DataExploration

September 24, 2020

[1]:
```
import nltk
import gensim
import json
import numpy as np
from tqdm import tqdm
from pprint import pprint
```

We define a function to load the json files for the dataset

[2]:
```
def load_data(path):
    with open(path, "r") as fp:
        data = json.load(fp)
    return data
```

[3]:
```
train_path = "train-v2.0.json"
val_path = "dev-v2.0.json"
```

[4]:
```
train_data = load_data(train_path)
data = train_data["data"]
```

We load the word2vec model into memory here

[5]:
```
w2vec_model = gensim.models.KeyedVectors.load_word2vec_format('../../
 ↪GoogleNews-vectors-negative300.bin', binary=True)
```

Now for each question in the dataset we compare each sentence in their respective context using the word mover's distance, a measure of the earth mover's distance for the word vectors between two sentences, and identify the sentence with the lowest score (most similar).

We record whether the answer (according to the dataset) is present in the sentence or not.

[6]:
```
ans_in_sim = []
for topic in tqdm(train_data["data"]):
    for para in topic["paragraphs"]:
        context = para["context"]
        sents = nltk.tokenize.sent_tokenize(context)
        for q in para["qas"]:
            question = q["question"]
            if q["answers"]:
                answer = q["answers"][0]["text"]
            else:
                continue
```

```python
            most_sim = min([(w2vec_model.wmdistance(question, sent), sent) for sent in sents], key=lambda x:x[0])
            ans_in_sim.append(answer in most_sim[1])
ans_in_sim = np.array(ans_in_sim)
```

```
100%|| 442/442 [1:08:34<00:00,  9.31s/it]
```

[7]:

```python
print("Number of questions evaluated:", ans_in_sim.shape[0])
print("Number of questions where the most similar sentence contained the correct answer:", np.sum(ans_in_sim))
print("Percentage correct:", np.sum(ans_in_sim)/ans_in_sim.shape[0])
```

```
Number of questions evaluated: 86821
Number of questions where the most similar sentence contained the correct
answer: 49091
Percentage correct: 0.5654277191002177
```