

courpus_V3.json 為 HSMM 模型所產出的模板。
(其他.json 為早期訓練出的版本，格式也皆相同)

詳細規格如下：

以 courpus_V3.json 為例
內共計 4560 個樣式模板



每個樣式的模板內有 1.references、2.state2words_references、
3.selected_template、4.predictions_result 四樣資訊

```
▼ 4 : {  
  ▶ 1.references : [ 10 items ]  
  ▶ 2.state2words_references : { 5 props }  
  3.selected_template : [ _時間A_ , 公司 ] 237, [ 余额 分别为 _時間A公司应交税费_] 140, [ 万元 、 _公司货币资金_ 万元 ] 80, [ 和 ] 45, [ <eos> ] 13,  
  4.predictions_result : _時間A_ , 公司余额分别为_時間A公司应交税费_万元、_公司货币资金_万元和<eos>  
}
```

1.references 為該樣式模板中，模型有參考原數據集的哪些句子，通常以此資訊來判定
模型是否有把相似句構、相似敘述內容的句子分群在一塊。

```
▼ 1.references : [ 10 items  
  0 : _時間A_ , 公司 主营业务收入 分别为 _時間A公司主营业务收入_ 万元 、 _公司主营业务收入_ 万元 、 <eos>  
  1 : _時間A_ , 发行人 营业收入 分别为 _時間A公司营业收入_ 万元 、 _公司营业收入_ 万元 、 <eos>  
  2 : _時間A_ , 发行人 营业成本 分别为 _時間A公司营业成本_ 万元 、 _公司营业成本_ 万元 、 <eos>  
  3 : _時間A_ , 公司 应交税费 余额 分别为 _時間A公司应交税费_ 万元 、 _公司应交税费_ 万元 、 <eos>  
  4 : _時間A_ , 公司 <unk> 余额 分别为 <unk> 万元 、 <unk> 万元 、 <eos>  
  5 : _時間A_ , 公司 货币资金 余额 分别为 _時間A公司货币资金_ 万元 、 _公司货币资金_ 万元 、 <eos>  
  6 : _時間A_ , 公司 应付票据 金额 分别为 _時間A公司应付票据_ 万元 、 <unk> 万元 和 <eos>  
  7 : _時間A_ , 公司 无形资产 净额 分别为 <unk> 万元 、 <unk> 万元 、 <eos>  
  8 : _時間A_ , 公司 境内 主营业务收入 分别为 _時間A公司主营业务收入_ 万元 、 _公司主营业务收入_ 万元 、 <eos>  
  9 : _時間A_ , 公司 应付账款 余额 分别为 _時間A公司应付账款_ 万元 、 _公司应付账款_ 万元 、 <eos>  
]
```

2.state2words_references 為模型將內些句子排列分成各個詞狀態，每個詞狀態也都
會將相似字詞、句構分群在同狀態下。以下圖為例，共有 237、140、80、45、13 這幾個
詞狀態。



3.selected_template 表示此樣式模板為：

詞狀態 237->詞狀態 140->詞狀態 80->詞狀態 45->詞狀態 13

此順序組成的

```
3.selected_template : [_時間A_ , 公司]237,[余额 分别为 _時間A公司应交税费_]140,[万元 、 _公司货币资金_ 万元]80,[和]45,[<eos>]13,
```

4.predictions_result 為從每個詞狀態挑選字詞後，即可組合成一段句子。

根據所選的字詞不同，可組合成不同敘述的句子。