LING 388: Computers and Language

Lecture 13
Adminstrivia

• Homework 5 review
• New Homework 6 on Python regex
Python regex

• Summary (so far):
  • \w a character [A-Za-z0-9_]
  • \d [0-9]
  • \b word boundary
  • \s space character [ \t\n\r\f\v]

• Operators:
  • * zero or more repeats
  • + one or more repeats
  • ( ) grouping

• Raw string (avoid escaping \\):
  • r"\w+"

• Negation:
  • \W anything not in \w
  • \D anything not in \d

• Anchors:
  • ^ beginning of the line
  • $ end of the line

• Backreferences:
  • \n n is group number

• Full Documentation:
  https://docs.python.org/3/library/re.html
Python regex

- **Methods:**
  - **Key:** RE = regex raw string, String = where to search
  - `re.match(RE, String)` matching must start from start of String
  - `re.search(RE, String)` searches anywhere in String
  - `re.findall(RE, String)`
  - `re.finditer(RE, String)` use with loop for `m in re.finditer()`
  - `re.sub(RE, SUB, String)` SUB = regex raw string to substitute for RE
Review examples

Backreferences and grouping

```python
>>> re.search(r"b\(w\)+s+(\w+)s+and\s+2\s+s+1\b", "I know Elton John and Elton John")
<_sre.SRE_Match object; span=(7, 32), match='Elton John and John Elton'>
>>> re.search(r"b\(w\)+s+(\w+)s+and\s+2\s+s+1\b", "I know Elton John and John Elton")
<_sre.SRE_Match object; span=(9, 16), match='the the'>
>>> re.search(r"b\(w\)+s+(\w+)s+1\b", "Paris in the the Spring")
group()
'the the'
>>> re.search(r"b\(w\)+s+(\w+)s+1\b", "Paris in the the Spring").group(0)
'the the'
>>> re.search(r"b\(w\)+s+(\w+)s+1\b", "Paris in the the Spring").group(1)
'the'
>>> re.search(r"b\(w\)+s+(\w+)s+1\b", "Paris in the the Spring").group(2)

```
Review examples

Substitution using `.sub()` with backreferences and grouping:

```python
>>> p = re.compile("section{([^}]*)}\)"
>>> p.match(text).group(1)
'one'
>>> p.sub(r"subsection\{\1\",text)
'subsection{one}'
```

[^}\ means any character but }
Review examples

Using ^ and $ meta-characters:

```python
>>> import re
>>> text = "abc-def"
>>> re.match(r"\w+", text)
<_sre.SRE_Match object; span=(0, 3), match='abc'>
>>> re.match(r"[\w-]+", text)
<_sre.SRE_Match object; span=(0, 7), match='abc-def'>
>>> re.match(r"^\w+", text)
<_sre.SRE_Match object; span=(0, 3), match='abc'>
>>> re.match(r"\w+$", text)
>>> re.match(r"^\w+$", text)
>>> re.search(r"\w+$", text)
<_sre.SRE_Match object; span=(4, 7), match='def'>
```
Homework 5 Review

- NER program:
  - [http://cogcomp.org/page/demo_view/NERextended](http://cogcomp.org/page/demo_view/NERextended)

---

1. Russian curler Alexander Krushelnitsky has been stripped of his Winter Olympics bronze medal after being found guilty of doping.
2. The 25-year-old was one of 168 Russians allowed to compete as neutrals at the Games despite the country being banned for previous doping offences.
4. The Court of Arbitration for Sport said he admitted the anti-doping violation.
5. With Krushelnitsky and Bryzgalova’s third place annulled, when the bronze medal is officially reallocated it will go to Norway.
6. The International Olympic Committee banned Russia from this month’s Games in Pyeongchang, South Korea over “systemic” doping at Sochi 2014, which Russia hosted.
7. But athletes who could prove they were clean were allowed to compete as Olympic athletes from Russia (OAR). The team of 168 was the third biggest behind Canada and the United States.
8. On Tuesday, the OAR team said a criminal investigation had been opened into Krushelnitsky's positive test.
9. It said there was “no evidence of the systematic usage of meldonium”, and that he returned a clear sample on 22 January, when the team’s curlers were tested before flying out to South Korea.
The Illinois Extended Named Entity Recognizer has identified the following named entities.

[NORP Russian] curler [PERSON Alexander Krushelnitsky] has been stripped of his Winter Olympics bronze medal after being found guilty of doping.
The 25-year-old was [CARDINAL one] of [CARDINAL 168] [NORP Russians] allowed to compete as neutrals at the Games despite the country being banned for previous doping offences.
[PERSON Krushelnitsky], who won mixed-doubles bronze with wife [PERSON Anastasia Bryzgalova], tested positive for meldonium.
[ORG The Court of Arbitration for Sport] said he admitted the anti-doping violation.
With [PERSON Krushelnitsky] and [PERSON Bryzgalova]'s [ORDINAL third] place annulled, when the bronze medal is officially reallocated it will go to [GPE Norway].
[ORG The International Olympic Committee] banned [GPE Russia] from [DATE this month's] Games in [GPE Pyeongchang], [GPE South Korea] over “systemic” doping at [GPE Sochi] [DATE 2014], which [GPE Russia] hosted.
But athletes who could prove they were clean were allowed to compete as Olympic athletes from [GPE Russia] (OAR).
The team of [CARDINAL 168] was the [ORDINAL third] biggest behind [GPE Canada] and [GPE the United States].
On [DATE Tuesday], the OAR team said a criminal investigation had been opened into [ORG Krushelnitsky]'s positive test.
It said there was “no evidence of the systematic usage of meldonium”, and that he returned a clear sample on [DATE 22 January], when the team’s curlers were tested be
Homework 5 Review

**Missing/Wrong** (in green: optional for this homework – needed for anaphora resolution)

- Line 1:
  - Winter Olympics (Event)
  - his -> he (Person)

- Line 2:
  - The 25 year-old (Person)
  - the Games (Event)
  - the country (GPE)

- Line 4:
  - he (Person)

More advanced:
- curler (Person?)
- neutrals
- wife
- who
Homework 5 Review

**Missing/Wrong** (in green: optional for this homework – needed for anaphora resolution)

With [PERSON Krushelnitsky] and [PERSON Bryzgalova]'s [ORDINAL third] place annulled, when the bronze medal is officially reallocated it will go to [GPE Norway].

[ORG The International Olympic Committee] banned [GPE Russia] from [DATE this month's] Games in [GPE Pyeongchang], [GPE South Korea] over "systemic" doping at [GPE Sochi] [DATE 2014], which [GPE Russia] hosted.

But athletes who could prove they were clean were allowed to compete as Olympic athletes from [GPE Russia] (OAR). The team of [CARDINAL 168] was the [ORDINAL third] biggest behind [GPE Canada] and [GPE the United States].

• Line 5:
  • it (Product?)

• Line 6:
  • Games (Event)
  • Pyeongchang (LOCATION)

• Line 7:
  • they (Person)
  • OAR (ORG/GPE?)

More advanced:
• athletes < Olympic athletes < {Krushelnitsky, Bryzgalova}
• team < OAR
Homework 5 Review

**Missing/Wrong** (in green: optional for this homework – needed for anaphora resolution)

On **[DATE Tuesday]**, the OAR team said a criminal investigation had been opened into **[ORG Krushelnitsky]**'s positive test. It said there was "no evidence of the systematic usage of meldonium", and that he returned a clear sample on **[DATE 22 January]**, when the team's curlers were tested be

• Line 8:
  • OAR (ORG/GPE?)
  • Krushelnitsky (Person NOT ORG)

• Line 9:
  • it (ORG/GPE?)
  • he (Person) the Games (Event)

More advanced:

• the team's curlers (Person)
Homework 6

• Recall file I/O Python code? (see Lecture 9)
• Template:

```python
# assume argv[1] will be input filename
import sys
import re
if len(sys.argv) != 2:
    print('Error: usage infile')
    sys.exit(1)

infile = open(sys.argv[1], "r")
for line in infile:
    # do something to line
infile.close()
```

hw6-template.py

replace this with your code!
Homework 6

File: hw6.txt

(Full text of the BBC News article used in Homework 5 as processed by the U. of Illinois Extended NER demo)
Homework 6

• Question 1:
  • Using your Python regex coding skills, modify the template code to parse hw6.txt to create Counters for categories ORG, PERSON and GPE.
  • Have your code print the Counters after reading the file.
  • Recall Counter? (See Lecture 7)

• Sample output:

```bash
Sandyways-MacBook:ling388-18 sandiway$ python3 hw6-q1.py hw6.txt

ORG:
[['the World Anti-Doping Agency', 1], ['The International Olympic Committee', 1], ['The Court of Arbitration for Sport', 1], ['Krushelnitsky', 1]]

PERSON:
[['Krushelnitsky', 3], ['Anastasia Bryzgalova', 1], ['Alexander Krushelnitsky', 1], ['Bryzgalova', 1]]

GPE:
[['Russia', 3], ['South Korea', 2], ['Norway', 1], ['the United States', 1], ['Pyeongchang', 1], ['Sochi', 1], ['Canada', 1]]

Sandyways-MacBook:ling388-18 sandiway$
```
Homework 6

• Template for reading and writing:

```python
# assume argv[1] will be input filename
# assume argv[2] will be output filename
import sys
import re
if len(sys.argv) != 3:
    print('Error: usage infile outfile')
    sys.exit(1)

infile = open(sys.argv[1], "r")
outfile = open(sys.argv[2], "w")

for line in infile:
    outfile.write(line)
infile.close()
outfile.close()
```

hw6-template2.py

copies each line from infile to outfile
Homework 6

Question 2:

"OAR just so happens to be a country you've already heard of ... in disguise. It's the Russians!"

• Using .sub(), modify the template code to parse hw6.txt and create a modified file hw6new.txt where abbreviations like OAR are labeled as GPEs.

• Sample from hw6new.txt:

   But athletes who could prove they were clean were allowed to compete as Olympic athletes from [GPE Russia] ([GPE OAR]). The team of [CARDINAL 168] was the [ORDINAL third] biggest behind [GPE Canada] and [GPE the United States].

   On [DATE Tuesday] , the [GPE OAR] team said a criminal investigation had been opened into [ORG Krushelnitsky]'s positive test.
Homework 6

• Question 2 (continued):
  • Now re-run your code from Question 1 on the modified hw6new.txt file.
  • Show your output.

• Sample:

```
$ python3 hw6-q1.py hw6new.txt
ORG:
[('Krushelnitsky', 1), ('The Court of Arbitration for Sport', 1), ('Doping Agency', 1), ('The International Olympic Committee', 1)]
PERSON:
[('Krushelnitsky', 3), ('Alexander Krushelnitsky', 1), ('Bryzgalova\nia Bryzgalova', 1)]
GPE:
[('Russia', 3), ('OAR', 2), ('South Korea', 2), ('Sochi', 1), ('Pyeong\n'Canada', 1), ('the United States', 1), ('Norway', 1)]
```
Homework 6

• Question 3:
  • Along the lines of Question 2, further modify your program to also mark all pronouns and possessive pronouns as PERSON.
  • Show your latest hw6new.txt file.
  • Re-run your code from Question 1. Show your new output.

• Sample:
Homework 6

• Spring break next week
• Due by Monday March 12th by midnight
• Usual rules:
  • one PDF file to TA Patricia Lee
  • submit all code and sample output in the PDF file