Da un appello di POO

package **poo.testo**

Progetto di una classe Testo per applicazioni sulla similarità di testi

Si deve progettare una classe Testo che a tempo di costruzione riceve un oggetto File associato ad un file di tipo testo. Mediante un'opportuna struttura dati collezione, la classe estrae le parole uniche del file testo, mantiene per ciascuna parola la frequenza di ripetizione, e organizza la successione in ordine alfabetico delle parole. Metodi pubblici della classe includono i seguenti:

String toString() che ritorna, sotto forma di stringa, la successione ordinata delle parole distinte (normalizzate maiuscolo), ciascuna parola essendo accompagnata dalla sua frequenza

int frequenza(String parola)
String parolaPiuFrequente()
Testo retainAll(Testo t)

che crea e restituisce un nuovo oggetto Testo contenente le parole dell'oggetto this (con le relative frequenze) presenti *anche* in t.

double similaritaCoseno(Testo t)

che restituisce la *similarità coseno* tra l'oggetto testo this e l'oggetto testo t, *preliminarmente ridotti alle sole parole comuni*, definita come:

$$\boldsymbol{similaritaCoseno}(\boldsymbol{t1}, \boldsymbol{t2}) = \frac{\boldsymbol{t1} \cdot \boldsymbol{t2}}{|\boldsymbol{t1}||\boldsymbol{t2}|}$$

Si nota che ogni oggetto testo è assimilabile ad un vettore in uno spazio n-dimensionale in cui le coordinate <x,y,z,…> sono le parole e i valori sono le quantità associate alle coordinate (frequenze di parola). A numeratore c'è il prodotto scalare (prodotto interno) dei due vettori/testo. A denominatore c'è il prodotto dei moduli dei due vettori, il modulo essendo:

$$|\boldsymbol{t}| = \sqrt{\boldsymbol{f}(\boldsymbol{p1})^2 + \boldsymbol{f}(\boldsymbol{p2})^2 + \cdots + \boldsymbol{f}(\boldsymbol{pn})^2}$$

dove p1, p2, … sono le parole distinte del testo t, e f(p) denota la frequenza di p.
La similarità coseno di due testi è compresa nell'intervallo reale [0,1]. 0 dice che i due testi non hanno parole in comune. 1 riflette il caso che i due testi hanno le stesse parole anche se in ordine diverso.
Spesso la similarità coseno è usata per dedurre se due testi parlano di uno stesso argomento o sono stati scritti da uno stesso autore etc.

Aggiungere una classe Main col ***main(…),*** che legga i nomi esterni di due file di tipo testo e provveda a calcolare e scrivere la similarità coseno dei due testi.