alle macchine viene insegnato a leggere e interpretare il testo così come fanno gli esseri umani, ma tramite l'elaborazione del linguaggio naturale è possibile anche "comprendere", oltre che analizzare, un testo o un intero discorso. Questo tipo di capacità umana è estremamente utile per analizzare grandi volumi di dati testuali.
Con la crescita esponenziale di dati provenienti da diversi canali, come quelli social o mobile, le aziende hanno bisogno di solide tecnologie per valutare il *sentiment* e le reazioni dei clienti. Finora, le aziende sono state propense ad analizzare le azioni dei clienti, ma nell'attuale clima competitivo, quel tipo di analisi dei clienti risulta obsoleto.
Ora le aziende devono analizzare e comprendere atteggiamenti, preferenze e persino stati d'animo dei clienti, che rientrano nell'ambito delle analisi del *sentiment*. Senza l'elaborazione del linguaggio naturale, i titolari delle aziende si troverebbero in estrema difficoltà nel condurre anche le più elementari analisi di mercato.
Se avete deciso di leggere questo libro, suppongo che vogliate comprendere il funzionamento di alcuni strumenti di comprensione del linguaggio naturale e di vederne l'utilizzo in applicazioni e software. I sistemi di comprensione del linguaggio sono costruiti su un complesso insieme di regole che governano l'analisi del testo, ovvero un insieme di tecniche e metodi che combinano manipolazione di parole, avverbi e quant'altro.

Piuttosto che essere definiti solamente da regole, i linguaggi naturali sono definiti dall'uso e devono essere decodificati per essere calcolati; sono infatti molto distanti dai linguaggi di programmazione o dai linguaggi formali, che sono invece basati su regole matematiche e logiche rigide e ben definite. Noi esseri umani, in quanto tali, siamo in grado di decidere che cosa significa una frase elaborata in una comunicazione scritta o orale, anche se capita spesso che si creino ambiguità a seconda della cultura e della tradizione linguistica degli interlocutori.

Questo libro si concentra sul tema dell'elaborazione del *linguaggio naturale*. Per linguaggio naturale intendiamo il linguaggio usato per comunicare abitualmente, che sia in lingua inglese, hindi o portoghese. A differenza dei linguaggi artificiali, come i linguaggi di programmazione e le notazioni matematiche, i linguaggi naturali si sono evoluti passando da una generazione all'altra e sono difficili da definire con regole esplicite. In questo senso, intendiamo quindi come linguaggio naturale qualsiasi linguaggio che ci permetta di manipolare – e in un certo senso anche comprendere – un dato quantitativo di informazioni tramite un sistema digitale.

Le tecnologie basate sull'elaborazione del linguaggio naturale sono sempre più diffuse. Per esempio, smartphone e computer supportano il riconoscimento del testo e della scrittura a mano; motori di ricerca come Google danno accesso a informazioni contenute anche in dati non strutturati, come immagini; i sistemi di traduzione automatica ci permettono di fotografare cartelloni aeroportuali in cinese e leggerli in italiano in pochi secondi.

Questo libro vuole quindi fornire un'introduzione altamente accessibile al campo dell'elaborazione del linguaggio naturale; può essere preso come riferimento per lo studio individuale o come libro di testo di integrazione per un corso di intelligenza artificiale o di *text mining* o anche di linguistica computazionale. Il libro è il più possibile pratico e cerca di andare "al sodo" di ogni concetto, dal momento che contiene molteplici esempi spiegati passo per passo, anche se l'obiettivo è stato quello di trovare un equilibrio tra teoria e applicazione. Il tutto è basato sul linguaggio di programmazione Python, insieme a diverse librerie open source, come NLTK o spaCy.

Lo studio dell'analisi del linguaggio naturale trova applicazione nelle discipline scientifiche, economiche, sociali e culturali. Al momento sta vivendo una rapida crescita, in quanto le sue teorie e i suoi metodi sono implementati in un'ampia varietà di nuove tecnologie linguistiche. Nel mondo dell'industria, le applicazioni includono analisi di informazioni commerciali e sviluppo di software web; all'interno del mondo accademico e scientifico includono settori dall'informatica umanistica alla linguistica dei *corpora*, fino ad arrivare all'informatica e all'intelligenza artificiale.

Non avete mai programmato?

I primi due capitoli del libro sono adatti ai lettori che non hanno una precedente conoscenza della programmazione e che non temono di affrontare nuovi concetti e sviluppare nuove capacità informatiche. Il libro è ricco di esempi che potete copiare e provare, insieme a decine di esercizi di difficoltà appropriata. Se avete bisogno di un'introduzione più generale a Python, troverete un valido elenco delle risorse Python qui di seguito.

Avete programmato, ma non conoscete Python?

Python è relativamente facile da imparare ed è considerato uno dei linguaggi migliori per iniziare ad approcciarsi alla programmazione, quindi non è necessario avere un'esperienza di programmazione precedente.
Ha un vocabolario molto semplice. Se lo conoscete almeno un po', sapete che alcune (quasi tutte) le regole sono molto rigide e se commettete anche il minimo errore, avrete molti problemi a eseguire il codice; d'altronde delle regole così solide rendono più semplice imparare l'intera sintassi. Python rende molte cose davvero facili, con una sintassi essenziale ma potente.
Un esempio? È un linguaggio tipizzato dinamicamente. Potete evitare il mal di testa da *typecasting* e dichiarare il tipo di variabile mentre dichiarate la variabile stessa.
Se però volete comprendere appieno i contenuti di questo libro, ci sono moltissime risorse online cui far riferimento. Ecco alcuni esempi:

`https://www.programmareinpython.it/video-corso-python-base`
`https://www.python.it/doc/newbie`
`https://www.sololearn.com/Course/Python`

Siete esperti di Python?

Potete cominciare subito a immergervi nell'elaborazione del linguaggio naturale. Questo libro non è un testo informatico avanzato; il contenuto va dall'introduttivo all'intermedio, ed è rivolto ai lettori che vogliano imparare ad analizzare il testo usando Python e i suoi strumenti di analisi del linguaggio. Per informazioni sugli algoritmi avanzati implementati in NLTK e nelle altre librerie, vi consiglio di dare un'occhiata al codice Python presente nella documentazione ufficiale.
Seguite anche i vari consigli che vengono proposti durante la lettura; sono utili per approfondire e stimolare la curiosità.
Vi auguro quindi buona lettura; che sia divertente, ricca di soddisfazioni ed entusiasmante almeno quanto è stato per me scrivere questo libro.

Codice degli esempi

Il codice completo di tutti gli esempi, così come i vari dataset utilizzati, si trova all'indirizzo `https://github.com/serenasensini/analisi-del-linguaggio-con-Python`.