README

Python Version Used:

Use Python3 for Data fetching and processing(except for **G15/CODE/preprocess_fetch_data/fetchwikiArticles.py**), use python2 everywhere else.

Procedure to fetch Data and process it:

1. **G15/DAT/mids.txt** consists of line separated mids(Freebase Id's) of entities.

2. To fetch entities present in this file , run **G15/CODE/preprocess_fetch_data/fetch_freebase_triples.py**
   This requires proxy settings to be modified.
   Inside "**G15/CODE/preprocess_fetch_data/fetch_freebase_triples.py**" modify "**http_proxy=http://USER:PASSWORD@PROXYIP:PORT**" ,"**https_proxy=https://USER:PASSWORD@PROXYIP:PORT**","**ftp_proxy=ftp://USER:PASSWORD@PROXYIP:PORT**" to your details

3. **G15/DATA/freebase_entity/** will contain raw entity corresponding to Id in **G15/DATA/mids.txt** .

4. **G15/DATA/freebase_entity_processed/** will contain processed entity corresponding to Id in **G15/DATA/mids.txt**.

5. Get wiki Links of all the entities , run **G15/CODE/preprocess_fetch_data/extract_wiki_links_from_entities.py**

6. Get raw wiki of all the entities ,run **G15/CODE/preprocess_fetch_data/fetchwikiArticles.py** -- use python2 for this

7. Process wikipedia articles of all the entities , run **G15/CODE/preprocess_fetch_data/process_wiki_files.py**

Procedure to run NeuralNetworkCode:

1. The **NeuralNetworkCode** folder contains 4 files.

2. The first file is **Preprocess.py**, to preprocess the data required for NTN model you should run **Preprocess.py** file and you can set various Train/Test percentage so that the model trains on the appropriate data based on the percentage you mentioned.

3. The **Preprocess.py** file writes set of possible entities in the whole data into entity folder, randomly picks triples form relation folder based on the percentage you have mentioned and dumps them into training folder.The remaining triples are dumped into testing file .

4. The **relationcount.txt** file contains the all relations and no of triples for that relation which is also created during preprocessing.

5.The **wikiData** folder contains the data extracted from wikipedia articles of that entity which will later be used for training the gensim model.

6. The Second file is **NTN.py** , this file trains the neural network and generates models for the relations you mention.The created models(relations) are stored in NNParameters folder and trained entityvectors in EntityVectors file.

7. The Third file is **Testing.py** which reads the testing file and scores the triples for the relation you mention.The final accuracy measures are displayed on the prompt.

8. The Fourth file is **HiddenRule.py** which takes the relations you mention and based on the hidden rule you define and finds accuracy measure accordingly.

NOTE : Please train before testing.

Result:
1.All entities are present in **G15/DATA/freebase_entity/**.
2.Entities with their facts are present in **G15/DATA/freebase_entity_processed/**.
3.The processed wikipedia articles for the entities with freebase Id's in mids.txt have been stored in **G15/DATA/wikiDataParsedText/** . Each file name is the entity name.
4. To check for more entities insert freebase people/person/ mids in mids.txt in line separated format and run all the above 5 steps again.
Freebase entities mids can be found on : http://www.freebase.com/people/person?instances= . Hover over a person's name and copy mid of the form '/m/******'.
5. The results of the NTN model are mentioned in the report.