ITC-InfoTech
Internship Report

By Debashis Karmakar

Project Title: FAQ-Chatbot

Project Duration: 4th June 2018 to 14th June 2018

Scope:

- Implements a Chat-Bot given a FAQ set, to print top 'k' Question/Answers for users.
- Upon every wrong output and corresponding correction, 'learning' takes place and the next output has improved results.
- Upon incorrect spelling entry, an auto-correct module corrects the input.
- Any technical jargons like 'fa' for 'financial year' are represented appropriately using 'tags'.
- Import Q/A set from an Excel file
- Compare input query with all the available Questions in the dataset using Natural language processing.

Technology Used:

- Python
- NLTK – natural language tool kit
- JSON – data structure used

Implementation Details:

- **Python Packages Used:**
  - JSON
  - Sys
  - Nltk.corpus.stopwords
  - Nltk.corpus.wordnet
  - Nltk.tokenize.RegexpTokenizer
  - Autocorrect
  - Pandas
  - Numpy

❖ **Data Sets used/given:**

- Travel Expense FAQ
- Reimbursement FAQ

❖ **Files developed:**

- TopK2.py -- main implementation file. Prints top likely questions, also updates json data incase of an incorrect output.
- Similarity_calc.py -- Calculating similarity score b/w user Question and given dataset Question
- Make_json.py -- Coverts given xlsx data to json file. One time run file.

❖ **Working:**

- Given xlsx file containing Q/A set. We load the data using pandas dataframe. Some customizations need to be made according to the format of data given.

```
'''
this part needs to be customized according to xlsx file
'''
#loading data_frame --- this part needs to be changed heavily depending on your xlsx file format
df0 = pd.read_excel('DATA/Travel_Chatbot.xlsx', sheet_name='Sheet2')
df0 = df0.drop(columns=["Input"])
df0 = df0.dropna()
df0["Serial No."]=df0["Serial No."].astype(int)
df0 = df0.set_index("Serial No.")
df0 = df0.drop([25])
df0.index = np.arange(1,len(df0)+1)
df0.index.name = "Serial No."
# ------------------------------------------------------------------------
```

- For each question , we tokentize the “Question” string, using nltk package in python. Following which we remove stop words and correct the spellings using autocorrect. For each word, we then check if a corresponding “Synset” exists in nltk’s wordnet . Words like “ITC” which are proper nouns and jargons which are not found in the wordnet, are set as TAGs. The other words are set as “Question-list”.

  Using these two lists and the existing Question/Answer pairs, a dictionary is created which is at the end of the program turned into a JSON file, as follows:

```
    "11": {
        "Question ": "Is parking bill reimbursable?",
        "Question_list": [
            "parking",
            "bill"
        ],
        "Tags": [
            "reimbursable"
        ],
        "Answer": "No parking bills are not reimbursed."
    },
    "12": {
        "Question ": "If I am retiring in middle of Fy the can I claim the entire sampling amount eligible?",
        "Question_list": [
            "retiring",
            "middle",
            "claim",
            "entire",
            "sampling",
            "amount",
            "eligible"
        ],
        "Tags": [
            "fy"
        ],
        "Answer": "yes"
    },
```

Index is set as the Question numbers. The file is stored as data_Reimbursement.json. The "Make_json.py" file handles the above mechanizations.

- The main program , TopK2.py, returns the top 'k' questions based on a user's query. If it fails to do so it requests the user to point out the question "they" would have expected. Following the user's direction it updates the json file with new tags and new list of the users query string.

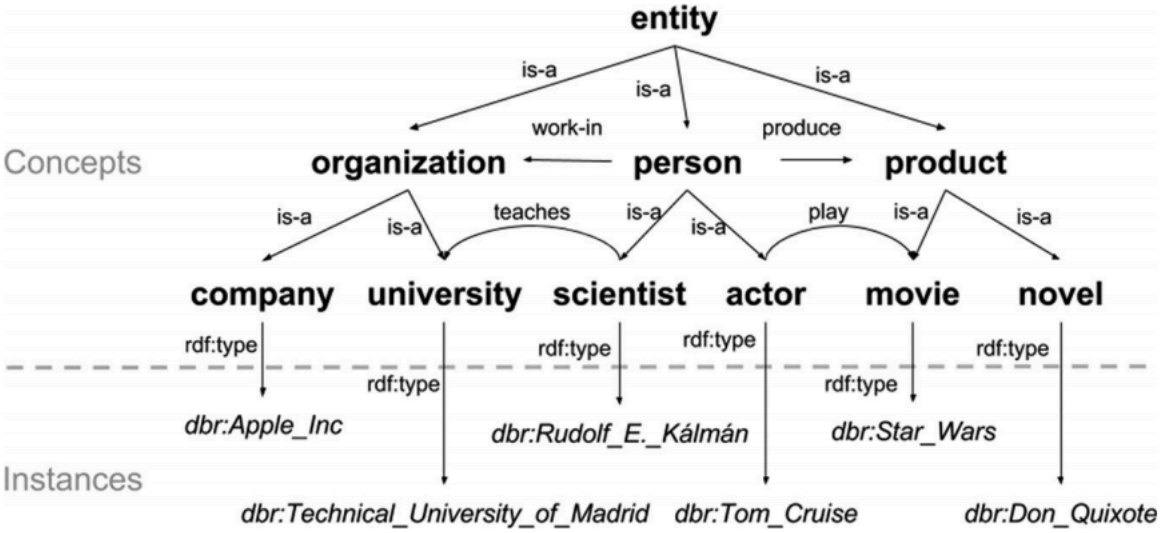
  The program begins by asking the user for their said question. Upon receiving the input, the program tokenizes, removes stop words, corrects spellings and prepares a user-list and a user-tag, similar to the Make_json file.

```
Debashiss-MBP:Code debashiskarmakar$ cd Chatbot
Debashiss-MBP:Chatbot debashiskarmakar$ python topk2.py

----------------------WELCOME TO REIMBURSEMENT FAQ CHAT FOR ITC-----------------------

I take k = 5 :P


Ask a Question >What all shops can i apply reimbursement for?
Tokenized string : ['what', 'all', 'shops', 'can', 'i', 'apply', 'reimbursement', 'for']
After stopword removal : ['shops', 'apply', 'reimbursement']
Checking for spelling errors ...
Corrected/Understood Spellings : ['shops', 'apply', 'reimbursement']
Ulist: ['shops', 'apply', 'reimbursement']
Utags: []

Do you find tag list incorrect?
Ans in 'y' or 'n'>n
Thank you for your time. :)
Utags : []
```

At times, some words are found which are not defined in the wordnet provided by the nltk package. These words get classified as "weird words" and are added to the json file as a list. These words often get filtered into the user-tags , hence before any word makes it to the tags we check for any weirdness.

Weird Words not present in Wordnet

- The calculation of similarity of two sentences is done using similarity_calc.py. Here we use nltk package's path_similarity() function. The given function takes two words "defined" in wordnet and returns a value that depicts the 'closeness' of two words based on meaning, synonym, root words they originate form and distance from each other in the wordnet tree.



An instance of Paths in Wordnet

The path_similarity() function is not commutative(a*b != b*a) hence we normalize the function as [(a*b+b*a)/2].

Hence we calculate word to word similarity score for every word in the User-list and Question-list to give a final score of similarity between the two sentences.

The Tag lists are calculated in a different way however, if the tag is present in both the question and the user tags then a lump sum amount is added to the score.

- These scores are calculated as the main program calls the calc2() function from similarity_calc.py. The scores are stored in a list. The list is then sorted and the top ‘k’ Q/A sets are displayed.

```
Final Score on Q1 = 0.1880952380952381
Final Score on Q2 = 0.09666666666666666
Final Score on Q3 = 0.13888888888888887
Final Score on Q4 = 0.05502136752136751
Final Score on Q5 = 0.05854700854700855
Final Score on Q6 = 0.11666666666666664
Final Score on Q7 = 0.056944444444444436
Final Score on Q8 = 0.11626984126984127
Final Score on Q9 = 0.05787037037037037
Final Score on Q10 = 0.0970973470973471
Final Score on Q11 = 0.08888888888888889
Final Score on Q12 = 0.10158730158730159
Final Score on Q13 = 0.049999999999999996
Final Score on Q14 = 0.041666666666666664
Final Score on Q15 = 0.09722222222222222
Final Score on Q16 = 0.049652777777777775
Final Score on Q17 = 0.046296296296296294
Final Score on Q18 = 0.06474358974358974
Final Score on Q19 = 0.16934156378600823
Final Score on Q20 = 0.16666666666666666
Final Score on Q21 = 0.030555555555555555
```

Scores calculated against every question on the set

- In case expected question does not show up in the top ‘k’, the program goes on to display the next k questions.
  If the expected question is still not found, the program exits.

```
----------------------Displaying top 5 Q/A-----------------------------
Q30) What is this Sampling Reimbursement?

Ans.  All the itc products can be claimed for reimbursement from the company this is done to increase branding of the itc products . Reimburs
ement can be done even if friends and family also buy itc products.

Q23) How we can claim the reimbursement for online shopping?

Ans.  For online product reimbursement under sampling scheme , employees need to provide proof of payment like :
I) for cash on delivery then money receipt or the print of the cash payment screen obtained from the website should be attached with the clai
m.
ii) If payment made through credit card then credit card bank statement copy  showing the debit of the specific ITC product purchased should
be attached with the claim.
iii)If payment made through debit card then the bank statement copy showing the debit of the specific ITC product purchased should be attache
d with the claim.
iv) If any cash back is provided on the credit card or offered by the online website then it needs to be highlighted in the credit card bank
statement and the correct balance amount only should be claimed by the employee.

Q1) Will I get the full amount of the claim or a part of the sample or a part of it ?

Ans.  Yes, you will get the full amount of the sampling according to the entitlement as per your grade.

Q19)   Total time requires towards getting the payment for one claim?

Ans.  15 days will be needed to get the payment of one claim

Q20) What is the payment mode which is used?

Ans.  Payment options are available in the portal option needs to be selected as per continence.



Please check if relevant Q/A is given above (0 if none of them matched)
If answer is there press 1 >0
```

```
 ----Showing 5 to 10 questions in sorted order of         possible prefernce-----

Q29) Can I submit invoices from Spencer

Ans. Yes invoices can be sent for sampling

Q35) Which hotels are eligible for reimbursement?

Ans. Hotels which are under the scheme:
ITC Mughal Agra
ITC Grand Chola
ITC Windsor,Bangaluru
My Fortune , Chennai
My Fortune , BangaloreThe Kakatiya, Hyderabad
The Kakatiya, Hyderabad
ITC Rajputana, Jaipur
ITC Sonar, Kolkata
ITC Maratha, Mumbai
ITC Grand Central, Mumbai
ITC Maurya, New Delhi
Sheraton New Delhi, New Delhi
WelcomHeritage Umed Bhawan Palace, Kota
ITC Gardenia hotel, Bengaluru
WelcomHotel Vadodara, Vadodara
Fortune Bay Island, Port Blair
Classic Golf Resort, Delhi

Q3) Will I get reimbursement if the product is in discount?

Ans. Reimbursement would be permissible even if the products are purchased in discount scheme on offer.

Q32) Where do I access the sampling reimbursement form?

Ans. http://eclaims.net.itc/

Q6) What happens if my bill is lost and the claim is of a huge amount

Ans. If the bill is lost the claim cannot be made
```

- If a question is pointed out by the user, then the user's question list as well as their tags are added to the json file under that question. If a similar question is asked again, it will show up higher up in the top of the list.

```
Q32) Where do I access the sampling reimbursement form?

Ans. http://eclaims.net.itc/

Q6) What happens if my bill is lost and the claim is of a huge amount

Ans. If the bill is lost the claim cannot be made


Enter the Question number that was close to what you expect (0 if none of them matched)

SCROLL UP to check a Bit Please

>29

Thank you for your time :-)
We will be adding this query to our file!

adding shops to Q29

adding apply to Q29

------------------------------BYE BYE-------------------------------
Debashiss-MBP:Chatbot debashiskarmakar$ _
```

Updating Our Json data to match user Preference

- When we type a similar question a second time, our first output is exactly what we want!

```
Debashiss-MBP:Chatbot debashiskarmakar$ python topk2.py

----------------------WELCOME TO REIMBURSEMENT FAQ CHAT FOR ITC----------------------

I take k = 5 :P


Ask a Question >Can i shop in Spencers and claim a reimbursement?
Tokenized string : ['can', 'i', 'shop', 'in', 'spencers', 'and', 'claim', 'a', 'reimbursement']
After stopword removal : ['shop', 'spencers', 'claim', 'reimbursement']
Checking for spelling errors ...
Corrected/Understood Spellings : ['shop', 'spencers', 'claim', 'reimbursement']
Ulist: ['shop', 'spencers', 'claim', 'reimbursement']
Utags: []

Do you find tag list incorrect?
Ans in 'y' or 'n'>n
Thank you for your time. :)
Utags : []
Final Score on Q1 = 0.15337301587301586
Final Score on Q2 = 0.07704545454545456
Final Score on Q3 = 0.10464015151515152
Final Score on Q4 = 0.13050213675213676
Final Score on Q5 = 0.04668803418803419
```

```
Final Score on Q35 = 0.11259920634920635
-----------------------Displaying top 5 Q/A-----------------------------
Q29) Can I submit invoices from Spencer

Ans.  Yes invoices can be sent for sampling

Q23) How we can claim the reimbursement for online shopping?

Ans.  For online product reimbursement under sampling scheme , employees need to prov
I) for cash on delivery then money receipt or the print of the cash payment screen ob
```

```
Q1) Will I get the full amount of the claim or a part of the sample

Ans.  Yes, you will get the full amount of the sampling according to



Please check if relevant Q/A is given above (0 if none of them match
If answer is there press 1 >1
Hope I was useful :)

-----------------------------BYE BYE-------------------------------
```

Limitations:

- Calculation Process is not optimized.

Conclusion:

The Python service required for the given specifications have been implemented.

Further Development:

- A Web/App based UI needs to be prepared.
- Better wordlist needs to be provided for the auto-correct module.
- Optimization of calculation process, presently it is $O(n^4)$.