

```
1 import re
2 from importlib.resources.readers import remove_duplicates
3
4
5 class Data:
6
7     # list of possible type of questions that will be used in conditional statements [non-exhaustive]
8     possibleList = ["capital", "distanc", "weather", "movie", "forecast", "cit", "length", "climat", "humidit", "director", "actor"], # direct-answer question
9     ["exercis", "diet", "cook", "workout", "routin", "gym", "activit", "food", "nutri", "wellness", "recipi", "fitness"], # health-related questions
10    ["calendar", "remind", "task", "event", "deadlin", "project", "checklist", "alert", "notifi", "organ"], # productivity questions
11    ["scor", "gam", "jok", "song", "challeng", "puzzl", "music", "lyric", "match", "adventur", "humor"], # entertainment questions
12    ["plu", "minus", "multipl", "divid", "formula", "concept", "ratio"], # mathematical questions
13    ["pric", "mean", "fact", "happen", "latest"], # knowledge-building question
14    ["best", "advic", "help", "tip", "plan"] # advice-seeking questions
15
16    # list of possible cities that the user might reference [non-exhaustive]
17    specificPlaceList = ["Los Angeles", "Chicago", "San Francisco", "Miami", "Austin", "Las Vegas", "Paris", "London", "Tokyo", "Sydney", "Rome", "Barcelona",
18    "Berlin", "Dubai", "Toronto", "Seoul", "Bangkok", "Mexico City", "Riverside", "Cape Town", "California", "Florida", "Texas", "New York", "Nevada", "Hawaii", "Colorado", "Alaska", "Arizona",
19    "Utah", "Illinois", "Michigan", "Washington", "Georgia", "North Carolina", "Tennessee", "South Carolina", "Oregon", "New Jersey", "Virginia", "United States", "Canada", "United Kingdom",
20    "France", "Italy", "Spain", "Mexico", "Germany", "Australia", "Brazil", "Japan", "India", "South Korea", "Thailand", "South Africa", "China", "Russia", "Egypt", "Argentina", "New Zealand", "Pakistan"]
21
22    # list of possible pop-culture references [non-exhaustive]
23    specificPopCultureList = ["The Avengers", "Star Wars", "The Matrix", "Harry Potter", "Jurassic Park", "Titanic", "The Godfather", "Pulp Fiction", "Back to the Future", "The Lion King"], # movies
24    "Apple", "Nike", "Tesla", "Coca-Cola", "McDonald's", "Amazon", "Google", "Adidas", "Disney", "Microsoft", # brands
25    "Friends", "Game of Thrones", "The Office", "Stranger Things", "Breaking Bad", "The Simpsons", "The Mandalorian", "The Crown", "The Walking Dead", "Westworld", # tv shows
26    "The Beatles", "Beyonce", "Kanye West", "Taylor Swift", "Elvis Presley", "Michael Jackson", "Ariana Grande", "Drake", "Lady Gaga", "Eminem", # artists
27    "Super Mario Bros.", "Minecraft", "Fortnite", "The Legend of Zelda", "Call of Duty", "Grand Theft Auto", "Pokémon", "League of Legends", "FIFA", "The Witcher", # video games
28    "Rolls-Royce", "Ferrari", "Lamborghini", "Porsche", "Maserati", "Bentley", "Aston Martin", "Bugatti", "McLaren", "Mercedes-Benz"] # automotive brands
29
30    # method that removes accidental duplicates found by regular expression
31    def __remove_duplicate(self, list):
32        accList = []
33        for i in list:
34            if i not in accList:
35                accList.append(i)
36        return accList
37
38
39    # uses Python Regular Expressions to derive key data in a structural format, replicating a basic version of Natural Language Processing
40    # "QT" = Question Type && "I" = Identifier
41    def parsedData(self, userInput):
42        question_type = self.__remove_duplicate(re.findall(r"(what|who|why|where|when|how|will|can|play|should|is)", userInput))
43        identifiers = self.__remove_duplicate(re.findall(r"(capital|best|cit|length|climat|humidit|director|actor|task|schedul|event|deadlin|project|checklist|alert|notifi|organ|advic|help|tip|pl|distanc|plan|weather|forecast|latest"
44            r"|happen|movi|exercis|song|diet|workout|routin|gym|activit|food|nutri|wellness|recipi|fitnes|calendar|remind|cook|scor|pric|mean|plu|ratio|minus|multipl|divid"
45            r"|jok|gam|fact|formula|concept|challeng|puzzl|music|lyric|match|adventur|humor)", userInput))
46
47        # returns error if either one of the variables above is empty, else, normal dictionary returned
48        if len(question_type) + len(identifiers) < 2:
49            issue = ("QT": "ERROR", "I": "ERROR")
50            return issue
51        else:
52            result = {"QT": question_type, "I": identifiers}
53            return result
54
55    # takes the original sentence inputted by the user and then removes suffixes; local change not global
56    def stemWord(self, userInput):
57        return re.sub(r'b(?:[l]es|[i]ng|[e]d|s|se|l|cation|ization|isation|ized|ised|ied|ous|y|les|tion|ent|ents|er|ers|b)', "", userInput)
```