About:

This Project is about taking pairs of sentences or any other combination of pairs (X1…. Xn, Y2…. Ym) and learn the relation between the pairs.

The main object is to return a prediction Y for a new input X Using Neural Networks.

It might be related to a lot of subjects in our lives, you can understand the idea by seeing the datasets we chose to work with.

The System can work and generate pairs from 3 types of datasets:

1. Only Positive Pairs (Dynamic Generating Negative samples)

2. Mixed Positive and Negative (Organize Positive and Negative samples).

3. Already Organized Pairs with class. (No change needed)

The System have 3 pre-processing pipelines to use, from soft preprocessing to hard preprocessing.

The System 5 Neural Networks Models, some work better some less, it more depends on the data itself.

You can save all the preprocessing files and load to train on any Neural Model you want.

The Neural Model also saved after the Learning process.

Models:

Machine-Learning , Neural-Network:

1) Deep Convolutional Neural Network with language Embeddings Representation.
2) Siamese Neural networks with language Embeddings Representation.

    *embedding_model
    *embedding_model2
    *embedding_lstm_model_manhattan_dist (Siamese)
    *embedding_lstm_model (Siamese)

Evaluation:

K-Fold.

Data-Sets:

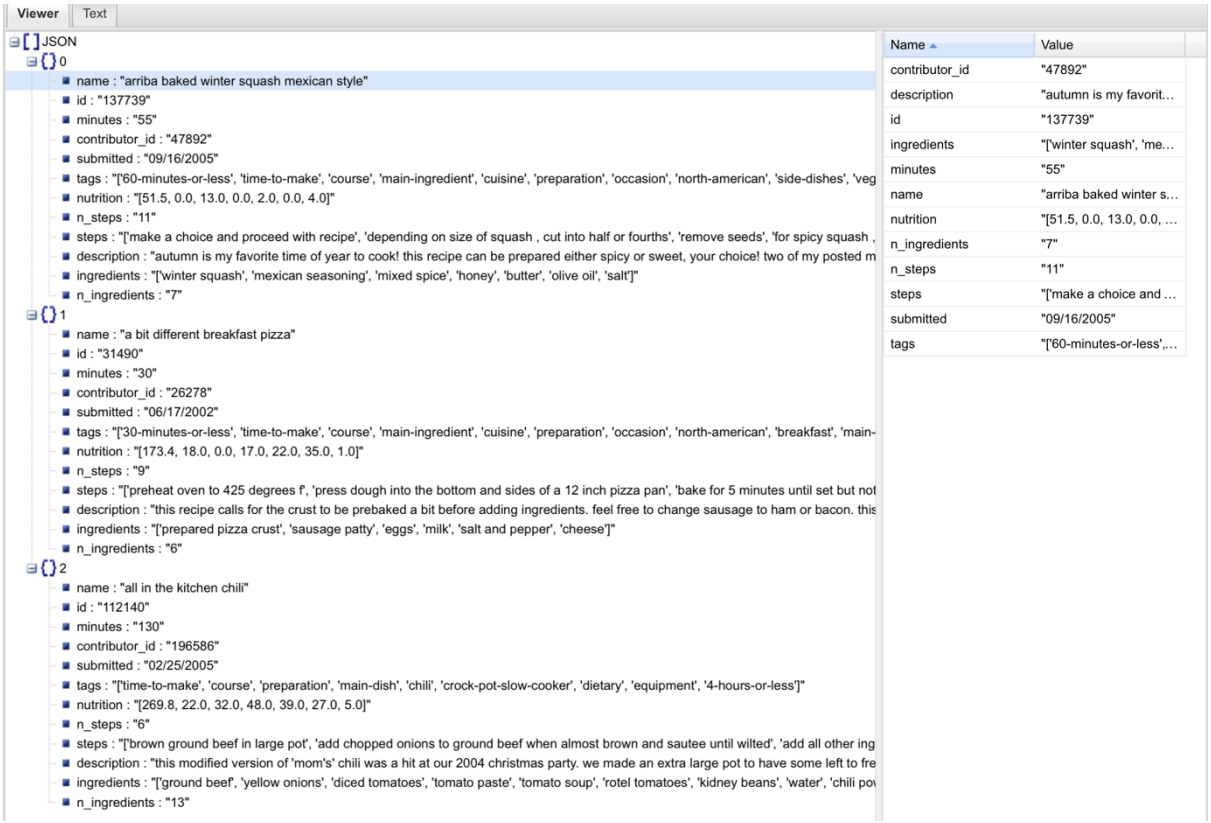
Dataset 1: Recipe description vs Recipe review

Dataset2: Covid19-questions vs Covid19-answer

Dataset 3: Questions-Pairs. (question1 vs question2)

Dataset 1: Recipe description vs Recipe review

About: Recipe dataset, very nice dataset to work with, a lot of options.



Original Files: RAW_recipes.csv and RAW_interactions.csv

Preprocess: Remove unwanted chars.

Export: pairs file **recpies_DS_1.csv**

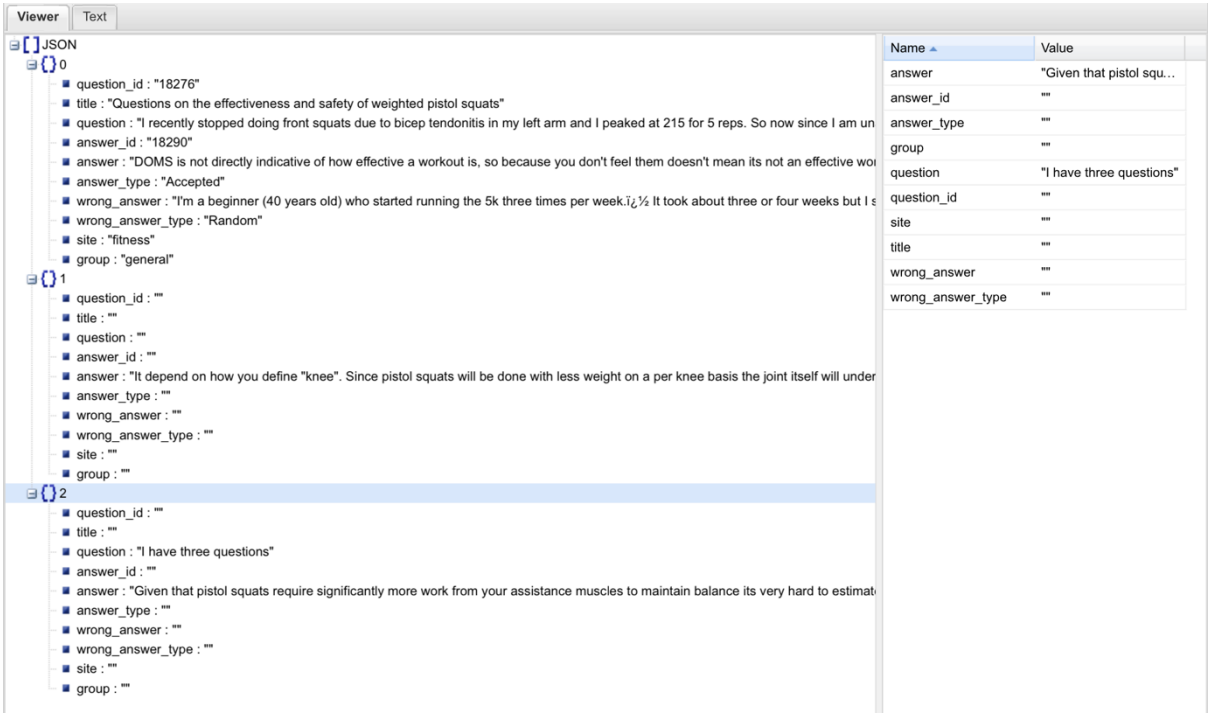
Out[8]:

	description	review
0	autumn is my favorite time of year to cook! th...	I used an acorn squash and recipe#137681 Swee...
1	autumn is my favorite time of year to cook! th...	This was a nice change. I used butternut squas...
2	autumn is my favorite time of year to cook! th...	Excellent recipe! I used butternut squash and ...
3	this recipe calls for the crust to be prebaked...	Have not tried this, but it sounds delicious. ...
4	this recipe calls for the crust to be prebaked...	This recipe was wonderful. Instead of using t...

Dataset2: General questions vs right or wrong answer

Original Files: general.csv

View:



Preprocess:

1. Remove unwanted words that does not fit the regression rule: [A-Za-z0-9] and between 2-50 chars.

2. Remove stop words using the stopwors.txt file

3. Lemmatization

4. Replace capital letters

Export: file **precovid_data.csv**

In [7]:
```
data.head()
```

Out[7]:

	Unnamed: 0	question	answer	wrong_answer
0	0	I recently stopped doing front squats due to b...	\nDOMS is not directly indicative of how effec...	I'm a beginner (40 years old) who started runn...
1	1	I have a lot of fat lose. I weigh 210 pounds, ...	Yes, its theoretically possible. Are you going...	Welcome to the Fitness SE!\n\nMaking your own ...
2	2	After going through this link I am confused ab...	It appears that this article is aimed more at ...	There is a myth that exercise increases the se...

Dataset 3: Questions-Pairs. (question1 vs question2)

Original Files: questions.csv

View:

```
In [6]: data.head()
```

Out[6]:

	Unnamed: 0	question1	question2	is_duplicate
0	0	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0

Preprocess 1:

1. Remove unwanted words that does not fit the regression rule: [A-Za-z0-9] and between 2-50 chars.

2. Remove stop words using the stopwors.txt file

3. Lemmatization

4. Replace capital letters

Export: file **p_c_questions_data.csv**

Preprocess 2:

1. Remove unwanted words that does not fit the regression rule: [A-Za-z0-9] and between 1-50 chars.

2. Remove stop words using the stopwors.txt file (updated the stopwors.txt)

3. Lemmatization

4. Replace capital letters

Export: file **p_c_questions_data2.csv**

Preprocess3:

1. Remove unwanted words that does not fit the regression rule: [A-Za-z0-9] and between 1-50 chars.

2. Lemmatization

3. Replace capital letters

Export: file **p_c_questions_data3.csv**

Project Files description

FinalDataGenerator1.py :

Knows how to work with this type of scheme :

Out[8]:

	description	review
0	autumn is my favorite time of year to cook! th...	I used an acorn squash and recipe#137681 Swee...
1	autumn is my favorite time of year to cook! th...	This was a nice change. I used butternut squas...
2	autumn is my favorite time of year to cook! th...	Excellent recipe! I used butternut squash and ...
3	this recipe calls for the crust to be prebaked...	Have not tried this, but it sounds delicious. ...
4	this recipe calls for the crust to be prebaked...	This recipe was wonderful. Instead of using t...

Only positive samples.

At the main function:

You can generate batches with your own parameters.

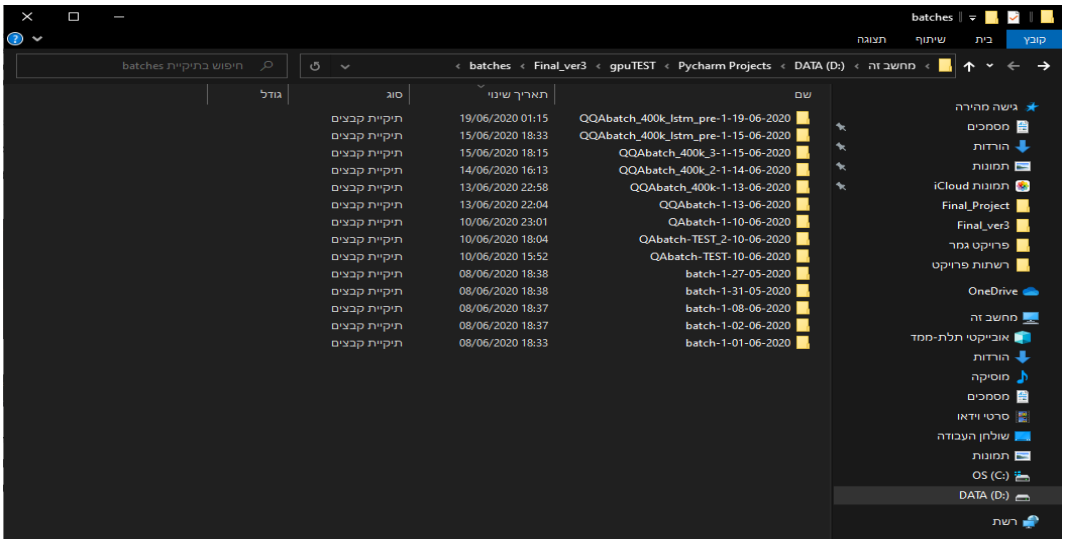
```
batch1 = next(self.generate_batch(1000,1))  # 2000
batch1 = self.arrange_batch(batch1,self.batch_length)
self.save_batch(batch1,1000,1)
batch2 = next(self.generate_batch(self.num_of_rows,1))
batch2 = self.arrange_batch(batch2,self.batch_length)
self.save_batch(batch2,self.num_of_rows,1)
```

```
def generate_batch(self, n_positive=None,
negative_ratio=1):
```

you can change the negative ratio as you like. The default is 1 to 1.

This function yields the batch.

The save_batch function saves the batch to the “batches” folder and creates there your batches.



FinalDataGenerator2.py :

Knows how to work with this type of scheme :

```
In [7]: data.head()
```

Out[7]:

	Unnamed: 0	question	answer	wrong_answer
0	0	I recently stopped doing front squats due to b...	\nDOMS is not directly indicative of how effec...	I'm a beginner (40 years old) who started runn...
1	1	I have a lot of fat lose. I weigh 210 pounds, ...	Yes, its theoretically possible. Are you going...	Welcome to the Fitness SE!\n\nMaking your own ...
2	2	After going through this link I am confused ab...	It appears that this article is aimed more at ...	There is a myth that exercise increases the se...

With positive and negative samples at the same file.(without labels)

At the main function:

```
def main(self):
    #self.load_files(500)
    self.pre_process()
    self.load_files(130000)
    self.data_preperation()

    batch2 = next(self.generate_batch(130000,1))
    batch2 = self.arrange_batch(batch2,self.batch_length)
    self.save_batch(batch2,self.num_of_rows,1)
```

Notice the pre_process() Function , It takes the data and run it through the preprocess pipe line, you can remove or add functions as you like.

It exports the data to this kind of scheme:

```
In [7]: data.head()
```

Out[7]:

	Unnamed: 0	question	answer	wrong_answer	clean_question	clean_answer	clean_wrong_answer	p_c_question	p_c_answer	p_c_wrong_answer
0	0	I recently stopped doing front squats due to b...	\nDOMS is not directly indicative of how effec...	I'm a beginner (40 years old) who started runn...	recently stopped doing front squats due to bic...	DOMS is not directly indicative of how effecti...	beginner 40 years old who started running the ...	[recently, stop, squat, bicep, tendonitis, lea...	[doms, directly, indicative, effective, workou...	[beginner, old, start, run, start, develop, sh...
1	1	I have a lot of fat lose. I weigh 210 pounds, ...	Yes, its theoretically possible. Are you going...	Welcome to the Fitness SE\n\nMaking your own ...	have lot of fat lose weigh 210 pounds and was ...	Yes its theoretically possible Are you going t...	Welcome to the Fitness SE Making your own prog...	[lot, fat, lose, weigh, pound, hope, start, lo...	[yes, theoretically, possible, able, decide, s...	[welcome, fitness, se, program, program, gener...
2	2	After going through this link I am confused ab...	It appears that this article is aimed more at ...	There is a myth that exercise increases the se...	After going through this link am confused abou...	It appears that this article is aimed more at ...	There is myth that exercise increases the secr...	[link, confuse, believe, link, claim, true, af...	[appear, article, aim, condition, activity, ru...	[myth, exercise, increase, secretion, testoste...

Generate batch works a little bit different because of the data set format but yields and saves the results same as the first one.

FinalDataGenerator3.py :

Knows how to work with this type of scheme :

```
In [6]: data.head()
```

Out[6]:

	Unnamed: 0	question1	question2	is_duplicate
0	0	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0

With positive and negative samples at the same file already have labels to each pair.

The preprocess part same as: FinalDataGenerator2.py

```
In [6]: data.head()
```

Out[6]:

	Unnamed: 0	question1	question2	is_duplicate	clean_question1	clean_question2	p_c_question1	p_c_question2
0	0	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	[step, step, guide, invest, share, market, india]	[step, step, guide, invest, share, market]
1	1	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	What is the story of Kohinoor Koh-i-Noor Diamond	What would happen if the Indian government sto...	[story, kohinoor, koh, noor, diamond]	[happen, indian, government, steal, kohinoor, ...
2	2	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	[increase, speed, internet, connection, vpn]	[internet, speed, increase, hack, dns]

Generate batch works a little bit different because of the data set format but yields and saves the results same as above.

FinalLoadBatch.py

loading the saved batch from the batches folder.

```
class LoadBatch:
    def __init__(self,dir_name):
```

FinalModels.py

Run:

```
a = Models("embedding_lstm_model2","QQAbatch_400k_lstm_pre-1-2-06-2020")
model = a.model
print(model.summary())
model = a.start()
```

The first parameter is the model you want to run.

The second parameter is the name of the batch file at the “batches” directory. Loads the data using “LoadBatch” Object.

At the “start” function you can set the number of K at k-fold evaluation.

FinalGUI.py

```
def loadData(self):
    # self.data = LoadBatch('batch-1-07-05-2020')
    self.data = LoadBatch('QQAbatch_400k_lstm_pre-1-19-06-2020')
    self.tokenizer = self.data.tokenizer
    self.max_len_description = self.data.left_vector_max_size
    self.max_len_review = self.data.right_vector_max_size
    self.max_words = self.data.vocab_size
    self.df = self.data.convert_to_dataframe()
    # self.model = load_model('embedding_model2_07-05-2020.h5')
    self.model = load_model('400k_model_embedding_TEST-LIAD_3_QQAbatch-1-20-06-2020.h5')
    # self.model = load_model('400k_model_embedding_TEST-LIAD_3_QQAbatch-1-21-06-2020.h5')
```

First, you need to load the batch you used to train the model.

Second, you need to load the Neural Network model you have trained as h5 file.