Text Analytics

Negli ultimi anni si sta progressivamente affiancando all'analisi dei dati strutturati - presenti nella banche dati aziendali o amministrative - l'analisi dei dati non strutturati presenti nei documenti, sul web o all'interno dei social media.

L'avvento dei Big Data e l'enorme mole di informazioni prodotte e raccolte quotidianamente sul web rappresentano una fonte informativa nuova e dalle enormi potenzialità. Essa, rispetto all'informazione strutturata, richiede nuove metodologie di raccolta, trattamento e analisi dei dati.

Obiettivi

L'obiettivo dei Text Analytics è raccogliere, trattare ed analizzare le informazioni provenienti da fonti informative non strutturate al fine di estrarne conoscenza utilizzabile all'interno di processi decisionali o di valutazione.

A livello generale possono essere previsti diversi possibili approcci:

- L'analisi di documenti non strutturati di cui non si conoscano i contenuti, al fine di classificarli e identificare i temi e gli argomenti trattati al loro interno, riconducendoli ad apposite classificazioni standard o pensate appositamente per tali documenti
- L'analisi di documenti con l'obiettivo di individuare al loro interno contenuti di interesse definiti a priori, per comprendere come tali argomenti vengano trattati e applicando ad essi metodologie di analisi che ne valutino il contenuti (Esempio: *Sentiment Analysis*)

Fasi

Gli approcci descritti prevedono l'applicazione di metodologie differenti ma condividono lo stesso processo di trattamento dell'informazione.

REPERIMENTO DELL'INFORMAZIONE

Il dato destrutturato può essere messo a disposizione sotto forma di documenti o essere presente sul web, ma in ogni caso è necessario un passaggio che consenta di accedere all'informazione, di raccoglierla e di conservarla in un formato che consenta le successive elaborazioni. A tal fine sono previste attività di *crawling* che permettono, una volta definiti gli ambiti della raccolta (canali informativi, formati, ecc.) e le specifiche dei documenti (documenti riguardanti un certo tema, in un certo intervallo di date, ecc.) di reperire tali documenti e di raccoglierli in modo automatico, prevedendo anche attività periodiche di aggiornamento.

TRASFORMAZIONE DEL TESTO

Una volta raccolto il dato si presenta nella sua forma destrutturata, ma per poter procedere è necessario trattarlo e riportalo ad una forma più semplice che ne consenta l'analisi depurandoli di tutti i fattori di "disturbo" che l'utilizzo della lingua corrente in luogo di un linguaggio formalizzato può comportare.

Nel corso di questo passaggio il testo viene semplificato (*parsing*), eliminando tutte le parole che non portano con sé significato (articoli, preposizioni, ecc.); viene standardizzato, riconducendo alla stessa forma tutte le parole che possono presentare diverse declinazioni o forme verbali (*stemming*), o riconducendo allo stesso termine eventuali sinonimi (*synonym detection*) o concetti (*parts-of-speech and noun group detection*); viene valutato, assegnando pesi diversi alle parole in base all'importanza che si intende dare loro in fase di analisi o alla loro posizione nel testo (*term and frequency weighting*); e viene classificato, riconoscendo al suo interno le parole chiave che fanno riferimento ad un determinato tema.

STRUTTURAZIONE DEI CONTENUTI

I documenti si presentano a questo punto in una forma semi strutturata, composti cioè in parte da informazioni strutturate, ad esempio le classificazione assegnate al passo precedente, e in parte da informazioni destrutturate.

In questa fase il testo destrutturato viene analizzato e a seconda delle esigenze possono essere applicate diverse metodologie di *text mining* finalizzate all'analisi del testo al fine di valutarne i contenuti (*content categorization*), l'orientamento (*sentiment analysis*) o la relazione con altri documenti analizzati.

L'analisi semantica del testo consente inoltre di generare a partire da esso ontologie che ne descrivano i contenuti o di inserirlo all'intento di ontologie esistenti.

ANALYTICS

Una volta strutturati in tal modo i testi possono essere analizzati in base alle esigenze, adottando - quando necessario - strumenti visuali che facilitino sia la consultazione del risultati aggregati, sia l'analisi di dettaglio dei singoli documenti e delle relazioni esistenti tra diversi documenti.

La *Social Media Analytics* rappresenta una porzione in costante crescita dei Text Analytics. Essa focalizza la propria attenzione sulle informazioni messe a disposizione all'interno dei social media (Twitter, Facebook, LinkedIn, ecc.).

La particolare struttura dei social media consente infatti non solo di reperire il testo al loro interno - come ad esempio post, commenti e *tweet* - ma anche diverse informazioni legate agli utenti che immettono tali informazioni (genere, età, rete di conoscenze, ecc.), oltre a una serie di metadati legati al documento (date di pubblicazione, numero di *retweet*, i *like*, ecc.): tali informazioni aggiuntive, in forma strutturata, consentono di articolare le analisi successive fornendo diversi elementi grazie ai quali caratterizzare maggiormente i soggetti oggetto dell'indagine e la relazione tra le loro caratteristiche e i commenti forniti.