NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 7: Lecture 1
Exam Review – Case Studies

How can text be a kind of data?

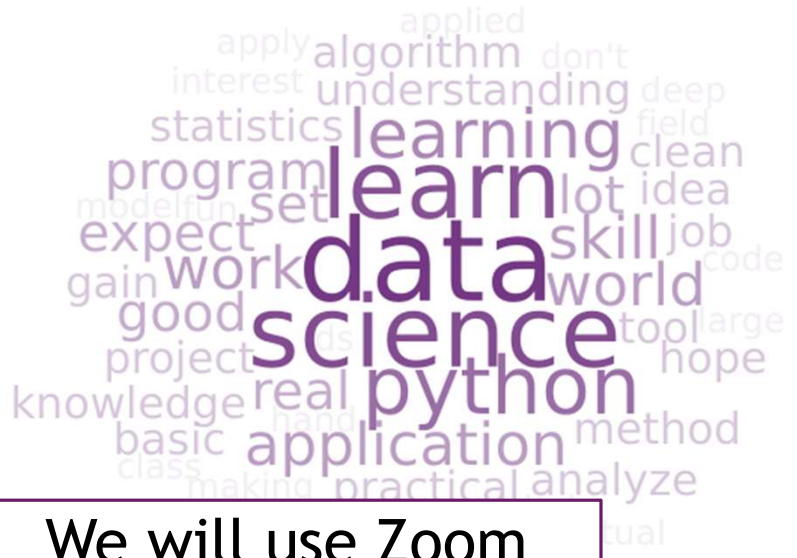
DS-UA 112
Introduction to Data Science

Week 7: Lecture 1
Exam Review – Case Studies

Adapted from Adhikari, DeNero, Hug

Announcements

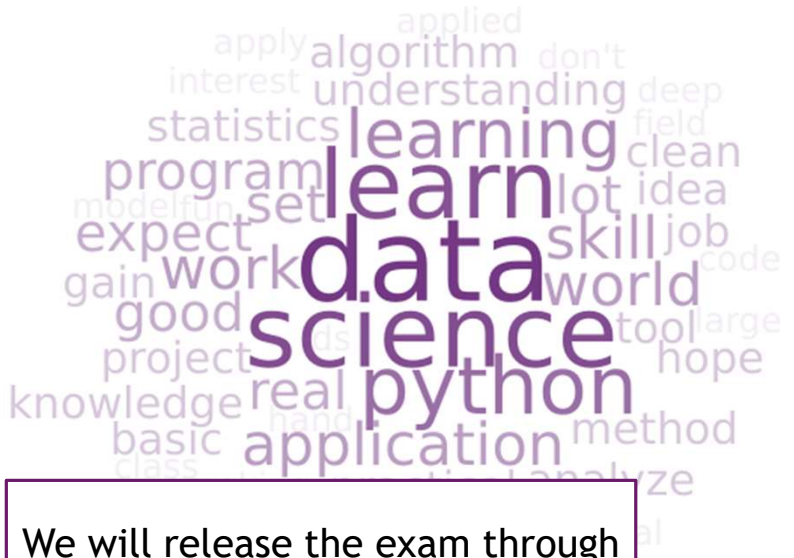
- Please check Week 7 agenda on NYU Classes
  - Lab 6
    - Due on Friday
  - Recordings
    - Lecture and Section
- Remember to post to Piazza



We will use Zoom Conference linked to NYU Classes starting Monday March 23

Announcements

- Midterm
  - Wednesday March 11
    Take-Home
  - Reference Sheet
- Practice
  - Exam
  - Problems
  - Examples



We will release the exam through JupyterHub like homework

Review

operation	order	example	matches	does not match
concatenation	3	AABAAB	AABAAB	every other string
or	4	AA|BAAB	AA BAAB	every other string
closure (zero or more)	2	AB*A	AA ABBBBBBA	AB ABABA
parenthesis	1	A(A|B)AAB	AAAAB ABAAB	every other string
		(AB)*A	A ABABABABA	AA ABBA

Review

operation	example	matches	does not match
any character (except newline)	`.U.U.U.`	`CUMULUS` `JUGULUM`	`SUCCUBUS` `TUMULTUOUS`
character class	`[A-Za-z][a-z]*`	`word` `Capitalized`	`camelCase` `4illegal`
at least one	`jo+hn`	`john` `jooooooohn`	`jhn` `jjohn`
zero or one	`joh?n`	`jon` `john`	`any other string`
repeated exactly {a} times	`j[aeiou]{3}hn`	`jaoehn` `jooohn`	`jhn` `jaeiouhn`
repeated from a to b times: {a,b}	`j[ou]{1,2}hn`	`john` `juohn`	`jhn` `jooohn`

Review

Examples

regex	matches	does not match
`.*SPB.*`	RASPBERRY CRISPBREAD	SUBSPACE SUBSPECIES
`[0-9]{3}-[0-9]{2}-[0-9]{4}`	231-41-5121 573-57-1821	231415121 57-3571821
`[a-z]+@([a-z]+\.)+(edu|com)`	horse@pizza.com horse@pizza.food.com	frank_99@yahoo.com hug@cs

Review

operation	example	matches	does not match
built-in character classes	`\w+` `\d+`	`fawef` `231231`	`this person` `423 people`
character class negation	`[^a-z]+`	`PEPPERS3982` `17211!↑å`	`porch` `CLAmS`
escape character	`cow\.com`	`cow.com`	`cowscom`

Review

operation	example	matches	does not match
beginning of line	`^ark`	`ark two` `ark o ark`	`dark`
end of line	`ark$`	`dark` `ark o ark`	`ark two`
lazy version of zero or more *?	`5.*?5`	`5005` `55`	`5005005`

5.*5 would match this!

Agenda

- Case Study
  - Working with Text as Data
  - Regular Expressions
  - Determining Patterns