RawData

```
1.riedel_train.json
2.riedel_test.json
3.rel2id.json (NYT 2010. Riedel, 2010. 53 types)
4.etype2id.json (FIGER. Xiao Ling and Daniel S. Weld, 2012. 112 types)
5.dep2id.json (Universal Stanford Dependencies v1. de Marneffe et al., 2014. 40 types Stanfo
```

RawData->Bags

INPUT:

1. riedel_train.json
2. riedel_test.json
3. rel2id.json

CODE:

make_bags.py

OUTPUT:

1. train_bags.json
2. test_bags.json
3. dep2id.json

Bags->Pkls

INPUT:

1. rel2id.json
2. dep2id.json
3. type2id.json
4. entity2typeid.json
5. train_bags.json
6. test_bags.json

CODE:

final_process.py

OUTPUT:

```
1. train.pkl
2. test.pkl
3. params.pkl
4. pn1.pkl
5. pn2.pkl
6. pn3.pkl
```

Exception: Entity not found in entity2id.json

1 取消注释 add entity 代码块，输出 addentity 集合 empty__entity.pkl

2 识别未知实体的类型

2.1 找到这些未知实体的名称，包含这些未知实体的句子，以及这些实体在句子里的位置

INPUT:

```
1.rel2id.json
2.dep2id.json
3.empty_entity.pkl
4.train_bags.json
5.test_bags.json
```

CODE:

```
moreentity.py
```

OUTPUT:

```
1.addtrainid2name.json
2.addtestid2name.json
3.addtrainsent.txt
4.addtestsent.txt
5.addempty_train_sent.json
6.addempty_test_sent.json
```

2.2 根据实体在句子里的位置对句子进行 BIO 标注（除了实体其它都是 O）

INPUT:

```
1.addempty_train_sent.json
2.addtest_train_sent.json
```

CODE:

```
create_seg.py
```

OUTPUT:

```
1.addtrain.seg
2.addtest.seg
```

2.3 利用FIGER 工具进行实体类型标注

INPUT:

```
1.addtrain.seg
2.addtrainsent.txt
3.addtest.seg
4.addtestsent.txt
```

CODE:

```
tag.sh
```

OUTPUT:

```
1.addtrainsent.out
2.addtestsent.out
```

2.4 补充标注好的实体到 entity2id.json

INPUT:

```
1.addtrainsent.out
2.addtestsent.out
3.empty_entity.pkl
4.addtrainid2name.json
5.addtestid2name.json
6.etype2id.json
7.entity2id.json
```

CODE:

```
addentityid.py
```

OUTPUT:

```
entity2id.json
```

Pkls->Mdbs

INPUT

```
1.train.pkl
2.test.pkl
3.pn1.pkl
4.pn2.pkl
5.pn3.pkl
```

CODE:

```
creatmdb.py
```

OUTPUT:

```
1.train.mdb
2.test.mdb
3.pn1.mdb
4.pn2.mdb
5.pn3.mdb
```