REPORT

FREQUENTLY ASKED QUESTIONS - NLP

SIVAGURUNATHAN VELAYUTHAM
SXV176330@UTDALLAS.EDU

INTRODUCTION

Frequently asked questions popularly called as F.A.Q, provides a list of Questions and Answers, commonly asked in some context, and pertaining to a particular topic. FAQ are mostly used where questions tend to occur. The convenient way to share FAQ with others is writing an article and storing it in offline. In this case, the articles might not be FAQ – not necessarily questions and answers. However, FAQ used refer all those documents and postings which are offline.

With advancement in Internet, people tend to share the documents or articles in online. People prefer to ask questions in online forums, chat with customer support and reading reviews. These modes helped the user to find the right answer which are relevant for them. In recent times, users having access to a lot of data, where they could not find an appropriate answer for the question. This leads to FAQ as irrelevant if the answer provided in one or more FAQ, user does not get the answer what he/she looks for.

Natural Language processing (NLP) is a branch of artificial intelligence concerned with automatic interpretation and generation of human language like text, voice etc. It solves the problem of finding relevant question for user by applying NLP techniques like stemming, lemmatization and semantic features on the questions.

REQUIREMENTS

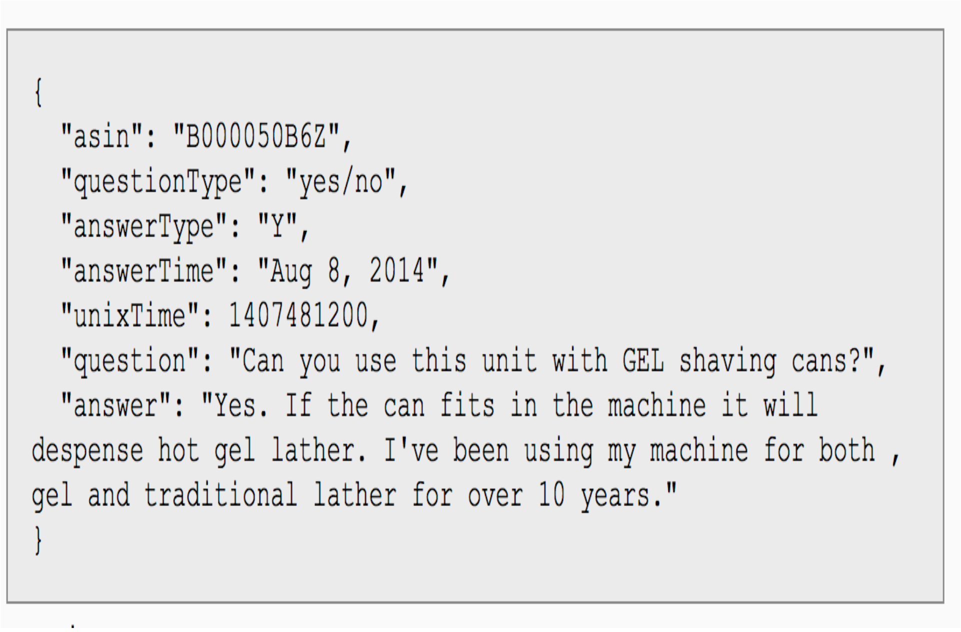
Implement a FAQ that will produce improved results using NLP features and techniques.

Input will be a set of FAQ's and answers. User's input natural language question/statement and generate one or more FAQ's that match the user's input question/statement.

DATASET:

This dataset contains Question and Answer data from Amazon by matching ASINs(Amazon Standard Identification Number).

Sample Question and Answer:

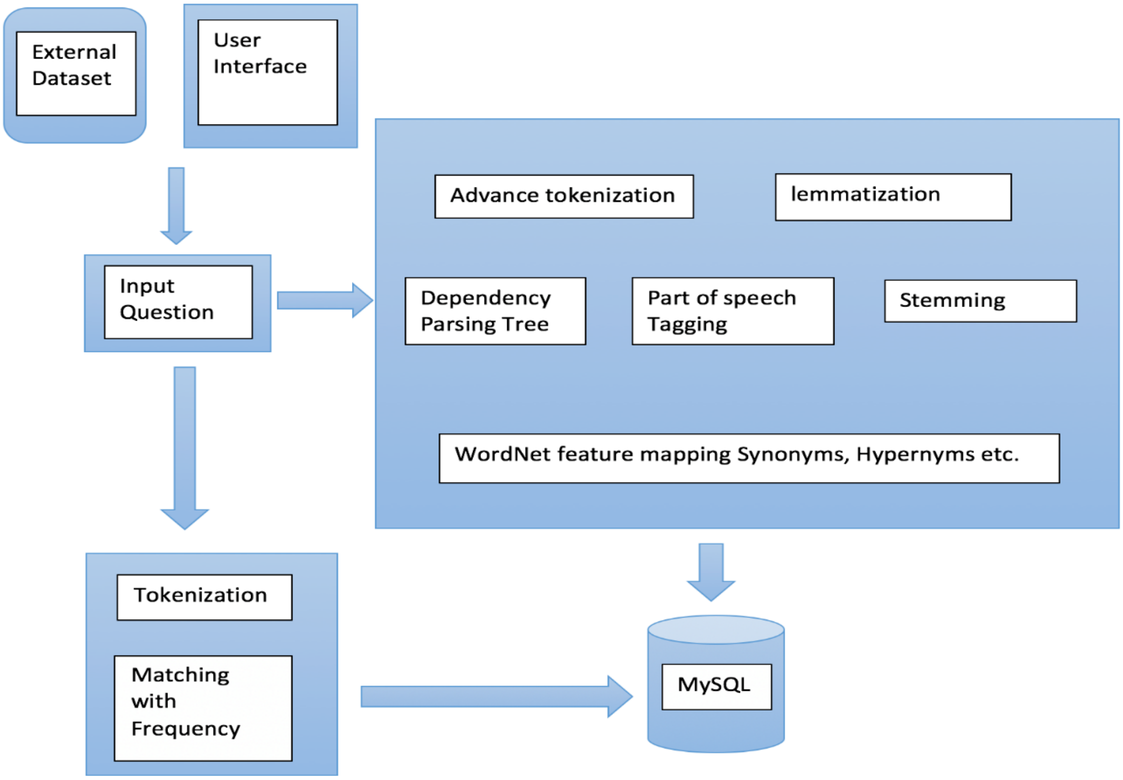


```
{
  "asin": "B000050B6Z",
  "questionType": "yes/no",
  "answerType": "Y",
  "answerTime": "Aug 8, 2014",
  "unixTime": 1407481200,
  "question": "Can you use this unit with GEL shaving cans?",
  "answer": "Yes. If the can fits in the machine it will
despense hot gel lather. I've been using my machine for both ,
gel and traditional lather for over 10 years."
}
```

1. asin- id of the product
2. questionType – type of question, could be yes/no or open-ended
3. answerType – type of answer, could be yes/no or '?' (if the polarity of the answer could not be predicted)
4. AnswerTime- raw answer time stamp
5. UnixTime: converted to unix timestamp
6. Question – question as text
7. Answer – answer as text

Here we went with Pet Supplies in the product category, since it contains more natural and distinguishes text to process than any other category.

ARCHITECTURE



IMPLEMENTATION

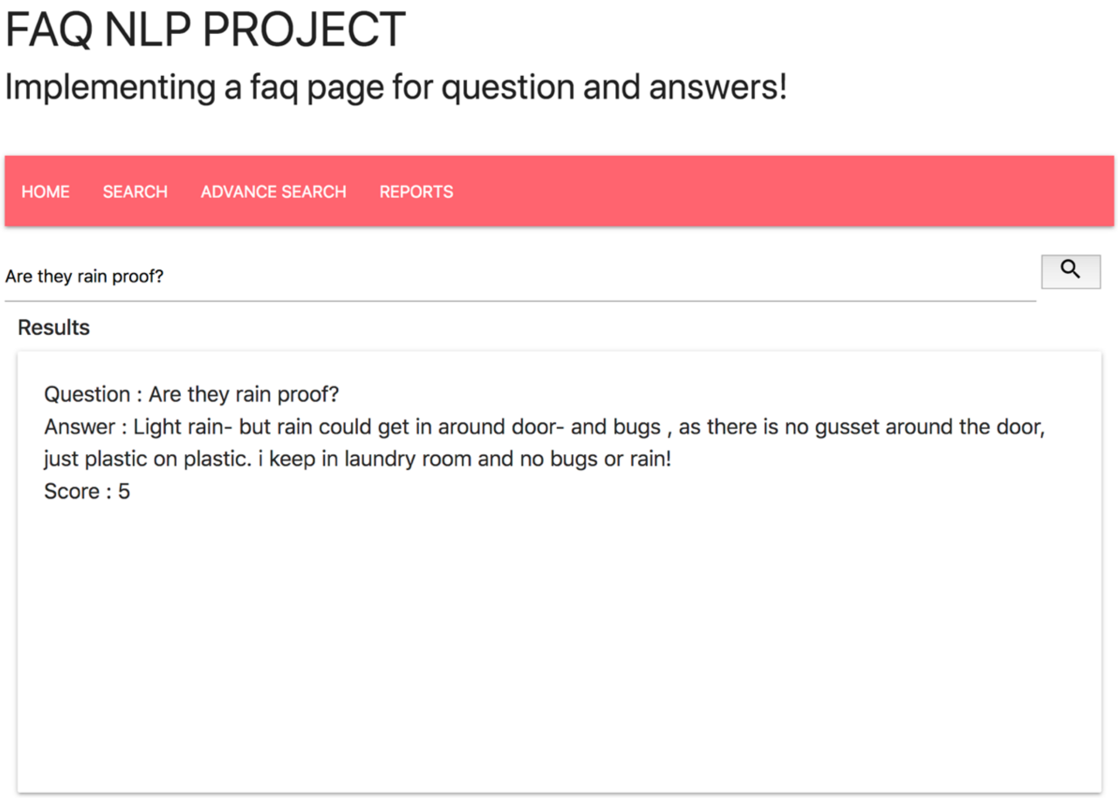
After getting the dataset, following are the steps involved in implementing the NLP Pipeline

1. From the input dataset, parse the JSON data and store the raw data in database
2. Extract the raw data from database, and parse that to the Tokenization using PBT and do unigram count probability and add weights to each of the question and store the result back to the database
3. From the SEARCH_UI Page, user types the question. It will flow through this pipeline and find the unigram probability for the user typed question and match the best probable from the database by brute force i.e. looking at all the records from the database

4. Above method is not efficient, as it is scanning the entire database and comparing it with every record and find the best one.
5. Build an advanced NLP pipeline that can extract features like lemma, stem, part of speech tag, dependency parse tree and synonyms (other meaning from WordNet) for the given sentence and store the results back to the database for building the model.
6. Built a model using the Word2Vec by aggregating the features extracted from the previous step. Here WordNet feature is used for training the model.
7. When user type the question from the ADVANCE_SEARCH_UI Page, the sentence will go through the advance NLP pipeline and extract the feature out of it.
8. Send this extracted feature to the model and predict the most similar words.
9. In the next step, we extract the questions based on the predicted words from the model and display it to the user
10. Update the model, after completing the request from the user. In this way model can be trained a lot more and its accuracy can be improved.

RESULTS

POSITIVE TEST CASE USING BRUTE FORCE



FAQ NLP PROJECT

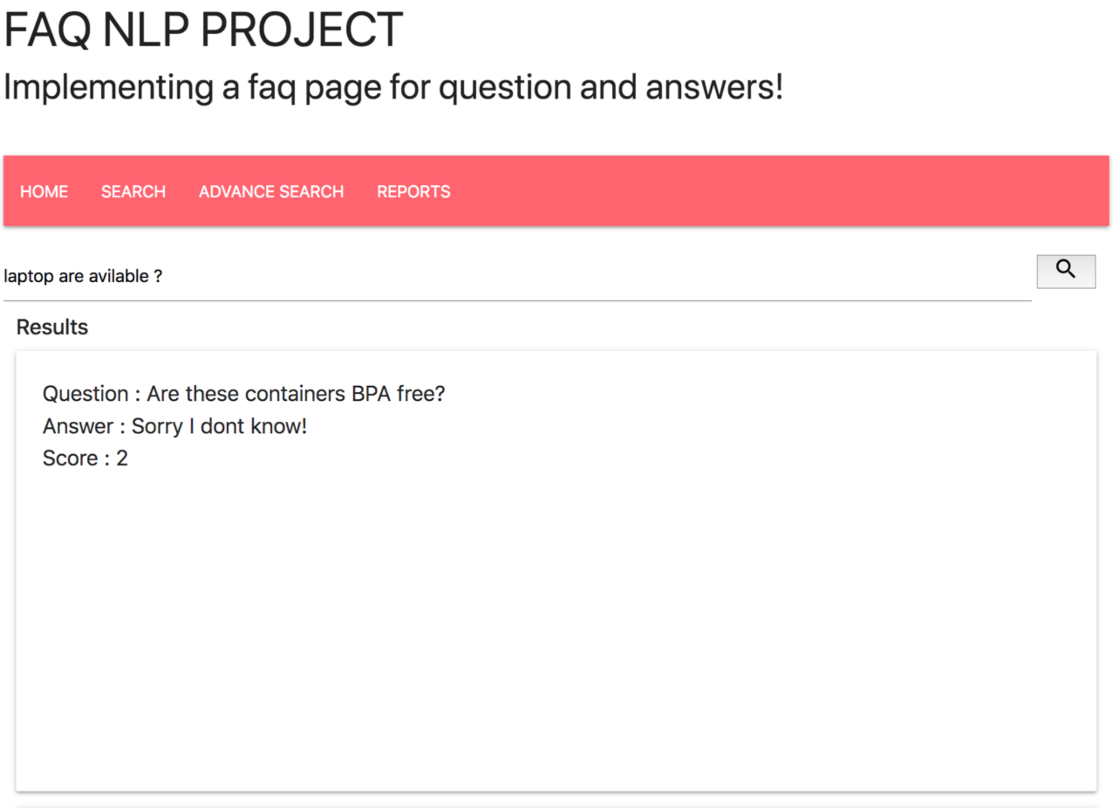
Implementing a faq page for question and answers!

Are they rain proof?

Results

Question : Are they rain proof?
Answer : Light rain- but rain could get in around door- and bugs , as there is no gusset around the door, just plastic on plastic. i keep in laundry room and no bugs or rain!
Score : 5

NEGATIVE TEST CASE USING BRUTE FORCE



FAQ NLP PROJECT

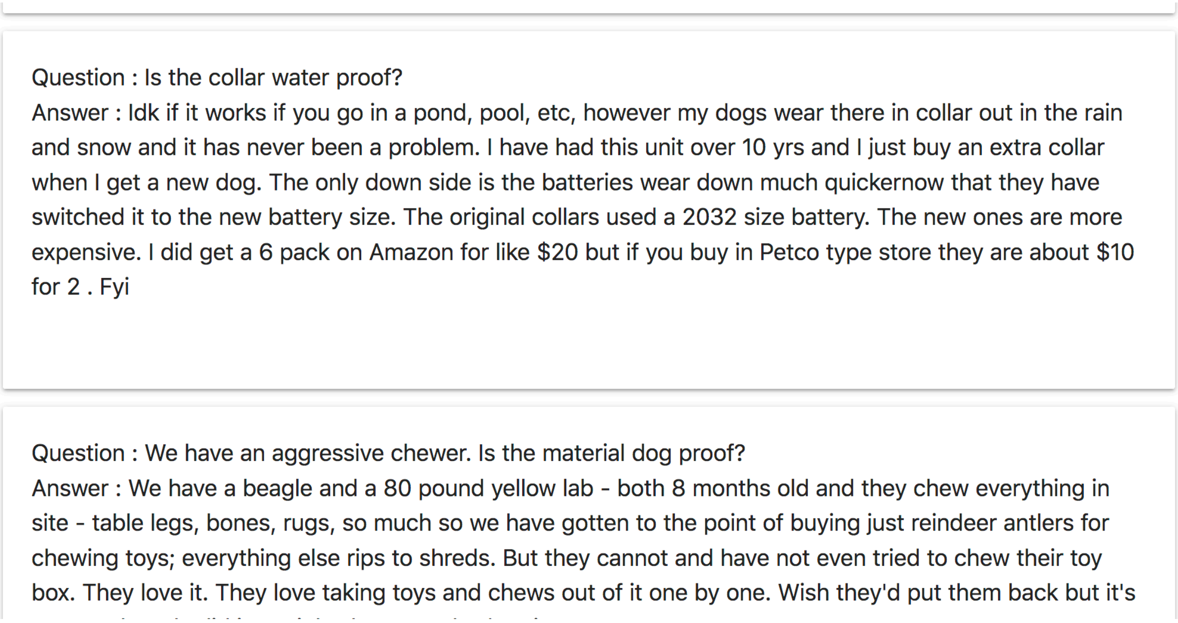
Implementing a faq page for question and answers!

laptop are avilable ?

Results

Question : Are these containers BPA free?
Answer : Sorry I dont know!
Score : 2

TEST CASE USING ADVANCE NLP TECHNIQUE



Question : Is the collar water proof?
Answer : Idk if it works if you go in a pond, pool, etc, however my dogs wear there in collar out in the rain and snow and it has never been a problem. I have had this unit over 10 yrs and I just buy an extra collar when I get a new dog. The only down side is the batteries wear down much quickernow that they have switched it to the new battery size. The original collars used a 2032 size battery. The new ones are more expensive. I did get a 6 pack on Amazon for like $20 but if you buy in Petco type store they are about $10 for 2 . Fyi

Question : We have an aggressive chewer. Is the material dog proof?
Answer : We have a beagle and a 80 pound yellow lab - both 8 months old and they chew everything in site - table legs, bones, rugs, so much so we have gotten to the point of buying just reindeer antlers for chewing toys; everything else rips to shreds. But they cannot and have not even tried to chew their toy box. They love it. They love taking toys and chews out of it one by one. Wish they'd put them back but it's

FUTURE WORK

With recent advancement in the deep learning, we can implement architecture based on the Recurrent Neural Networks (RNN's) and Convolutional Neural Networks (CNN) or combination of both.

Using genism, we implemented a word2vec and find the most similar word from that. Other options like we can build this from RNN combined with bi-LSTM using tensor flow which can give better performance and accuracy for larger datasets.

For improving the query performance, we can use push this data to elastic search or solar so that processing on partial text will be even faster.

TECHNOLOGIES USED

PROGRAMMING LANGUAGES: JAVA, PYTHON

TOOLS USED: STANFORD NLP, DROPWIZARD, HIBERNATE, GOOGLE GUAVA, JACKSON, JWI, FLASK

SERVER: JERSEY

UI COMPONENTS: HTML, CSS, JAVASCRIPT

RESOURCES AND LINKS

1. https://eng.uber.com/cota/
2. http://mccormickml.com/2016/03/25/lsa-for-text-classification-tutorial/
3. http://jmcauley.ucsd.edu/data/amazon/qa/ (DATASET)