Python per la Manipolazione dei Testi

Cezar Sas
University of Groningen
c.a.sas@rug.nl

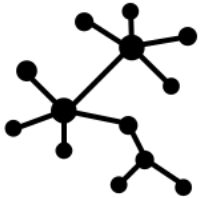
About Me

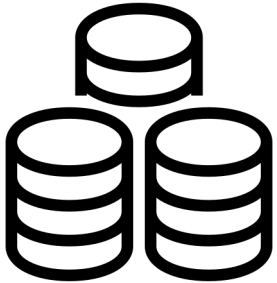

UNIVERSITA' DEGLI STUDI DI MILANO
BICOCCA


UNIVERSITY OF
COPENHAGEN


university of
groningen

About Me





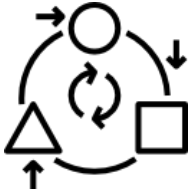
Python
per la *(mia)* ricerca

Environment e IDE

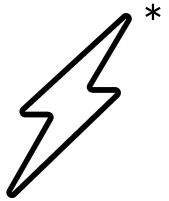
[1] Top 10 Companies using Python in Backend

Perchè Python?

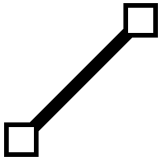
Prototipazione Rapida



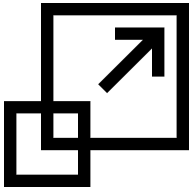
Flessibile



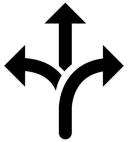
Veloce*



Facile e Leggibile



Scalabile[1]



Multi Paradigma

Perchè NO?

- Alto rischio di Spaghetti Code
- Problemi con il management dei package
- Poco leggibile (in riduzione con l'introduzione di `typing`)

Estrazione Dati Dal Testo

(Text Mining)

Estrazione Informazioni

Giuseppe Garibaldi (Nizza, 4 luglio 1807 – La Maddalena, 2 giugno 1882) è stato un generale, patriota, condottiero e scrittore italiano. Figura rilevante del Risorgimento, fu uno dei personaggi storici più celebrati della sua epoca. È noto anche con l'appellativo di «eroe dei due mondi» per le imprese militari compiute sia in Europa, sia in America Meridionale.

Estrazione Informazioni

Giuseppe Garibaldi (Nizza, 4 luglio 1807 – La Maddalena, 2 giugno 1882) è stato un generale, patriota, condottiero e scrittore italiano. Figura rilevante del Risorgimento, fu uno dei personaggi storici più celebrati della sua epoca. È noto anche con l'appellativo di «eroe dei due mondi» per le imprese militari compiute sia in Europa, sia in America Meridionale.

Estrazione Informazioni

Giuseppe Garibaldi (Nizza, 4 luglio 1807 – La Maddalena, 2 giugno 1882) è stato un **generale, patriota, condottiero e scrittore** italiano. Figura rilevante del Risorgimento, fu uno dei personaggi storici più celebrati della sua epoca. È noto anche con l'appellativo di «eroe dei due mondi» per le imprese militari compiute sia in Europa, sia in America Meridionale.

Pipeline Analisi Testo

Espressioni Regolari

Definizione

Un espressione regolare **R**, definita su un alfabeto **Σ** e un insieme di operazioni {∪, ∩, ∘, *, -}, è definita come una stringa **R** tale che una delle seguenti sia vera:

- **R = ∅, R ∈ Σ**
- **R = S ∪ T - UNIONE**
- **R = S ∩ T - INTERSEZIONI**
- **R = S ∘ T - CONCATENAZIONE**
- **R = S* - STELLA DI KLEEN**
- **R = Σ* - S - DIFFERENZA**

con **S, T** espressioni regolari su **Σ**

Definizione

Un espressione regolare **R**, definita su un alfabeto **Σ** e un insie
{∪, ∩, ∘, *, -}, è definita come una stringa **R** tale che una delle

- **R = Ø, R ∈ Σ**
- **R = S ∪ T - UNIONE**
- **R = S ∩ T - INTERSEZIONI**
- **R = S ∘ T - CONCATENAZIONE**
- **R = S* - STELLA DI KLEEN**
- **R = Σ* - S - DIFFERENZA**

con **S, T** espressioni regolari su **Σ**

Operazioni

Caratteri Speciali

- **.** (punto) - Qualsiasi carattere escluso **`\n`**
- \ (backslash) - Escape di caratteri speciali - e.g **\\n** non e piu newline
- | (pipe) - Oppure - e.g. **A | B**
- * **+ ?** (Quantificatori)
- **{min, max}** (Quantificatori)

Sequenza

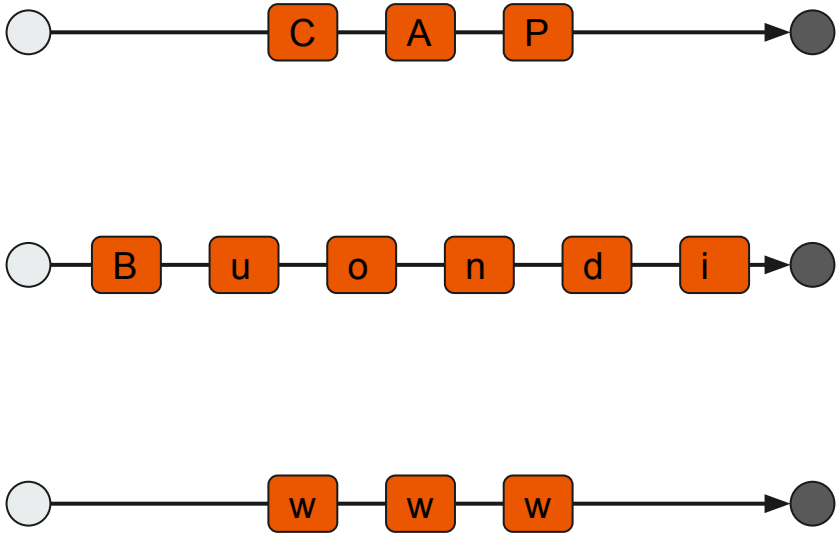
Match di tutta la sequenza

Sintassi:

sequenza

Esempi:

CAP: ⟶ CAP: 20010
Buondi ⟶ Buondi Si
www ⟶ awww



Gruppo

Match di quello che matcha il contenuto

Sintassi:

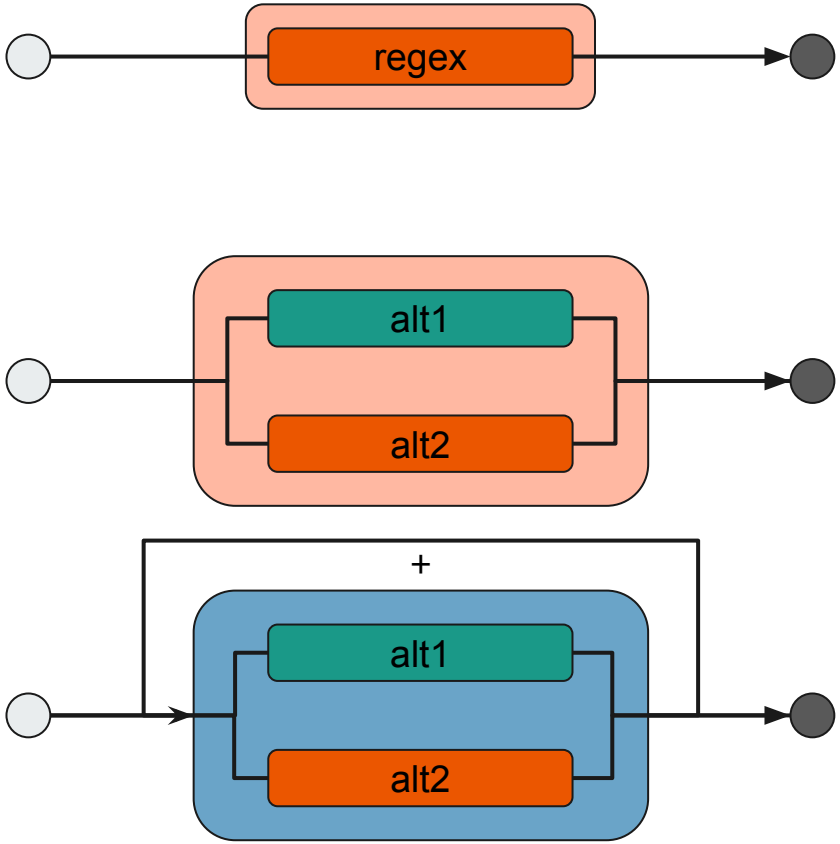
(regex)

Esempi:

(alt1 | alt2) ⟶ **kalt2alt1**

(alt1 | alt2)+ ⟶ **kalt1alt2**

(reg){2,} ⟶ **regreg**



Insieme

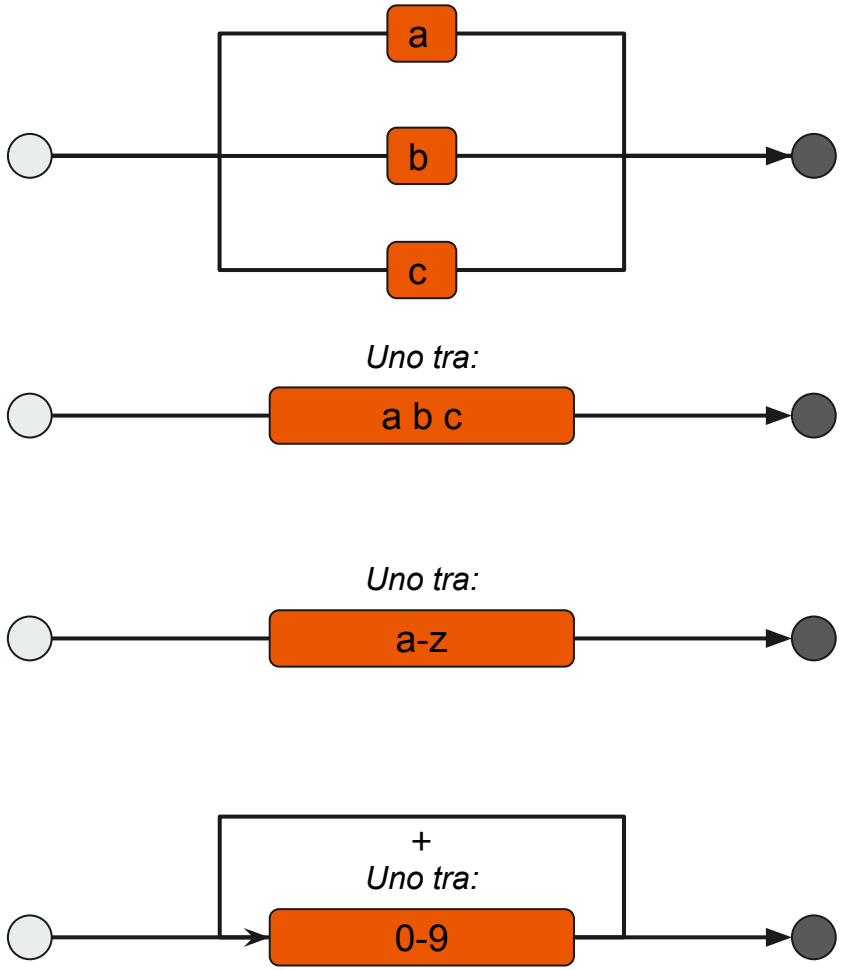
Match di un carattere all'interno del set

Sintassi:

[regex]

Esempi:

[abc] ⟶ **tabc**
[a-z] ⟶ **abc**
[a-zA-Z] ⟶ **Dabc**
[0-9]+ ⟶ **abc20**
[(regex)] ⟶ **he)llo**



Sequenze Speciali

- **\w** - Match di tutti i **caratteri alfanumerici e _ (**equivalente a **[a-zA-Z0-9_]**)
- **\d -** Match di tutti i **caratteri numerici**
- **\D** - Match di tutti i **caratteri NON numerici**
- **\s -** Match di tutti i **caratteri di spazio** (inclusi **\t, \n, \r**)
- **\b** - Match del **bordo di una parola** (vuoto incluso)

Operazioni Avanzate

(Cenni)

Gruppo Nominato

Come gruppo ma ha un nome invece di un indice

Sintassi:

(?<NOME>regex **)**

Backreference

Match di quello che e stato matchato nel gruppo referenziato

Sintassi:

(?P=NOME) oppure

\i - i e l'indice del gruppo

Lookahead

Match di una sequenza che è (o non) **susseguita** dal pattern definito. Non consuma il match

Sintassi:

(?= regex **)** - positivo

(!= regex **)** - negativo

Esempio:

Isaac **(?=** Asimov**)** ⟶

Isaac Asimov

Isaac Newton

Lookbehind

Match di una sequenza che è (o non) **preceduto** dal pattern definito. Non consuma il match

Sintassi:

(?<= regex **)** - positivo

(?<! regex **)** - negativo

Esempio:

(?<= CAP **)** \d{5} ⟶

CAP 20092

20092

Molte altre

Python doc: re module

Cheetsheet

Cheatsheet

LIVE! REC
CODING
(Finalmente)