Quora Question Pairs

Sukanya Behera [160714733008]
Chandrabhatta Sriram [160713733042]
Abul Faiz Mohammed Abdul Hadi Mouzzam [160714733108]

Project Guide: E. Shailaja [Asst. Prof. Dept. of CSE]
Project Co-ordinator: R. Sandeep [Asst. Prof. Dept. of CSE]

Methodist College of Engineering & Technology

Introduction [Problem Definition]

- More formally, the duplicate question detection problem can be defined as follows: given a pair of questions *q1* and *q2,* train a model that learns the function:
  - *f(q1, q2) → 0 or 1.*
  - where *1* represents that *q1* and *q2* have the same intent and *0* otherwise.
  - NOTE: *q1* and *q2* are given as string data.

Problem Input [test.csv]

id	qid1	qid2	question1	question2
447	895	896	What are natural numbers?	What is a least natural number?
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?

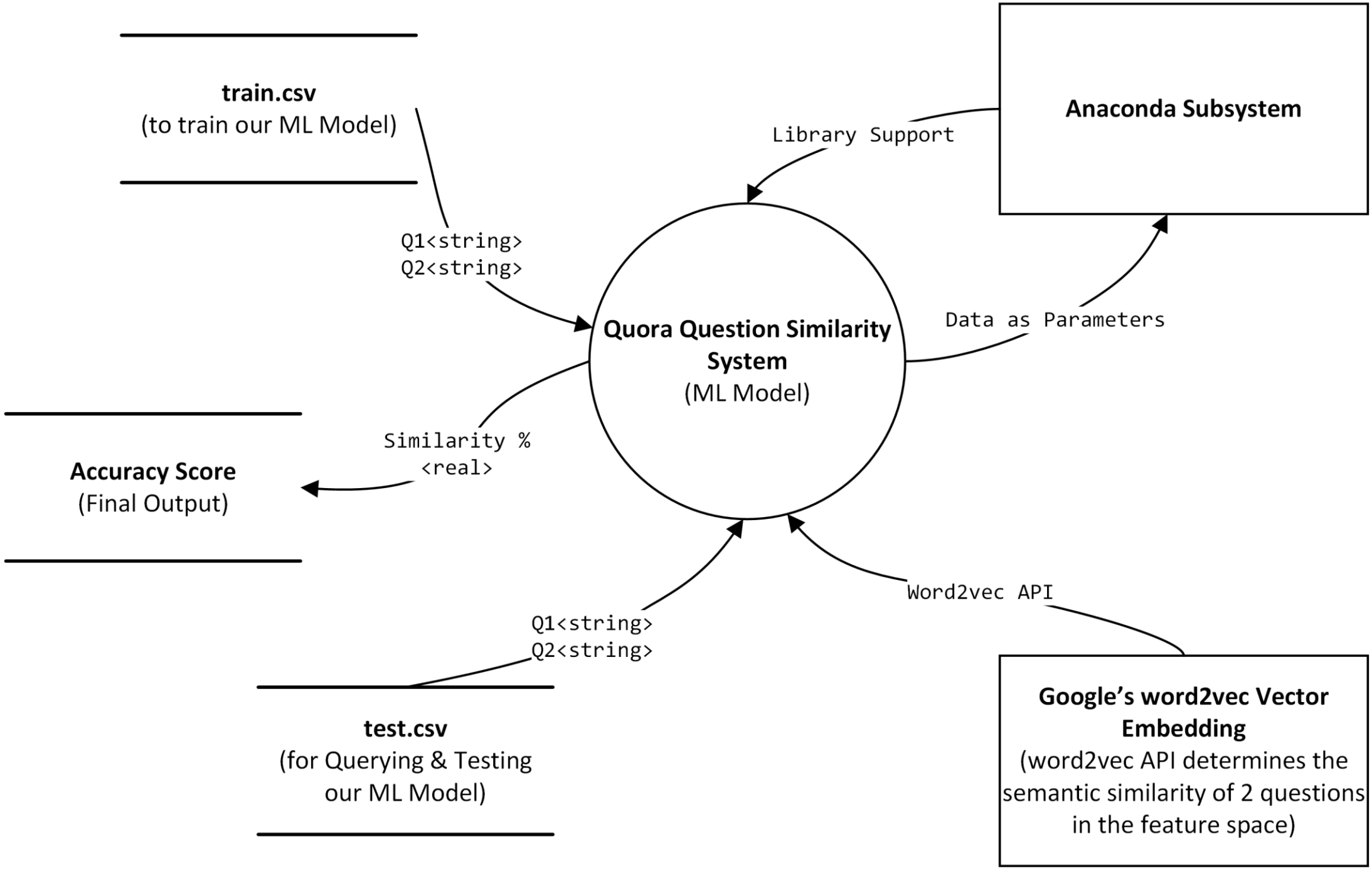
Problem Output [Expected]

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

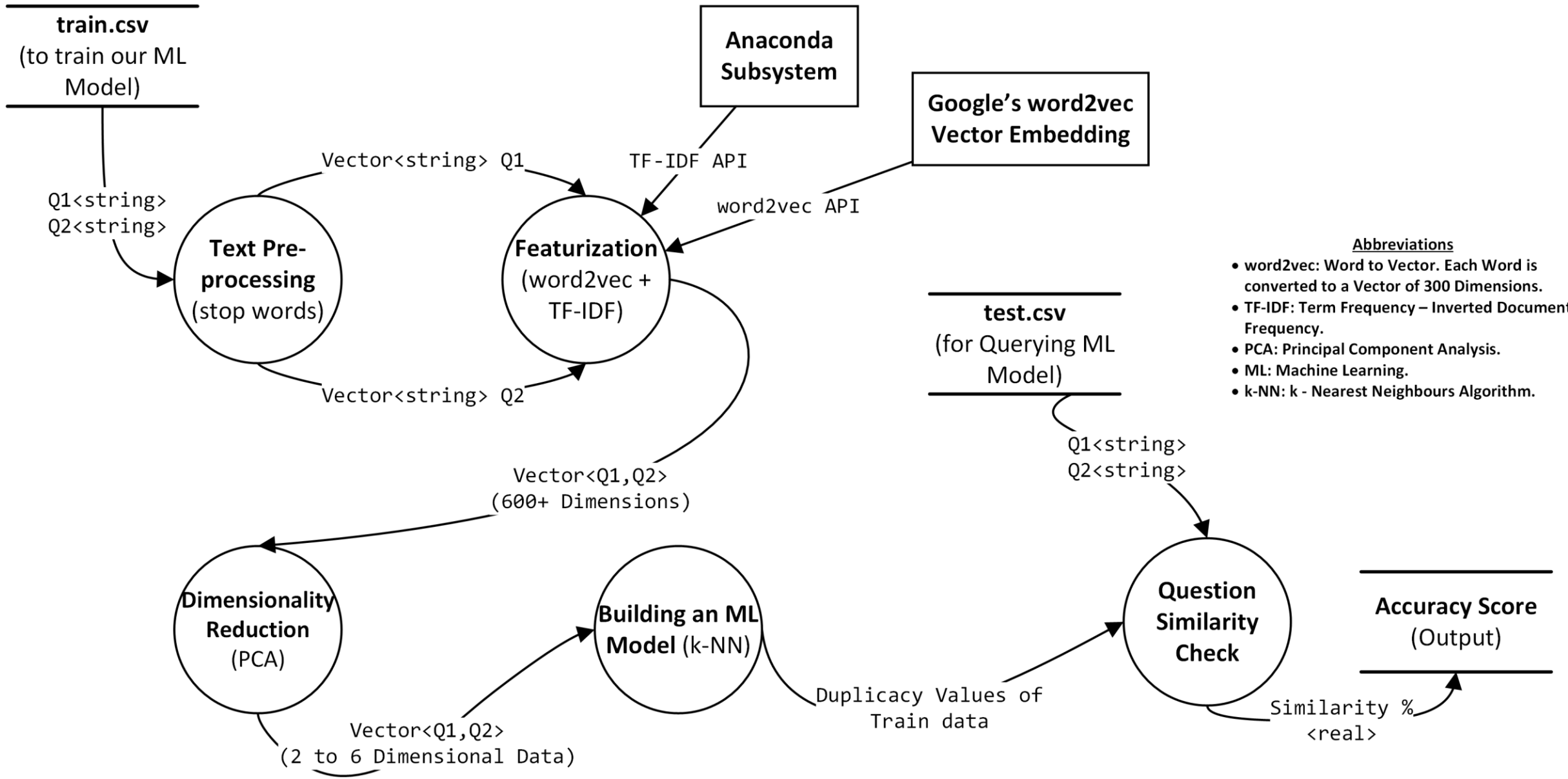
Modules

- Text Pre-processing: *text tokenization*, *uniform text casing*, *stop word removal*, *lemmatization*, *stemming*, etc.
- Featurization / Vectorization of the texts [*q1* and *q2*]: Convert the given *q1* and *q2* into Word Vectors using word2vec. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.
- Apply a Machine Learning Model [using k – Nearest Neighbours algorithm] to get the Expected Output.

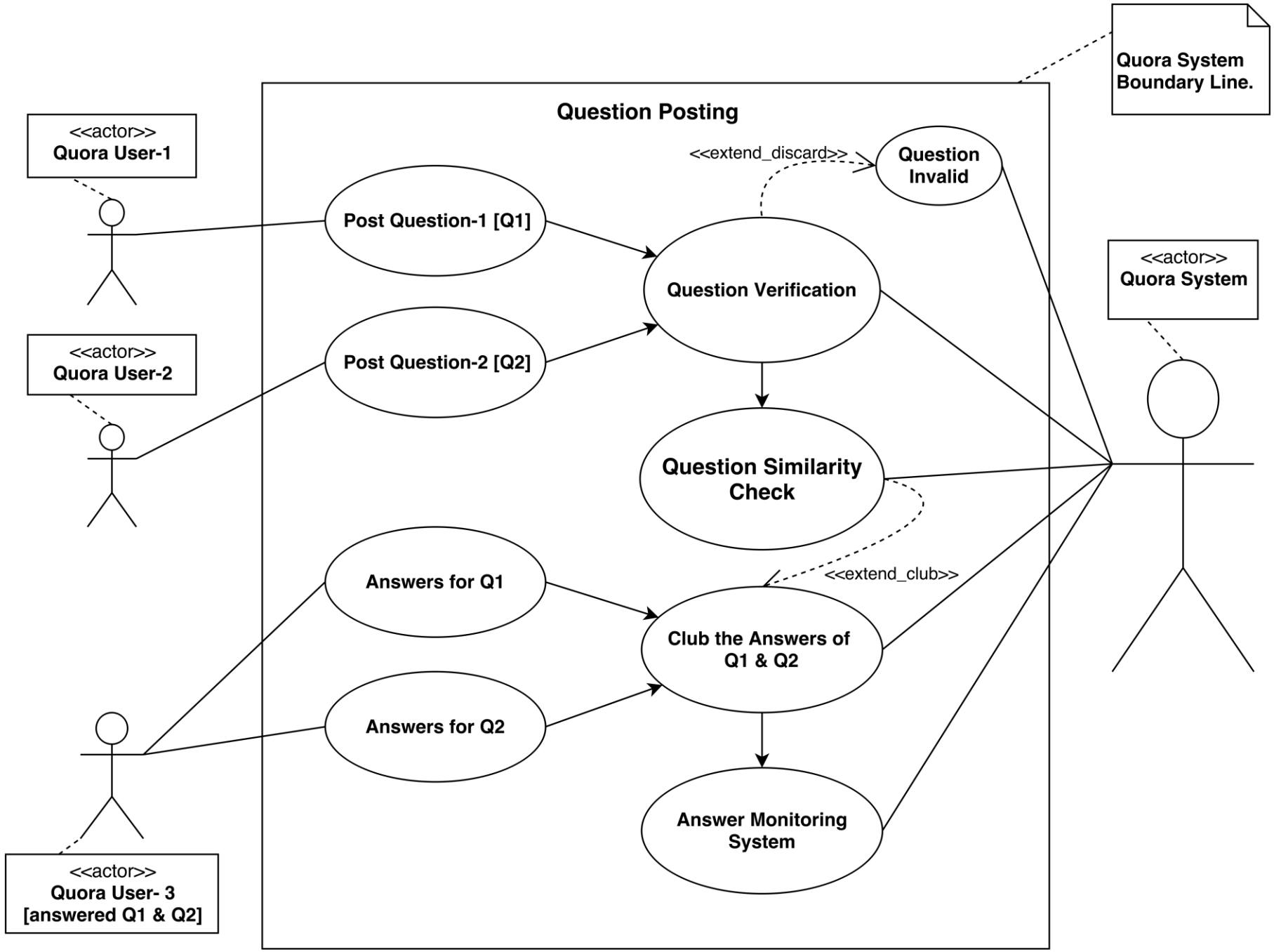
Data Flow Diagram – Level 0


train.csv
(to train our ML Model)
Anaconda Subsystem
Library Support
Q1<string>
Q2<string>
Quora Question Similarity System
(ML Model)
Data as Parameters
Similarity %
<real>
Accuracy Score
(Final Output)
Word2vec API
Q1<string>
Q2<string>
test.csv
(for Querying & Testing our ML Model)
Google's word2vec Vector Embedding
(word2vec API determines the semantic similarity of 2 questions in the feature space)

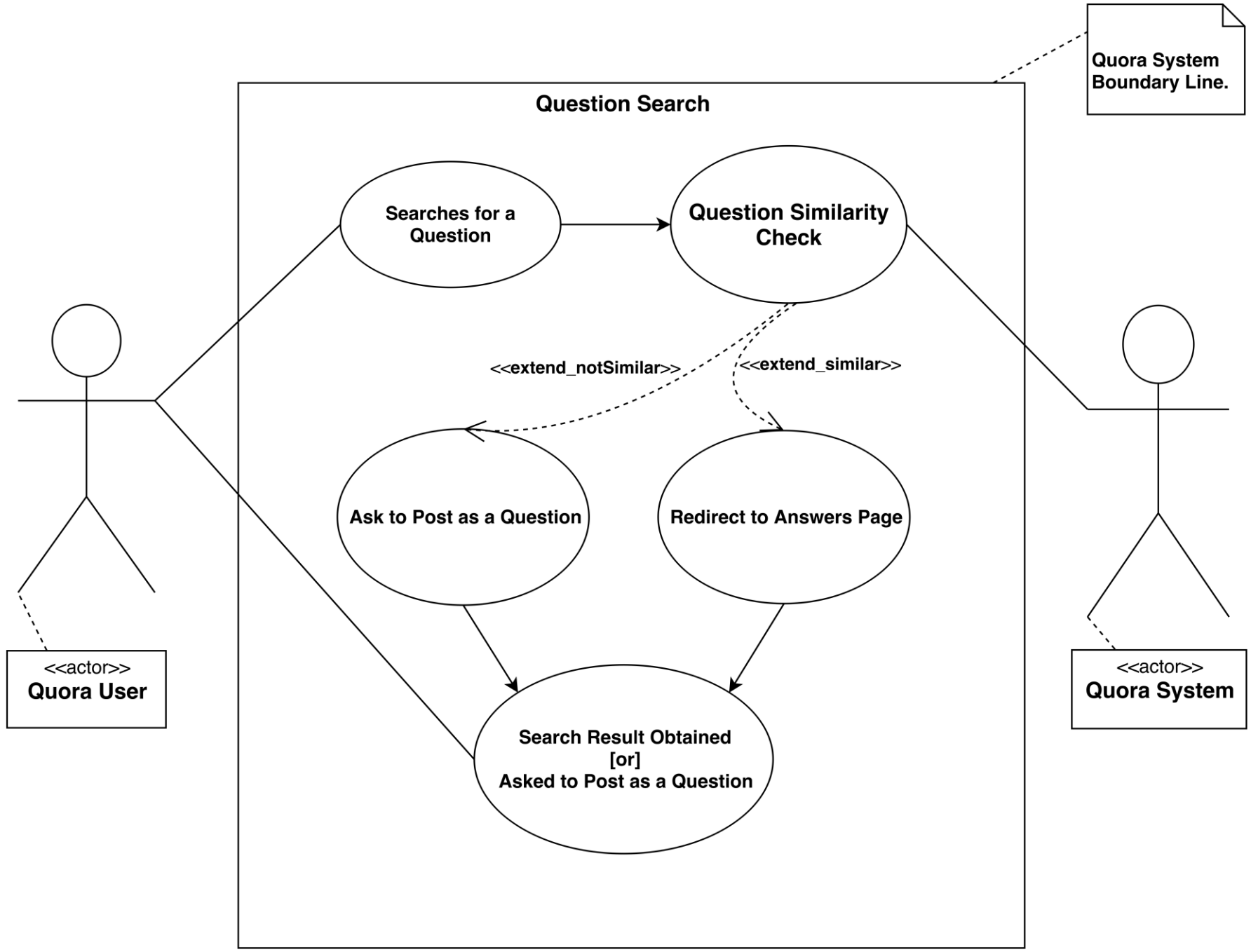
Data Flow Diagram – Level 1
Quora Question Similarity System


train.csv
(to train our ML Model)
Anaconda Subsystem
Google's word2vec Vector Embedding
Vector<string> Q1
TF-IDF API
Q1<string>
Q2<string>
word2vec API
Text Pre-processing
(stop words)
Featurization
(word2vec + TF-IDF)
Vector<string> Q2
Vector<Q1,Q2>
(600+ Dimensions)
test.csv
(for Querying ML Model)
Q1<string>
Q2<string>
Dimensionality Reduction
(PCA)
Building an ML Model (k-NN)
Question Similarity Check
Accuracy Score
(Output)
Vector<Q1,Q2>
(2 to 6 Dimensional Data)
Duplicacy Values of Train data
Similarity %
<real>
Abbreviations
word2vec: Word to Vector. Each Word is converted to a Vector of 300 Dimensions.
TF-IDF: Term Frequency – Inverted Document Frequency.
PCA: Principal Component Analysis.
ML: Machine Learning.
k-NN: k - Nearest Neighbours Algorithm.

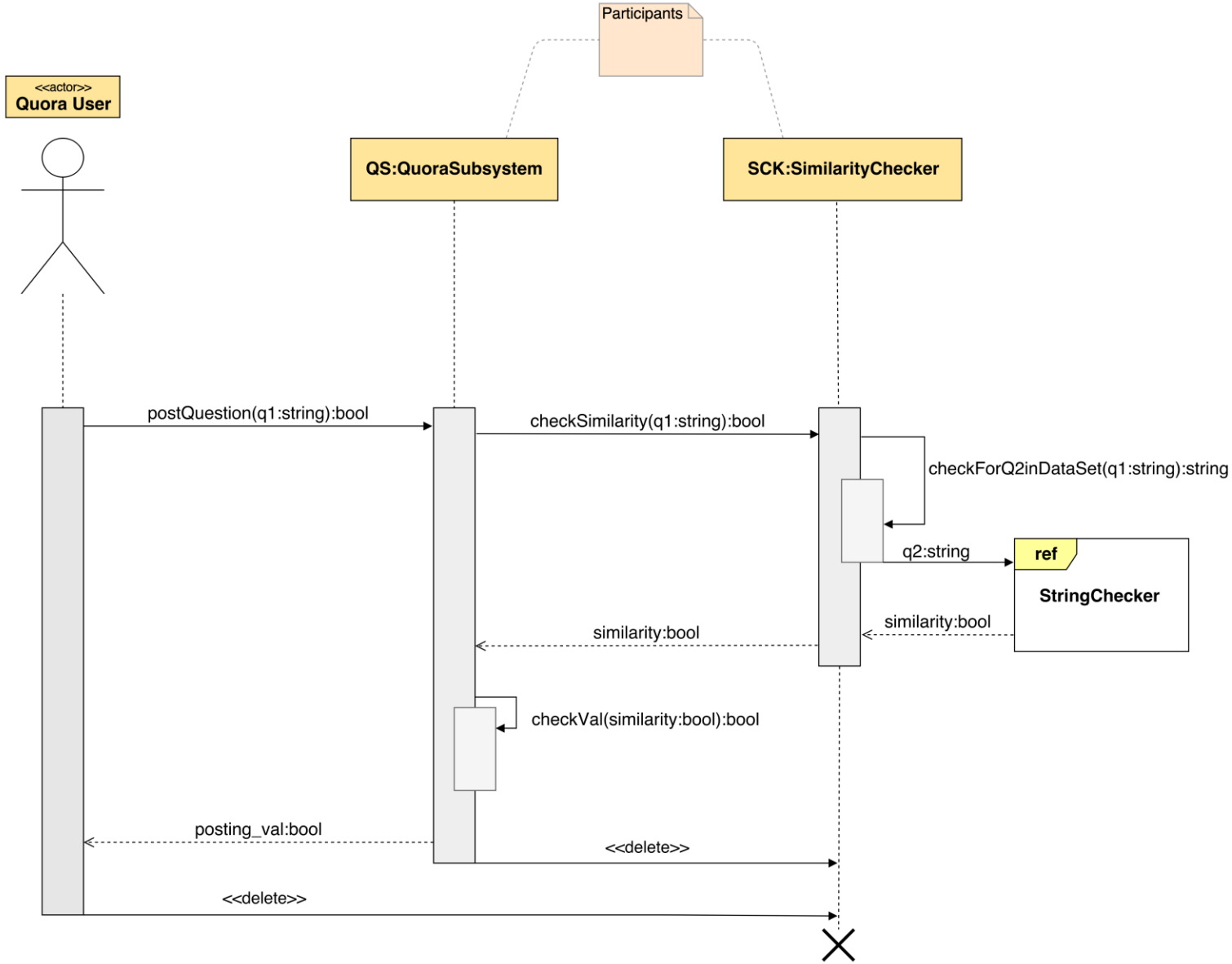
Use Case Diagram – 1 [Question Posting]

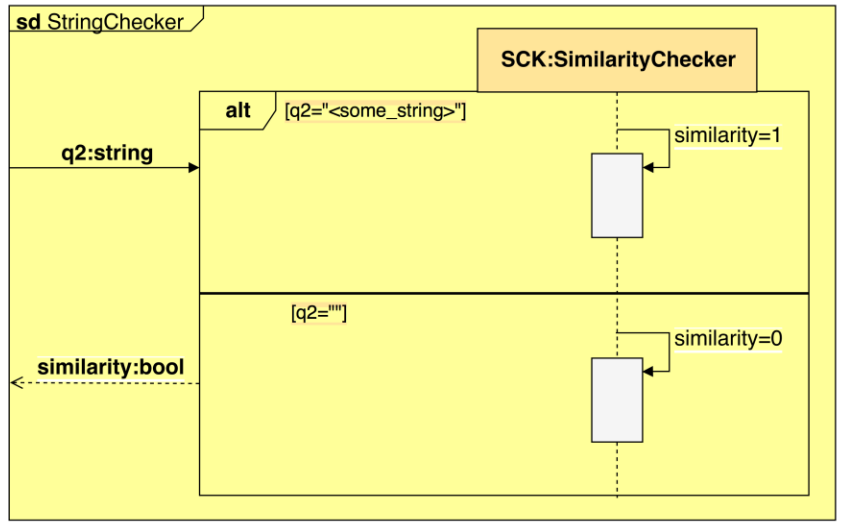

Quora System Boundary Line.
Question Posting
<<actor>>
Quora User-1
<<actor>>
Quora User-2
<<actor>>
Quora User- 3
[answered Q1 & Q2]
Post Question-1 [Q1]
Post Question-2 [Q2]
Answers for Q1
Answers for Q2
<<extend_discard>>
Question Invalid
Question Verification
Question Similarity Check
<<extend_club>>
Club the Answers of Q1 & Q2
Answer Monitoring System
<<actor>>
Quora System

Use Case Diagram – 2 [Question Search]

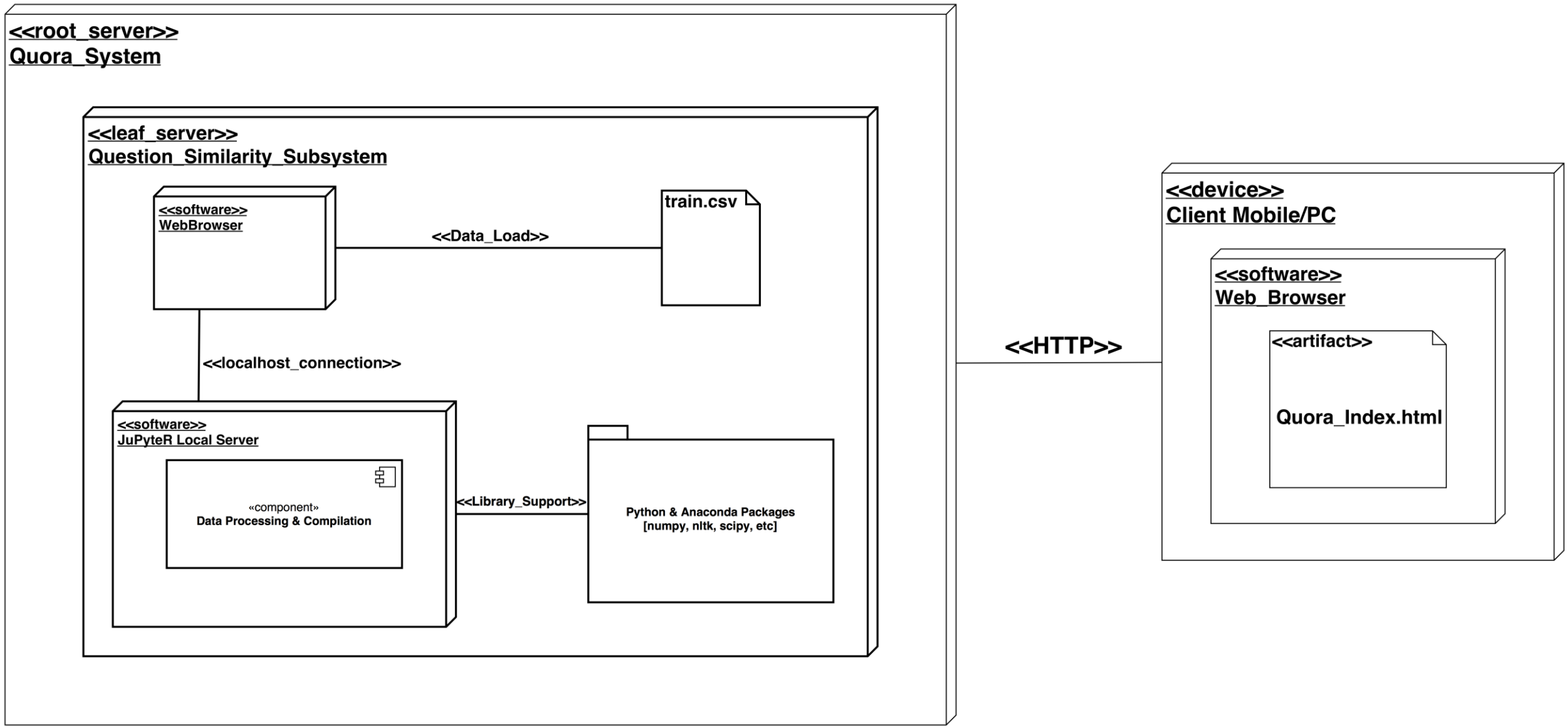

Quora System
Boundary Line.
Question Search
Searches for a
Question
Question Similarity
Check
<<extend_notSimilar>>
<<extend_similar>>
Ask to Post as a Question
Redirect to Answers Page
Search Result Obtained
[or]
Asked to Post as a Question
<<actor>>
Quora User
<<actor>>
Quora System

Sequence Diagram


Participants
<<actor>>
Quora User
QS:QuoraSubsystem
SCK:SimilarityChecker
postQuestion(q1:string):bool
checkSimilarity(q1:string):bool
checkForQ2inDataSet(q1:string):string
q2:string
ref
StringChecker
similarity:bool
similarity:bool
checkVal(similarity:bool):bool
posting_val:bool
<<delete>>
<<delete>>


sd StringChecker
SCK:SimilarityChecker
alt
[q2="<some_string>"]
similarity=1
q2:string
[q2=""]
similarity=0
similarity:bool

Deployment Diagram


<<root_server>>
Quora_System
<<leaf_server>>
Question_Similarity_Subsystem
<<software>>
WebBrowser
<<Data_Load>>
train.csv
<<localhost_connection>>
<<software>>
JuPyteR Local Server
«component»
Data Processing & Compilation
<<Library_Support>>
Python & Anaconda Packages
[numpy, nltk, scipy, etc]
<<HTTP>>
<<device>>
Client Mobile/PC
<<software>>
Web_Browser
<<artifact>>
Quora_Index.html

Algorithm Information

- In ***k*-NN** classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its ***k*** nearest neighbours (***k*** is a positive integer, typically small). If ***k*** = *1*, then the object is simply assigned to the class of that single nearest neighbour.
- Distance between 2 points in the feature space, can be found by taking Cosine Distance (= 1 - Cosine Similarity) between those 2 points.

Thank You!

Any Questions/Queries?