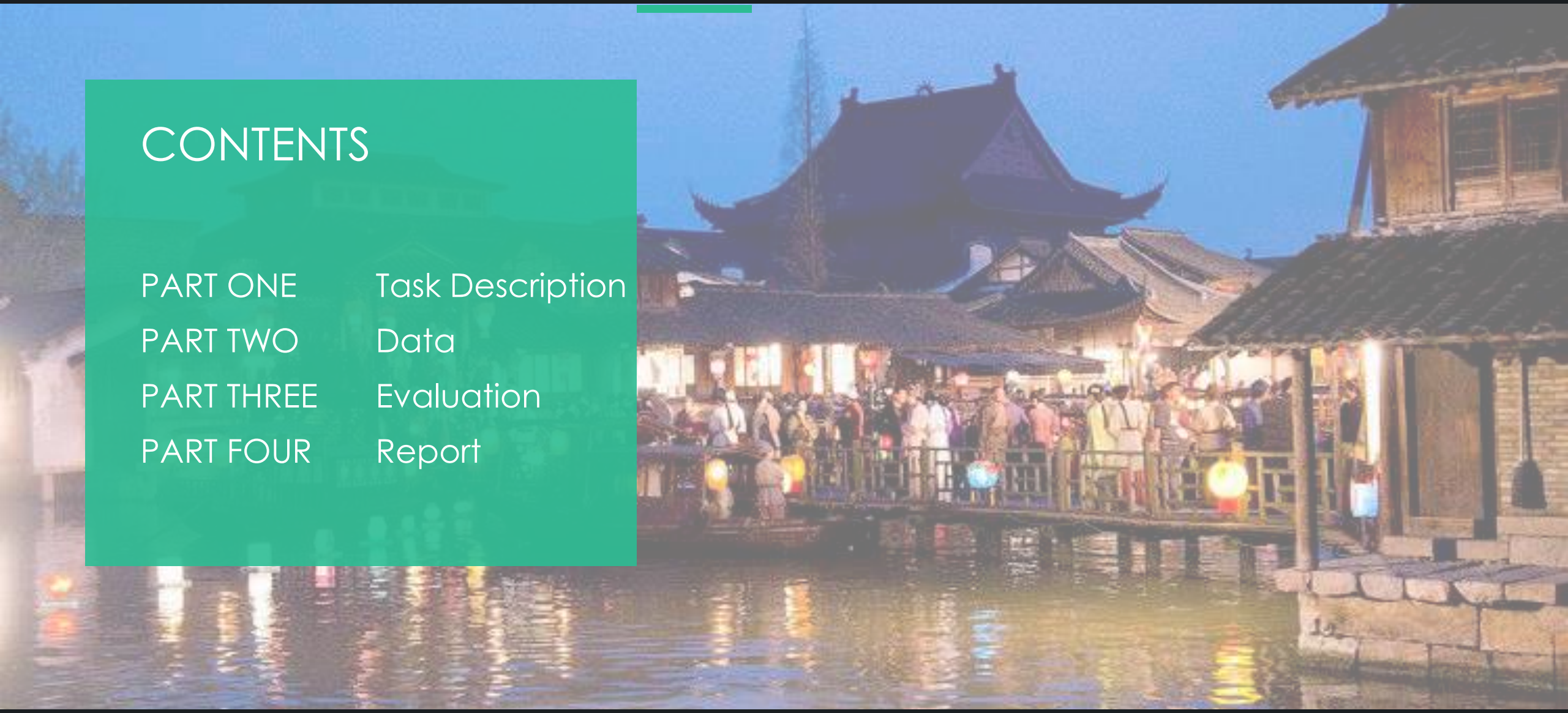
2017

Information Retrieval and Extraction
Project 2

NTU CSIE, Fall 2017 陳信希 教授
顏安孜 陳重吉 {d04922005, d05922016}@ntu.edu.tw



CONTENTS

Relation Extraction

Given Dream of the Red Chamber and a pair of entities, can you infer the relationship between the two entities?

- Resource: 紅樓夢
- 12 Relations:
  祖孫/ 母子/ 母女/ 父子/ 父女/ 兄弟姊妹/ 夫妻/ 姑叔舅姨甥侄/ 遠親/ 主僕/ 師徒/ 居處
- Ex: (賈代化,賈敷) = 父子

賈代化襲了官，也養了兩個兒子：長名賈敷，至八九歲上便死了，只剩了次子賈敬襲了官，如今一味好道，只愛燒丹煉汞，餘者一概不在心上。

Data Format

Dream_of_the_Red_Chamber.txt

Novel download from:
http://www.speedy7.com/cn/stguru/big5/redmansions.htm

train.txt & test.txt

ID	Entity1	Entity2	Relation

Separated by Tab

Ex:

1	賈演	賈源	兄弟姊妹
2	賈源	賈寶玉	祖孫

Dream_of_the_Red_Chamber_seg.txt

Novel already segment and has POS tag

Ex:
此_Nh 開_VHC 卷_Na 第_Nes 一_Neu 回_Nf 也_T 。_Po 作_VC 者_T8 自_Dh 云_VE

Download from:
http://lingcorpus.iis.sinica.edu.tw/kiwi/pkiwi/index.html

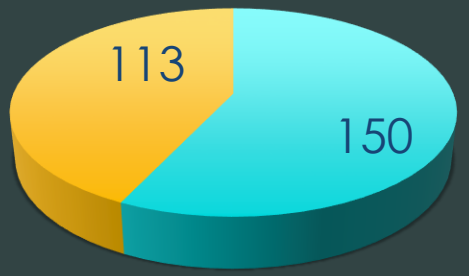
3

Evaluation

Accuracy = (# pairs correctly answered) / (# test pairs)

The accuracy would be 33.63% for a classifier that always predicted the majority class.

Data



Report

Report

- Language: Chinese or English (Be clear in meaning!)
- Pages: no more than 6 (with readable font size)
- Format: PDF
- Must include:
  - Name and student ID of each group member
  - Agree to share your report with your classmates? (YES/NO)
    - Will appear in 作業觀摩
  - Methodology
  - Experiments
  - Discussions

Code

- Describe your code
  - write the proper comment for each part and function

Uncommented Code

```
city=raw_input("Enter a city: ")
while city[-1]==" ":
    city = city[:-1]
temp=raw_input("Enter a temperature in Farenheit: ")
temp = float(temp)
temp = (temp - 32.0)*(100.0/180.0)
temp = round(temp,3)
temp = str(temp)
print "In "+city+" it is "+temp+" degrees Celcius!"
```

Commented Code

```
#Alyssa P. Hacker
#fah_to_celsius.py

#collect a city name from user
city=raw_input("Enter a city: ")

#trunacte whitespace
while city[-1]==" ":
    city = city[:-1]

#collect a temp from user
temp=raw_input("Enter a temperature in Farenheit: ")

#convert string to float
temp = float(temp)

#convert Farenheit temp to Celsius temp
temp = (temp - 32.0)*(100.0/180.0)

#truncate to 3 decimal places
temp = round(temp,3)

#recast as string so we can concatenate
temp = str(temp)

#print result!
print "In "+city+" it is "+temp+" degrees Celcius!"
```

Format

- Project2_team_<team number>.zip
  - Report_team_<team number>.pdf
  - Code_team_<team number>
    - readme.txt (description of each script)
    - $script_1$
    - …
    - $script_n$

Ex:

- Project2_team_0.zip
  - Report_team_0.pdf
  - Code_team_0
    - readme.txt
    - main.py

- Deadline: **2018/01/05 23:59**
- Upload to CEIBA
  - One submission per group
- Grading:
  - Performance: 30%
  - Report: 70%

Possible Directions

- As a pattern extraction problem
  - Consider the order of the two entities in sentences
  - Consider syntactic structure of the sentences
  - Select words/phrases that discriminate specific property from others
  - …
- As a classification problem
  - 12 properties → 12 classes
  - Features from sentence instances
  - Features from entity itself (type=Person/Place, known relations with others, …)
  - …

Baseline Algorithm – 45.53%

```
for each entity pair:
    for each sentence:
        if two entity in sentence:
            extract this sentence
            break
        elif two entity in context:# context = three sentences
            extract context
            break
        else:
            extract two sentences that two entities first occur
            break
```

After extracting sentences, we sum up word vectors by equal weight as features of entities pairs, and classify entities pairs into 12 relations by Random Forest classifier.


Have Fun