Team3_Sub Project 1

- 組員：

學號	姓名
107065501	蘇玫如
107065527	段凱文

Tools

1. Stanford NLP

- **Stanford NLP** 是一個知名的自然語言處理套件，支援多國語言，包含繁體中文
- 使用 StanfordNLP 訓練好的繁體中文模型
    - Stanford CoreNLP
    - StanfordNLP Python Library
    - Stanford Parser
    - Stanford POS Tagger
    - Stanford Named Entity Recognizer
    - Stanford Word Segmenter
    - Stanford Dependency Parser
- 不過此套件本身是由Java寫成，Stanford 也有在Python提供一些套件，但都不怎麼精準。因此，我們這組在多下點功夫：**用 java 的 model 跑 python**
- **由於建立此環境十分複雜，以及stanfordNLP檔案過大以致無法上傳至ilms，故在作業中提供ipynb供老師及助教參考**
    - 安裝 Java
    - 安裝 Python Package

        - NLTK
        - stanfordcorenlp
        - stanfordnlp
    - 安裝 Stanford NLP Java
        - Java code
        - model

> 欲知安裝詳情或Tag Label，請見 Reference

2. Jieba-zh_TW

- 結巴(jieba)斷詞台灣繁體版本
- 進行句子有效斷詞
- 原理：採用和原始jieba相同的演算法，替換其詞庫及HMM機率表製做出針對台灣繁體的jieba斷詞器

Method

1. Character

- Step1. 收集文章內所有可能角色
- Step2. 找出故事中親人關係
- Step3. 找出故事中第三方角色（路人、壞人、跑龍套等等）
- Step4. 計算可能角色 Top 3
- Step5. 增加「可能角色」、移除字典提及的特殊案例
- Step6. 確實移除角色串列中的重複值
- **得出「人物列表」**

2. Dialogue

- Step1. 選取故事中出現對話的句子
- Step2. 了解文本中對話內容可能形式：
    - 角色先被描述，然後單純說一句話
        - 豬二哥想了想，說：「稻草屋會被風吹倒，那我就用木頭來蓋好了，木屋較重，不怕風吹。」

    - 一個句子包含兩句話跟敘述狀態
        - 很快的，野狼就追來了。他生氣的說：「死小豬，看我把你們的房子撞倒，你們就要被我吃掉了。」野狼說著說的開始撞牆了。「呀！～～」他使出全身力氣，向磚牆猛撞過去！
    - 句首直接講話
        - 只有一句對話
            - 「萬歲！」三隻小豬很高興的叫起來。從此，他們三個兄弟住在一起，每天一起吃飯睡覺，也一起工作，日子過得很快樂，而且野狼一直沒再出現呢！
        - 有兩句對話
            - 「啪啦！」一聲，磚牆沒被撞倒，野狼的骨頭卻斷了，「哎唷，痛死我了，痛死我了！」野狼哭哭啼啼的回家去了。
- Step3. 由於對話形式的不同，在抓取對話時，要做不同的處理，才能抓取正確
- Step4. 抓完對話後，還要知道說話人
- **得出（對話內容,說話人）**

3. Location

- Step1. 從在人物角色那邊抓取出來的 **NN標籤**，去查找自己預設的loc_default（童話故事中常發生的地點列表），並存在 loc_temp
- Step2. 從 loc_temp 整理
- **得出「地點列表」**

4. Event

- 從非對話裡的句子，去偵測人物角色（主體/受體）
- **得出（人物,描述）**

Demo

1. 三隻小豬

> 老師提供在 ilms 上的範例

- Character

- ['媽媽', '大哥', '二哥', '小弟', '野狼']

- Dialogue

('你們都長大了，應該自己蓋房子，自己住，自己種田，自己生活。我要你們自己照顧自己。', '媽媽')
('蓋一棟稻草屋吧，那最簡單了。', '大哥')
('我的房子蓋好了，很漂亮吧～，你們也快一點蓋吧！', '大哥')
('哇，果然很漂亮，大哥，你真了不起啊！', '二哥')
('哥，你不擔心稻草屋會被風吹倒嗎？', '小弟')
('稻草屋會被風吹倒，那我就用木頭來蓋好了，木屋較重，不怕風吹。', '二哥')
('咚咚咚，咚咚咚！', '二哥')
('你們看，這麼漂亮的房子！而且釘得很牢固，不會被風吹倒，我真的好棒呀！', '二哥')
('木屋雖然不會被風吹倒，可是用力打，木頭會被打斷，房子就垮了。', '小弟')
('你以為你最聰明？看你搬磚頭搬一整天了，房子還沒蓋出來。笨蛋～～', '二哥')
('我蓋房子雖然比你們慢，但是我要蓋的房子不怕風吹，也不怕敲打，紅磚房子最牢固了。', '小弟')
('幸好房子蓋好了，我不怕大野狼。', '小弟')
('你這個大笨蛋，這種稻草屋，我吹一口氣就倒了。', '大哥')
('木頭屋子一樣擋不住我，我一定要把你吃掉！', '大哥')
('砰～～', '野狼')
('快把門鎖緊，不然我們會被吃掉的。', '大哥')
('呀！～～', '野狼')
('哎唷，痛死我了，痛死我了！', '野狼')
('萬歲！', '野狼')

- Location

  - ['房子', '稻草屋', '家', '木屋子', '磚屋']

- Event

('媽媽', '從前，有一隻胖胖的豬媽媽，她生了三隻小豬')
('大哥', '最大的小豬：豬大哥很貪睡，很懶惰')
('二哥', '第二個小豬：豬二哥很愛吃，他也很懶惰')
('小弟', '幸好最小的小豬：豬小弟是個勤勞的好孩子')
('小弟', '豬小弟不理會哥哥的嘲笑，他搬好了磚塊，又搬水泥，他把水泥和好了，開始堆砌磚塊，一塊一塊
('小弟', '豬小弟心理想：你們不要笑我，等我蓋好了，你們就知道了，我的房子比什麼都堅固，野狼來了我
('野狼', '豬小弟心理想：你們不要笑我，等我蓋好了，你們就知道了，我的房子比什麼都堅固，野狼來了我
('小弟', '豬小弟繼續加油工作，他趕呀趕，趕到天黑，月亮掛在天上了，他的紅磚房子才好不容易蓋好了'
('小弟', '豬小弟正想休息，卻聽到了大野狼的吼叫聲')
('野狼', '豬小弟正想休息，卻聽到了大野狼的吼叫聲')
('野狼', '這時兩個豬哥哥也聽到了野狼的吼叫聲，怕得發抖，他們怕野狼來，所以整夜都不敢安心睡覺')
('媽媽', '隔天，豬媽媽要三隻小豬到田裡工作，沒想到走到半路的時候，他們被一隻可怕的大野狼發現了，
('野狼', '隔天，豬媽媽要三隻小豬到田裡工作，沒想到走到半路的時候，他們被一隻可怕的大野狼發現了，
('大哥', '野狼決定先吃豬大哥')
('野狼', '野狼決定先吃豬大哥')

2. 小紅帽

我們這組在網路上找的範例

- Character
  - [‘奶奶’, ‘媽媽’, ‘獵人’, ‘小紅帽’, ‘狼’]
- Dialogue

("來，小紅帽，這裡有一塊蛋糕和一瓶葡萄酒，快給奶奶送去，奶奶生病了，身子很虛弱，吃了這些就會好一
('我會小心的。', '媽媽')
('你好，小紅帽', '狼')
('謝謝你，狼先生。', '小紅帽')
('小紅帽，這麼早要到哪裡去呀？', '狼')
('我要到奶奶家去。', '小紅帽')
('你那圍裙下面有什麼呀？', '狼')
('蛋糕和葡萄酒。昨天我們家烤了一些蛋糕，可憐的奶奶生了病，要吃一些好東西才能恢復過來。', '小紅帽
('你奶奶住在哪裡呀，小紅帽？', '狼')
('進了林子還有一段路呢。她的房子就在三棵大橡樹下，低處圍著核桃樹籬笆。你一定知道的。', '小紅帽'
('這小東西細皮嫩肉的，味道肯定比那老太婆要好。我要講究一下策略，讓她倆都逃不出我的手心。', '狼'
('小紅帽，你看周圍這些花多麼美麗啊！幹嗎不回頭看一看呢？還有這些小鳥，它們唱得多麼動聽啊！你大概
('也許我該摘一把鮮花給奶奶，讓她高興高興。現在天色還早，我不會去遲的。', '小紅帽')
('是誰呀？', '奶奶')
('是小紅帽，我給你送蛋糕和葡萄酒來了。快開門哪。', '狼')
('我身上沒有力氣，起不來。', '奶奶')
('天哪！平常我那麼喜歡來奶奶家，今天怎麼這樣害怕？', '小紅帽')
('早上好！', '小紅帽')
('哎，奶奶，你的耳朵怎麼這樣大呀？', '小紅帽')
('為了更好地聽你說話呀，乖乖。', '狼')
('可是奶奶，你的眼睛怎麼這樣大呀？', '小紅帽')
('為了更清楚地看你呀，乖乖。', '狼')
('奶奶，你的手怎麼這樣大呀？', '小紅帽')
('可以更好地抱著你呀。', '狼')
('奶奶，你的嘴巴怎麼大得很嚇人呀？', '小紅帽')
('可以一口把你吃掉呀！', '狼')
('這老太太鼾打得好響啊！我要進去看看她是不是出什麼事了。', '獵人')
('你這老壞蛋，我找了你這麼久，真沒想到在這裡找到你！', '獵人')
('真把我嚇壞了！狼肚子里黑漆漆的。', '小紅帽')
('要是媽媽不允許，我一輩子也不獨自離開大路，跑進森林了。', '小紅帽')
('我們把門關緊，不讓它進來。', '奶奶')
('奶奶，快開門呀。我是小紅帽，給你送蛋糕來了。', '狼')
('小紅帽，把桶拿來。我昨天做了一些香腸，提些煮香腸的水去倒進石頭槽里。', '狼')

- Location
  - [‘奶奶家’, ‘家’, ‘森林’, ‘房子’]
- Event

('奶奶', '從前有個可愛的小姑娘，誰見了都喜歡，但最喜歡她的是她的奶奶，簡直是她要什麼就給她什麼'
('奶奶', ' 一次，奶奶送給小姑娘一頂用絲絨做的小紅帽，戴在她的頭上正好合適')
('小紅帽', ' 一次，奶奶送給小姑娘一頂用絲絨做的小紅帽，戴在她的頭上正好合適')
('小紅帽', '" 從此，姑娘再也不願意戴任何別的帽子，於是大家便叫她'小紅帽'")
('奶奶', '奶奶住在村子外面的森林里，離小紅帽家有很長一段路')
('小紅帽', '奶奶住在村子外面的森林里，離小紅帽家有很長一段路')
('小紅帽', ' 小紅帽剛走進森林就碰到了一條狼')
('狼', ' 小紅帽剛走進森林就碰到了一條狼')
('小紅帽', ' 小紅帽不知道狼是壞家伙，所以一點也不怕它')
('狼', ' 小紅帽不知道狼是壞家伙，所以一點也不怕它')
('奶奶', '就在此時，狼卻直接跑到奶奶家，敲了敲門')
('狼', '就在此時，狼卻直接跑到奶奶家，敲了敲門')
('狼', '狼剛拉起門栓，那門就開了')
('奶奶', ' 狼二話沒說就衝到奶奶的床前，把奶奶吞進了肚子')
('狼', ' 狼二話沒說就衝到奶奶的床前，把奶奶吞進了肚子')
('奶奶', ' 然後她穿上奶奶的衣服，戴上她的帽子，躺在床上，還拉上了簾子')
('小紅帽', '可這時小紅帽還在跑來跑去地採花')
('奶奶', ' 直到採了許多許多，她都拿不了啦，她才想起奶奶，重新上路去奶奶家')
('奶奶', '看到奶奶家的屋門敞開著，小紅帽感到很奇怪')
('小紅帽', '看到奶奶家的屋門敞開著，小紅帽感到很奇怪')
('小紅帽', '狼剛把話說完，就從床上跳起來，把小紅帽吞進了肚子，狼滿足了食欲之後便重新躺到床上睡覺
('狼', '狼剛把話說完，就從床上跳起來，把小紅帽吞進了肚子，狼滿足了食欲之後便重新躺到床上睡覺，而
('奶奶', '他正準備向狼開槍，突然又想到，這狼很可能把奶奶吞進了肚子，奶奶也許還活著')
('狼', '他正準備向狼開槍，突然又想到，這狼很可能把奶奶吞進了肚子，奶奶也許還活著')
('獵人', '獵人就沒有開槍，而是操起一把剪刀，動手把呼呼大睡的狼的肚子剪了開來')
('狼', '獵人就沒有開槍，而是操起一把剪刀，動手把呼呼大睡的狼的肚子剪了開來')
('奶奶', '接著，奶奶也活著出來了，只是有點喘不過氣來')
('小紅帽', ' 小紅帽趕緊跑去搬來幾塊大石頭，塞進狼的肚子')
('狼', ' 小紅帽趕緊跑去搬來幾塊大石頭，塞進狼的肚子')
('狼', ' 狼醒來之後想逃走，可是那些石頭太重了，它剛站起來就跌到在地，摔死了')
('奶奶', ' 獵人剝下狼皮，回家去了；奶奶吃了小紅帽帶來的蛋糕和葡萄酒，精神好多了\n')
('獵人', ' 獵人剝下狼皮，回家去了；奶奶吃了小紅帽帶來的蛋糕和葡萄酒，精神好多了\n')
('小紅帽', ' 獵人剝下狼皮，回家去了；奶奶吃了小紅帽帶來的蛋糕和葡萄酒，精神好多了\n')
('狼', ' 獵人剝下狼皮，回家去了；奶奶吃了小紅帽帶來的蛋糕和葡萄酒，精神好多了\n')
('奶奶', '人們還說，小紅帽後來又有一次把蛋糕送給奶奶，而且在路上又有一隻狼跟她搭話，想騙她離開大
('小紅帽', '人們還說，小紅帽後來又有一次把蛋糕送給奶奶，而且在路上又有一隻狼跟她搭話，想騙她離開
('狼', '人們還說，小紅帽後來又有一次把蛋糕送給奶奶，而且在路上又有一隻狼跟她搭話，想騙她離開大路
('小紅帽', ' 可小紅帽這次提高了警惕，頭也不回地向前走')
('奶奶', '她告訴奶奶她碰到了狼，那家伙嘴上雖然對她說"你好"，眼睛里卻露著凶光，要不是在大路上，它
('狼', '她告訴奶奶她碰到了狼，那家伙嘴上雖然對她說"你好"，眼睛里卻露著凶光，要不是在大路上，它
('小紅帽', ' 這長著灰毛的家伙圍著房子轉了兩三圈，最後跳上屋頂，打算等小紅帽在傍晚回家時偷偷跟在
('奶奶', '可奶奶看穿了這家伙的壞心思')
('小紅帽', '小紅帽提了很多很多水，把那個大石頭槽子裝得滿滿的')
('狼', ' 香腸的氣味飄進了狼的鼻孔，它使勁地用鼻子聞呀聞，並且朝下張望著，到最後把脖子伸得太長了
('小紅帽', ' 小紅帽高高興興地回了家，從此再也沒有誰傷害過她')

Evaluation

1. 三隻小豬

Accuracy	Precision	Recall	F1-score
90%	97%	88%	92%

2. 小紅帽

Accuracy	Precision	Recall	F1-score
94%	93%	93%	93%

- 以這兩篇文本來說，四個評估值相當不錯，都達到九成的水準

Conclusion

- 人物、地點
  - 準確率高
  - 但存在一個問題
    - 某些詞在故事中是同義詞，但會被辨識為相異物件
- 對話
  - 準確率高
  - 狀聲詞會被誤以為是對話
  - 說話人辨識有時會被代名詞搞錯
  - 由於對話句子間的敘述，應該要記為「事件」，需要對此微調
- 事件
  - 準確率高
  - 主詞受詞辨識有時會顛倒、搞亂、受代名詞影響
  - 由於對話句子間的敘述，應該要記為「事件」，但會被忽略，需在「對話」處理時注意
- 以上是這次作業的簡單報告，而這次作業發生的缺點，希望能在 sub-project 2 進行改進

Reference

- Stanford

    - Stanford NLP Software (https://nlp.stanford.edu/software/)
    - StanfordNLP 安裝教學 1 (https://medium.com/@peilee_98185/python%E8%8B%B1%E6%96%87%E8%87%AA%E7%84%B6%E8%AA%9E%E8%A8%80%E8%99%95%E7%90%86-stanford-nlp%E5%AE%89%E8%A3%9D%E5%8F%8A%E6%B8%AC%E8%A9%A6-33818082384c)
    - StanfordNLP 安裝教學 2 (http://www.zmonster.me/2016/06/08/use-stanford-nlp-package-in-nltk.html)
    - StanfordNLP 安裝教學 3 (http://www.cnblogs.com/baiboy/p/nltk1.html)
    - Definitions of the Stanford typed dependencies (https://nlp.stanford.edu/software/dependencies_manual.pdf)
    - StanfordNLP POSTagger Label (https://www.cnblogs.com/tonglin0325/p/6850901.html)
- Jieba-zh_TW (https://github.com/ldkrsi/jieba-zh_TW)