Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

Words and prediction

Rocco Tripodi

rocco@unive.it

Lezioni precedenti

Sinclair: senza l'analisi dei dati reperibili in ampie collezioni di testi (corpora) non vi oggettività nella ricerca in quanto mancano strategie di misurazione. Una gran parte dei significati delle parole sono depositate nella parte superficiale del linguaggio
Studio delle strutture in cui i termini occorrono

Corpus di addestramento: per raccogliere dati quantitativi sui fatti osservati dal corpus (parole, significati, categorie lessicali) per trasformare le regolarità riscontrate in informazioni per effettuare previsioni (es: significati ambigui)

Modello: processo di astrazione che mira ad estrarre la struttura e le caratteristiche dell'oggetto per renderle adatte agli scopi dell'analisi.

Automa: macchina che compie un'operazione in un numero finito di passi.

Token: una qualsiasi sequenza di caratteri delimitata da un *separatore*

Metodo statistico

Statistica: scienza di derivazione matematica che si occupa di studiare e descrivere la realtà fenomenica nei suoi aspetti di rilevazione numerica

Regolarità statistiche: forme e strutture ricorrenti

Utilità

Presentare e descrivere in maniera appropriata le informazioni, trarre conclusioni dall'analisi dei dati e effettuare previsioni

Le fasi:

1. Definizione degli obiettivi
2. Pianificazione della raccolta dei dati
3. Rilevazione dei dati

} Statistica descrittiva

4. Elaborazione metodologica
5. Presentazione dei risultati

} Statistica inferenziale

Campionamento

Popolazione: è un insieme di elementi (persone, oggetti, eventi, ecc.) detti unità statistiche le cui caratteristiche o comportamenti rappresentano i dati oggetto di analisi.

Campione: modello in scala della popolazione

Teoria del campionamento

fornisce i metodi matematici per costruire i campioni
1. individuare i confini della popolazione (libri pubblicati in Italia nel 2010)
2. Estrarre in modo *casuale* il campione: ogni unità ha la stessa probabilità di essere inserito nel campione

Campionamento a strati: la popolazione è suddivisa in sottoinsiemi e su di essi viene effettuato il campionamento in maniera separata

Analisi quantitativa del testo

Tokenizzazione: unità atomica dell'analisi

Classificazione dei tokens

Es: token(posizione ="valore"; pos ="valore" ; ecc.)

le unità di analisi sono descritte mediante l'attribuzione di valori a delle variabili. La definizione delle variabili rispecchia gli scopi dell'analisi e richiede ulteriori analisi per ottenere i valori (lemmatizzazione, pos, ecc).

Modalità di una variabile: tipo di valore che può assumere

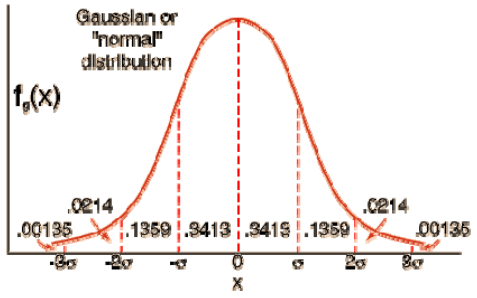
Variabili qualitative: accettano modalità nominali (M-F) o ordinali (livello)

Variabili quantitative: contengono valori numerici (range)

Distribuzione:

comportamento delle unità rispetto ad uno degli attributi

1. Prelevare il valore dell'attributo
2. Raggruppare in classi le unità aventi uguale attributo
3. Contare le unità presenti in ogni classe



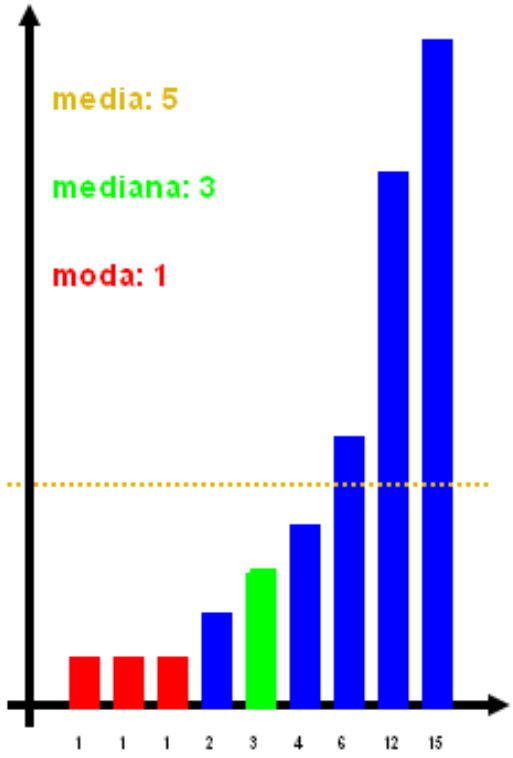
Tendenza verso il centro

Moda: il valore più alto di una determinata categoria

Mediana: punto centrale di una distribuzione
(i valori sono ordinati in modo crescente)

Media (frequenza)

il numero di volte che una data modalità si manifesta nel collettivo di riferimento.
Relativa: il numero di volte che si verifica un evento a prescindere dal numero totale delle prove
Assoluta: il rapporto tra la frequenza relativa e il numero di prove eseguite
f = numero esiti favorevoli /numero prove eseguite
(media aritmetica)
Cumulata: frequenza assoluta della modalità + f.a. della modalità che la precede



Tendenza alla dispersione

Criticità: X = (1; 2; 3; 4; 5) e Y = (3; 3; 3; 3; 3) hanno la stessa media

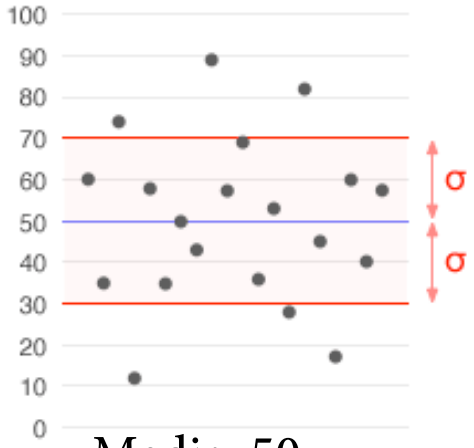
Bisogna calcolare l'imprecisione (scarto tra dato e media)

Range: l'intervallo tra il valore più basso e quello più alto di una classe

Deviazione standard: indica quanto un valore si discosta dalla media aritmetica.

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

Ad ogni valore $\mathbf{x_i}$ si sottrae la media **x**. Il risultato viene elevato al quadrato e accumulato (∑). Viene effettuata la media della sommatoria (n = numero di valori) e il risultato viene posto sotto radice quadrata



Media: 50

Deviazione: 20

Varietà lessicale

Lessico: l'insieme L di tutte le parole possibili di una lingua

Vocabolario di un testo: l'insieme V_T delle parole del testo T

Rapporto type/token = V_T / T

Es: 100 tokens della stessa parola rappresentano un unico type
Oscilla tra 0 e 1. Un valore basso indica che il testo è poco vario

Hapax: parole che ricorrono una sola volta ${V^1}_T$

Classi di frequenza: ${V^1}_T + {V^2}_T + \ldots + {V^{max}}_T$

Liste di frequenza: ordinamento del V_T in forma di lista in cui ad ogni entrata corrisponde una parola accompagnato dal numero di occorrenze

Lemmatizzazione: portare diverse forme di parole ad un unico lemma

Frequenza, rango e prevalenza

Indice di prevalenza: $Ip=[a/a+b]$ ***a*** è il numero di occorrenze di una parola e ***b*** quello di un'altra. ***a*** prevale su ***b*** se il risultato se è $Ip > 0{,}5$

Legge di Zipf: la frequenza di una parola è inversamente proporzionale al suo rango

$$f(P_i) = C/z^a$$

$f(P_i)$ = la frequenza della parola di rango ***i***
C = una costante corrispondente a $f(P_1)$ (varia in base alla lunghezza del testo e alla varietà del vocabolario)
a = indice inverso alla ricchezza lessicale (più è alto e meno è ricco il voc.)

prevede un decremento progressivo della frequenza di una parola proporzionale all'aumentare del suo rango

Dinamica del vocabolario

Crescita: V_T cresce in maniera non lineare. Le prime frasi di T accrescono velocemente V per poi diminuire ma mai annullarsi.

Ripetitività: parlare di un determinato argomento comporta l'uso di parole simili o uguali

Probabilità 1

Evento aleatorio: che accade maniera imprevedibile (lancio di un dado)

Probabilità: misura del grado di incertezza. Previsioni e predizioni

Spazio campionario: insieme degli esiti possibili (Es: testa/croce)

Eventi aleatori semplici: contiene un solo esito (Es: ottenere 6)

Eventi aleatori complessi: coinvolge più esiti (Es: ottenere un pari)

Calcolo della probabilità: $p(A) = |A|/|\Omega|$

$|\Omega|$ = spazio campionario

$|A|$ = numero di esiti che definiscono l'evento

Ottenere un 6 dunque ha probabilità 1/6

Ottenere un pari ha probabilità 1/2

Probabilità 2

Eventi congiunti (Es: ottenere due *3* di fila)

Insieme campionario (Ω) è formato da tutte le coppie di esiti possibili (e_1, e_2) e nel caso dei dati ci sono 36 coppie (6 * 6). Quindi la probabilità di ottenere due 3 di fila sarà uguale e 1/36

Criticità: questo metodo funziona solo quando gli esiti sono equiprobabili

Soluzione: usare la nozione frequenza

Frequenza relativa = f / n

n è il numero di volte che viene lanciato il dado (truccato)

f rappresenta il numero di volte che si verifica un determinato evento

Stima empirica della probabilità (approssimazione)

Lingua e probabilità

Modelli stocastici: descrizione del comportamento del sistema per spiegare e predire le sue manifestazioni

Corpus di addestramento: i dati da esso estratti sono usati per addestrare il sistema e prevedere i comportamenti futuri

Modello linguistico stocastico (MLS)

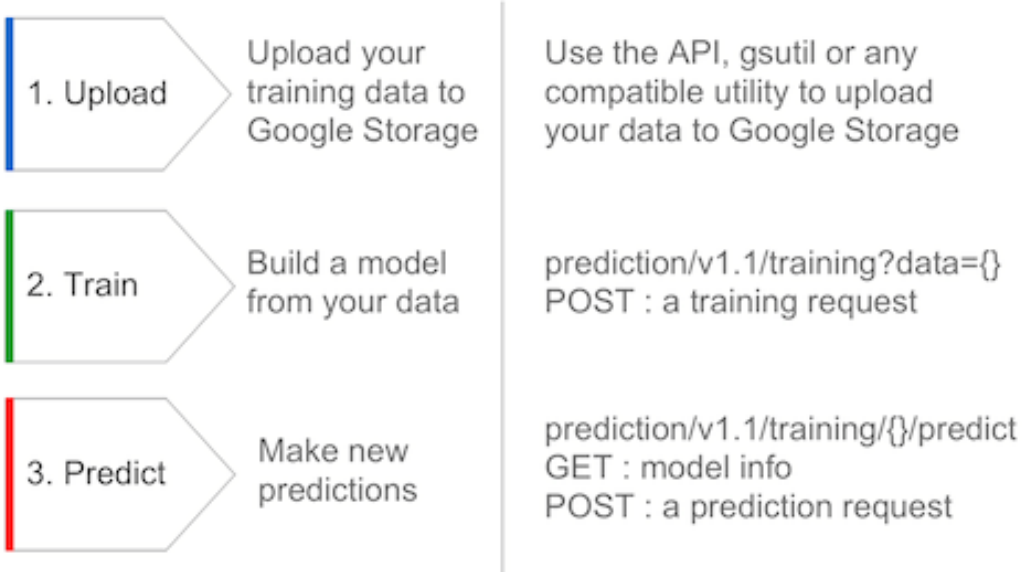
La lingua in questa prospettiva viene vista come un sistema aleatorio. Le parole all'interno di essa sono distribuite in base a vincoli probabilistici. È più probabile produrre una frase con una sintassi definita che una priva di regolarità

Pinocchio è un burattino di legno
un di Pinocchio legno burattino

Google prediction API

API: application programming interface

Machine learning: AI permettere all'elaboratore di sviluppare comportamenti partendo dalla valutazione di dati empirici



Catene markoviane 1

Sono una famiglia di MLS che stimano la probabilità di una parola a partire da un certo numero di parole che la precedono direttamente nel testo.

Ordine: specifica il numero esatto di parole che il modello considera per la stima della probabilità. Si dice di primo ordine se considera un solo antecedente.

Ordine zero: modello ad urna

La probabilità di generare una frase è uguale al prodotto delle probabilità di generare ognuna delle sue parole. Generare una frase come
Il cane dorme lì
con un lessico di 8 parole (il,un,pesce,cane,dorme,nuota,là,lì) è uguale a
$1/8 * 1/8 * 1/8 * 1/8 = 1/2^{12}$
Criticità: le parole non si combinano tra loro con la stessa probabilità. La scelta di una parola modifica la probabilità delle altre

Catene markoviane 2

Catene di ordine 1

Le parole nel testo non si combinano in maniera casuale ma secondo schemi precisi.

Probabilità condizionata

$P(e_1 = il\ ,\ e_{1+1} = cane) = p(e_1 = il) * p(e_{1+1} = cane \mid e_1 = il)$

Il lessico a nostra disposizione è composto da 8 unità dove l'articolo *il* può combinarsi solo con *cane* e *pesce*. Quindi la probabilità di formare la coppia *il cane* è:

*1/8 * ½ = 1/16*

probabilità di *il* * probabilità di *cane* dato *il*

Si riesce a ridurre l'incertezza quando gli eventi non sono indipendenti

Informazione e entropia

Entropia: misura il valore informativo di una classe di eventi esclusivi
Gli eventi più prevedibili (ripetitivi) sono meno informativi
Gli eventi rari contengono maggiore informazione

Bit: il numero delle combinazioni di n bit è uguale a 2^n

Parole in bit: per codificare k parole occorrono $log_2\ k$ bit

Entropia puntuale: si calcola sulla probabilità p di una parola v
$log_2\ 1/p(v)$
Le parole più frequenti ricevono una codifica corta, quelle meno frequenti un codifica lunga (codifica variabile)

Co-selezione delle parole in sequenze ricorrenti

Entropia massima: si ha quando le parole formano un flusso caotico in cui ogni parola ha probabilità costante, indipendentemente dal contesto che la precede (urna).

La creatività del linguaggio consiste nel creare frasi nuove partendo da un insieme finito di schemi ricorrenti

Progetti di ricerca

Read the Web - Link

Data Journalism – Link

Google Prediction API - Link