1.1 Analizzare il testo

1.2 Le informazioni nascoste del testo

Da una lettera, un romanzo, un saggio, un manuale e qualunque altro testo in linguaggio naturale si possono ricavare molte informazioni “nascoste” che vanno al di là dello specifico contenuto.

Per ricavarle servono elaborazioni particolarmente laboriose. Troppo laboriose per farle “a mano”.

Ma se il testo è in formato elettronico possiamo usare gli strumenti del Natural Language Processing, l’elaborazione del linguaggio naturale.

Cominceremo con l’operazione più semplice, cioè contare, per poi analizzare il testo in maniera sempre più raffinata…

1.3 Contare caratteri e token

Di solito, per prendere confidenza con un testo cerchiamo, per prima cosa, di capire

quanto è lungo. Con un'occhiata, valutiamo il "peso" di un libro o le pagine di un articolo.

Ma se il testo è elettronico possiamo conoscere facilmente:

- il numero di caratteri (ce lo dice qualunque programma di videoscrittura)
- e il numero di byte che occupa sul disco, che è di solito è maggiore.

"Il sabato del villaggio", per esempio, conta 1.555 caratteri che diventano 1.620 byte quando viene codificato con Unicode in formato UTF-8. Un carattere Unicode, infatti, può occupare a seconda dei casi da 1 a 3 byte.

Dopo aver tokenizzato il testo, segmentandolo in frasi e in token, possiamo fare due altre operazioni:

- contare i token;
- calcolare il rapporto tra il numero dei token e il numero dei caratteri, che però in buona misura è una caratteristica della lingua e non ci dà molte informazioni sullo stile di un singolo autore.

Naturalmente, questi sono solo i primi passi...

1.4 Calcolare la frequenza delle parole

Le prime informazioni davvero interessanti, che ci dicono molto sul modo di scrivere dell'autore, le otteniamo contando le occorrenze (ripetizioni) di una parola, cioè i token di cui la parola rappresenta l'astrazione.

Questo valore è chiamato "frequenza" della parola.

Da questa prima elaborazione possiamo sapere immediatamente quante parole diverse sono state usate. Un valore che possiamo dividere per il numero totale di token per capire la ricchezza del vocabolario, cioè la varietà del testo.

Se poi disponiamo le parole in ordine di frequenza decrescente otteniamo la "distribuzione di frequenza" delle parole nel testo, che ci permette di conoscere quelle più usate.

Vogliamo fare una prova?

Usiamo Voyant, uno strumento non proprio semplicissimo ma versatile e disponibile online per tutti. Una specie di coltellino svizzero per l'analisi testuale.

In realtà, Voyant è stato sviluppato per la lingua inglese e non tokenizza correttamente i testi italiani. Quindi fornisce risultati imprecisi, ma ci dà comunque un'idea degli strumenti che si possono usare per operazioni semplici come calcolare la frequenza delle parole e analizzare le loro co-occorrenze.

Proviamo a usarlo per analizzare proprio il "Sabato del villaggio".

Secondo Voyant, su 285 token ci sono ben 195 parole uniche. La ricchezza del vocabolario, chiamata qui "densità del vocabolario", è pari a 0,679.

E la distribuzione di frequenza ci dice che le parole più usate sono "festa" (5 volte) e "giorno" (4 volte).

1.5 Analizzare la co-occorrenza delle parole

Nell'analisi di un testo può essere interessante studiare il modo con cui le parole tendono a stare insieme.

Per questo, si possono analizzare le cosiddette "co-occorrenze", o "collocazioni", cioè la frequenza di gruppi di due o più parole, di solito contigue:

- "efferato delitto";
- "fibrillazione nella maggioranza";
- "notte buia e tempestosa"

e così via.

In cronaca nera possiamo supporre che "efferato delitto" sia più frequente di "feroce delitto", anche se il significato è più o meno lo stesso. E anche se il termine "feroce" è in assoluto più frequente di "efferato".

In questo caso la co-occorrenza tra "efferato" e "delitto" è di tipo idiomatico. Cioè non è dovuta a una regola, ma all'uso. È un modo di dire che si è affermato in parte per imitazione.

1.6 Normalizzare il testo

Il conteggio dei token, e quindi delle parole, può dare risultati falsati quando la stessa parola è scritta in modi diversi. Per esempio con un diverso uso delle maiuscole, che lascia spazi di discrezionalità.

Per esempio, la pubblica amministrazione, a volte ha le maiuscole, a volte no.

E poi, guidiamo una Fiat, che naturalmente ha l'iniziale maiuscola, o una FIAT scritta tutta in maiuscolo?

È una questione di stile personale: è evidente che si parla sempre della stessa automobile!

In questi casi potremmo "normalizzare" il testo, per esempio trasformando tutte le maiuscole in minuscole e poi ripetere il conteggio.

Ma dobbiamo fare attenzione, perché possiamo confondere "Italiano" nel senso di cittadino dell'Italia, con l'aggettivo "italiano".

Oppure la città di Brindisi con un brindisi di fine anno!

1.7 Lemmatizzare il vocabolario

La conversione delle maiuscole è solo una delle possibili operazioni di normalizzazione, forse la più semplice.

Talvolta, è più utile una normalizzazione un po' più spinta, chiamata "lemmatizzazione", che consiste nel sostituire a ogni token il lemma originario a cui fa riferimento:

- Se si tratta di un nome, un aggettivo o un pronome, si usa il termine singolare maschile. Per cui, "leoni", "leonessa" e "leonesse" diventano "leone".
- Se si tratta di un verbo, si usa l'infinito. Per cui, "sei", "fu" ed "eravamo" diventano "essere".

Un'operazione simile si fa scrivendo un dizionario, in cui le voci sono identificate da lemmi.

Dopo la lemmatizzazione, è possibile ricalcolare la frequenza dei lemmi, la loro distribuzione e la ricchezza lessicale, che è il rapporto tra il numero dei lemmi e la lunghezza del testo.

I risultati ci dicono molto sulla qualità e sulla ricchezza di un testo. Ma, come al solito, la lemmatizzazione è più complessa di quello che sembra...

1.8 Regole di lemmatizzazione

Per effettuare la lemmatizzazione, la prima cosa che viene in mente è usare le regole di declinazione e coniugazione e applicarle al contrario. Ciascuno di noi, e anche un bambino, sarebbe in grado di procedere senza troppi problemi.

Ma per il computer ci sono due grossi ostacoli:

- la presenza di un gran numero di casi irregolari;
- la mancata conoscenza del contesto.

Per cui di fronte al token “vado”, non saprebbe risalire al verbo “andare”. Probabilmente, applicando alcune regole inventerebbe il verbo “vadare”.

E penserebbe che “uomini” è il plurale di “uomino”.

In realtà, per effettuare la lemmatizzazione automatica di un testo, si usa, un “lessico morfologico”, cioè un vocabolario digitale che enumera per ciascun lemma le forme declinate o coniugate.

Questo risolve molti problemi, ma non tutti.

Di fronte al token normalizzato “mitra” non saprebbe se si tratta di un’arma, di un copricapo papale o di una divinità molto diffusa nella Roma imperiale.

Per questo, in molti casi, la lemmatizzazione fornisce non un risultato unico, ma una serie di alternative che richiedono una successiva disambiguazione. Ne riparleremo…

1.9 Ad ogni token il suo POS-TAG

Costruire un lessico morfologico è un’operazione manuale, lunga e complessa.

Anche perché a ogni voce (che può essere un lemma o una sua forma declinata o coniugata) viene associato un codice chiamato POS-TAG in cui POS sta per “part of speech” ("parte del discorso") e “tag” equivale al nostro “etichetta”.

Facciamo un esempio tratto da un grande progetto di analisi dei testi comparsi in diverse annate del quotidiano La Repubblica, in cui sono stati inseriti (a mano) i POS-TAG e altri attributi di 380 milioni di token.

I principali tag usati nel progetto La Repubblica sono questi...

I nomi dei tag, come puoi notare, sono abbreviazioni di categorie lessicali inglesi.

1.10 Raffinare il TAG-SET

Il TAG-SET è l'insieme dei codici a disposizione per etichettare ogni token e ogni lemma.

I TAG-SET possono essere più o meno ampi a articolati in funzione degli obiettivi e delle risorse a disposizione.

Di solito, in un lessico morfologico si usano TAG con almeno un secondo livello di dettaglio.

Per esempio:

- il tag "NOUN" (nome comune) si articola in "NOUN-M" (maschile) e "NOUN-F" (femminile);
- il tag "DET" può essere qualificato come "DET-POSS" (che indica un possesso, come "proprio" e "altrui") o "DET-INDEF" (che indica un numero indefinito, come in "qualche" e "alcuni");
- Il tag "PRO" può essere qualificato come "PRO-NUM" (pronome numerale), "PRO-PERS" (pronome personale), ecc.
- Il tag "CON" può essere qualificato come "CON-coo" o "CON-sub" per distinguere congiunzioni coordinanti e subordinanti

Si possono anche avere livelli di dettaglio ulteriori, come in "DET-NUM-CARD", che indica un aggettivo numerale cardinale.

Un TAG-SET completo può includere anche decine di tag.

Ma non è tutto, perché a ciascuna voce si possono aggiungere altre informazioni:

- un insieme di attributi, chiamati "feature", come numero e genere di un nome o aggettivo, tempo e modo e numero di un verbo, ecc.;

- la sua frequenza, calcolata in un corpus di testi sufficientemente ampio come nel caso de La Repubblica.

In questo modo, un lessico morfologico oltre a consentire l'analisi grammaticale, è di grande aiuto anche per disambiguare i casi di omonimia, perché indica la probabilità che un determinato termine abbia un certo significato.

1.11 Costruire un lessico morfologico

Come potrai immaginare, costruire un lessico morfologico in cui ogni termine è contrassegnato da un TAG-SET approfondito e da altre informazioni complementari è un'impresa notevole!

Si parte da ampi corpora di documenti, come quello de La Repubblica o il corpus WaCKy, sviluppato in parallelo per diverse lingue europee, e si procede affidando a persone esperte l'assegnazione di tag, attributi e frequenze.

È un lavoro immane, ma possibile grazie a una strategia di "bootstrapping". Funziona così:

- Si comincia con l'annotare a mano, parola per parola, una piccola parte del corpus.
- Si usa questo mini-corpus annotato per "addestrare" un programma chiamato un POS-tagger.
- Col POS-tagger si annota automaticamente una porzione più ampia del corpus.
- Si correggono gli errori (che all'inizio sono tanti!).
- E si ri-addestra il POS-tagger con il nuovo corpus.

Ripetendo più volte il procedimento, si ottiene un corpus di testi annotato sempre più ampio e il numero degli errori diminuisce.

Alla fine, ecco il risultato: un POS-tagger associato a un lessico morfologico. Insieme formano uno strumento prezioso per analizzare automaticamente qualunque documento.

O quasi, perché abbiamo ancora un aspetto da gestire...

1.12 Risolvere le ambiguità

Alla fine del procedimento, otteniamo uno strumento a grana fine, il cui fine ultimo è l'elaborazione automatica del linguaggio.

In questa frase, come vedi, ho usato più volte la parola "fine".

A cosa corrisponderà?

In un lessico morfologico italiano, il token "fine" appare almeno tre volte:

- "lemma: fine; nome comune maschile singolare"
- "lemma: fine; nome comune femminile singolare"
- "lemma: fine; aggettivo"

Per procedere, dobbiamo disambiguare il testo, decidendo volta per volta qual è la voce del lessico che lo descrive correttamente.

Noi che conosciamo la lingua di solito non abbiamo problemi a interpretare correttamente una frase come questa mettendo ogni "fine" al suo posto basandoci sul contesto e sul buon senso.

Ma se a disambiguare è un computer, come può farlo in maniera attendibile?

Come in altri compiti di analisi del testo, il metodo migliore è usare algoritmi di apprendimento automatico, che riescono ad estrarre regole generali da un insieme di esempi. Che però dev'essere piuttosto ampio.

È per questo che uno dei tipici lavori dei giovani linguisti è proprio quello di taggare manualmente pagine e pagine di testi appartenenti a un corpus!

1.13 Parole piene e parole vuote

Il POS-tagging di un testo permette di distinguere parole "piene" e parole "vuote":

- Le parole "piene", o parole "lessicali", sono quelle che hanno un significato autonomo, perché corrispondono a qualcosa o a qualcuno o a un'azione o a una qualità. In italiano, sono parole piene, salvo eccezioni, i nomi, gli aggettivi, gli avverbi e i verbi.

- La parole “vuote”, dette anche “grammaticali” o “funzionali”, invece, non hanno significato autonomo, ma servono a collegare, stabilire relazioni, richiamare, ecc. In italiano sono parole vuote i pronomi, gli articoli, le preposizioni semplici e articolate, le congiunzioni e le interiezioni.

Dopo il POS-tagging, quindi possiamo calcolare facilmente la cosiddetta “densità lessicale”, da non confondere con la ricchezza lessicale di cui abbiamo parlato in precedenza.

La densità lessicale di un testo è il rapporto tra il numero di token che costituiscono parole piene e il totale dei token.

In questi versi di Leopardi ci sono 13 parole piene su 29 token.

La densità lessicale è 0,44.

Perché questo valore è così importante?

Perché è fortemente correlata al tipo di linguaggio.

1.14 Densità lessicali diverse

È stato notato che la lingua parlata ha mediamente una bassa densità lessicale.

A causa della sua immediatezza il discorso è spesso frammentato, con interruzioni e riprese, ripetizione di parole vuote e una rarefazione di parole lessicali.

Il significato, quindi, è espresso più dalla grammatica che dal vocabolario!

È un discorso diverso nella lingua scritta, che in genere è costruita in modo più ponderato. Qui le parole “piene”, lessicali, occorrono con maggiore frequenza.

Ma, come mostrano alcuni studi statistici approfonditi, non per tutti i testi allo stesso modo:

- I più densi sono i testi scientifici. Anche troppo, secondo qualcuno, e questo li rende di difficile lettura.
- I testi letterari, sono in media meno densi, e presentano grandi differenze da autore a autore.

1.15 Parole più piene e meno piene

Un'ultima osservazione: la nozione di densità lessicale potrebbe essere raffinata.

Perché non tutte le parole “piene” sono proprio piene.

Usiamo tutti il verbo “fare” e il sostantivo “cosa”, che danno alla frase un contenuto informativo piuttosto basso.

(come in Ecce Bombo, di Nanni Moretti, con la celebre espressione "faccio cose, vedo gente, ...").

O come nella saga dei Puffi, che sostituiscono quasi ogni verbo con “puffare” e ogni sostantivo con “puffo”.

È per questo che qualcuno ha proposto di attribuire alle singole parole un indice di densità lessicale inversamente proporzionale alla frequenza.

Si sancirebbe così, matematicamente, che “eseguire”, “realizzare” o “costruire” hanno un significato più denso del generico “fare”!

1.16 Continua il percorso