3. Match Maker

The objective of the program is to find matching lines (MATCH_LINE) given an input line (TARGET) and a body of text (CORPUS).

TERMINOLOGY

Term	Meaning
CORPUS_LINE	A line of text from the corpus in which the matching is to be done
TARGET	The line of text that needs to be matched.
MATCHING_TERM	A term that appears in both TARGET and CORPUS_LINE
MATCH_SCORE	A score between 0 – 1 (two decimal points) indicating the extent of matching (1 indicates perfect match and 0 indicates no match)
VOCABULARY	A specific set of terms that will be used for matching

Version 1

Assumptions

1. CORPUS is available in the form of text made up of several lines.
2. Each line of the CORPUS is made of only terms from VOCABULARY
3. TARGET is one line of text made of several words ONLY from VOCABULARY
4. There will be no redundancy of words within a line, for both VOCABULARY and CORPUS.
5. Use input / output redirection only (No file handling is required).
6. Matching Criteria: Jaccard Similarity

$$\text{MATCH\_SCORE (Per line)} = \frac{\textbf{COUNT}(\text{MATCHING\_TERM})}{\textbf{COUNT}(\text{Terms in CORPUS\_LINE } \underline{\textbf{UNION}} \text{ Terms in TARGET})}$$

Data Format

Input format
First line contains TARGET terms Subsequent lines contain CORPUS_LINE

Output Format
First line contains TARGET terms Subsequent line contains MATCH_SCORE and the CORPUS_LINE

Example

Input lines		Output lines
database relation normalization		database relation normalization
database data document information	=1/6	x.xx:database data document information
relation database normalization	=3/3	x.xx:relation database normalization
relation normalization normal form	=2/5	x.xx:relation normalization normal form
normalization functional dependency	=1/5	x.xx:normalization functional dependency
1NF normalization	=1/4	x.xx:1NF normalization

File naming conventions

match_v1_**Rollno**.c (e.g., match_v1_MT2014001.c) - Roll number should be in upper case

Note: Print the rounded output till two decimals of precision.

Version 2

Assumptions

1. CORPUS is available in the form of text made up of several lines.
2. Each line of the CORPUS is made of many terms including terms from VOCABULARY
3. Each TARGET is a line of text made of several words ONLY from VOCABULARY
4. Matching Criteria: Jaccard Similarity (same as above)
    a. Eliminate the non-vocabulary terms from both CORPUS_LINE and TARGET
    b. Compute MATCH_SCORE

$$\text{MATCH\_SCORE (Per line)} = \frac{\textbf{COUNT}(\text{MATCHING\_TERM})}{\textbf{COUNT}(\text{Vocab Terms in CORPUS\_LINE } \underline{\textbf{UNION}} \text{ Vocab Terms in TARGET})}$$

Data Format

Corpus corpus.txt	Targets targets.txt	Matches matches.txt
First line contains VOCABULARY (CSV) Subsequent lines contain CORPUS	One TARGET per line	TARGET;Highest MATCH_SCORE;CORPUS_LINE with highest MATCH_SCORE (semi-colon separated lines)

Example

	corpus.txt
1	RDBMS,SQL,normal form,FD,transactions,ACID,entity,primary key,relational,database,data
2	relational,database,RDBMS,is,a,widely,used,technology,for
3	the,purpose,of,storing,large,amount,of,data
4	the,structuring,of,the,database,is,done,using,the,process
5	of,normalization,normalization,uses,FD,to,decide,appropriate,normal form
6	once,data,enters,the,database,transactions,play,a,crucial,role
7	data,integrity,is,maintained,using,ACID,properties,for,transactions
8	SQL,language,provides,transactions,to,query,and,manipulate,data,while,preserving,ACID,properties

	targets.txt
1	relational,database
2	database,transactions,SQL
3	transactions,data,ACID
4	SQL,ACID,transactions

Replace P/Q=X.XX below with appropriate numerator / denominator from the formula and the value rounded to 2 decimal places.

	matches.txt
1	relational,database;2/3=0.67;relational,database,…***{full text of line with highest matching score, line 2 }***
2	database,transactions,SQL;2/4=0.50;once,data,… ***{full text of line with highest matching score, line 6 }***
3	transactions,data,ACID;2/4=0.50;once,data,… ***{full text of line with highest matching score, line 6 }***
4	SQL,ACID,transactions;3/4=0.75;SQL,language…***{full text of line with highest matching score, line 8 }***

File naming conventions

match _v2_**Rollno**.c (e.g., match _v2_MT2014001.c) - Roll number should be in upper case