RegEx + SQL

Raguvir Kunani

Data 100

July 8, 2019

Data 100 in the News

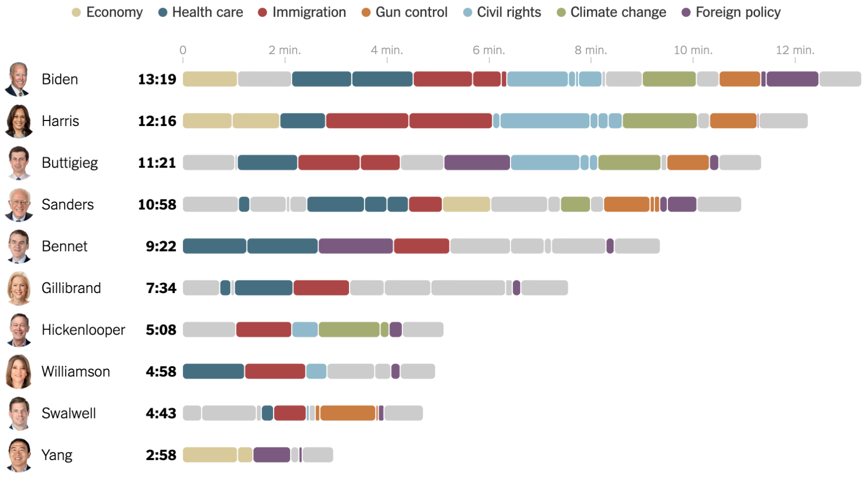
Granularity

- Disaggregation: Breaking up groups into their constituent parts
- The UC recently disaggregated enrollment data on race/ethnicity. This article explains the importance of disaggregating data.
- How does disaggregation relate to granularity?

Data 100 in the News

Data Visualization

One visualization about the 2020 Democratic presidential debate:



How Long Each Candidate Spoke

Each bar segment represents the length of a candidate's response to a question.

What are some good and bad things in the visualization above?

Regex: Overview

What is regex?

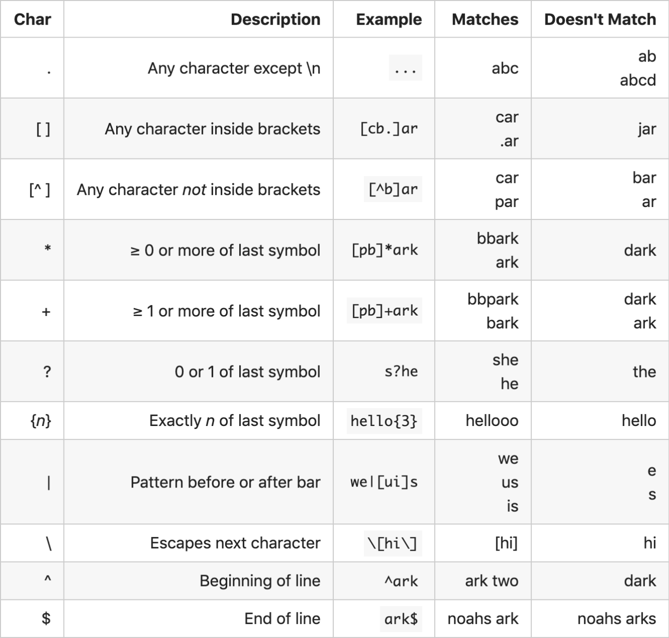
Using regex (regular expressions), we can write expressions that match a certain pattern within a string.

Why do we use regex?

With the tools of string matching and pandas, we can perform analysis or data cleaning on textual data.

Regex: Syntax

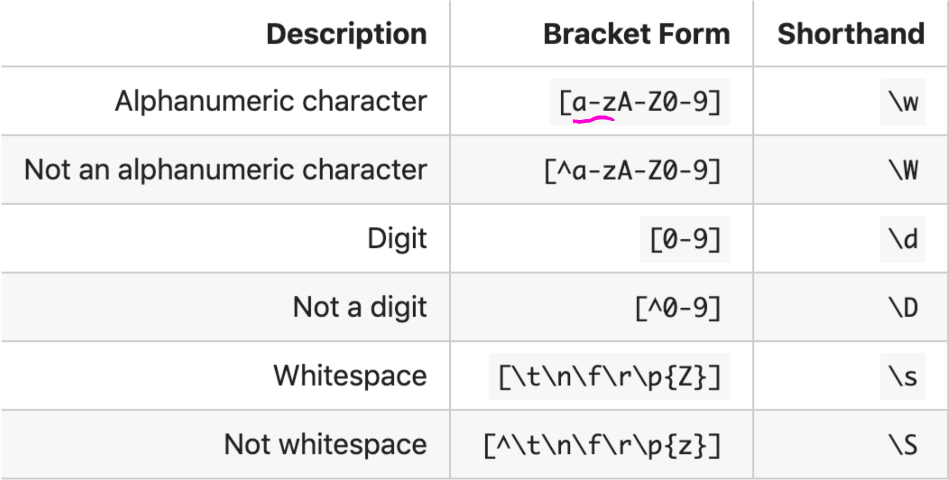
Here is a table summarizing the main parts of regex syntax:



Char	Description	Example	Matches	Doesn't Match
.	Any character except \n	`...`	abc	ab abcd
[ ]	Any character inside brackets	`[cb.]ar`	car .ar	jar
[^]	Any character *not* inside brackets	`[^b]ar`	car par	bar ar
*	≥ 0 or more of last symbol	`[pb]*ark`	bbark ark	dark
+	≥ 1 or more of last symbol	`[pb]+ark`	bbpark bark	dark ark
?	0 or 1 of last symbol	`s?he`	she he	the
{*n*}	Exactly *n* of last symbol	`hello{3}`	hellooo	hello
|	Pattern before or after bar	`we|[ui]s`	we us is	e s
\	Escapes next character	`\[hi\]`	[hi]	hi
^	Beginning of line	`^ark`	ark two	dark
$	End of line	`ark$`	noahs ark	noahs arks

Regex: Shorthand

Here is a table containing common shorthands in regex:



Description	Bracket Form	Shorthand
Alphanumeric character	`[a-zA-Z0-9]`	`\w`
Not an alphanumeric character	`[^a-zA-Z0-9]`	`\W`
Digit	`[0-9]`	`\d`
Not a digit	`[^0-9]`	`\D`
Whitespace	`[\t\n\f\r\p{Z}]`	`\s`
Not whitespace	`[^\t\n\f\r\p{z}]`	`\S`

Final Thoughts on Regex

- Regex is something you just have to learn by doing; there isn't much of a conceptual element to it.
- This resource is super helpful in learning regex. I used it when I took the course and still use it every time I need to use regex.
- Don't worry about learning every minute detail about regex. If you can do most of the past exam questions on regex, that's all you'll need.

SQL Overview

What is SQL?

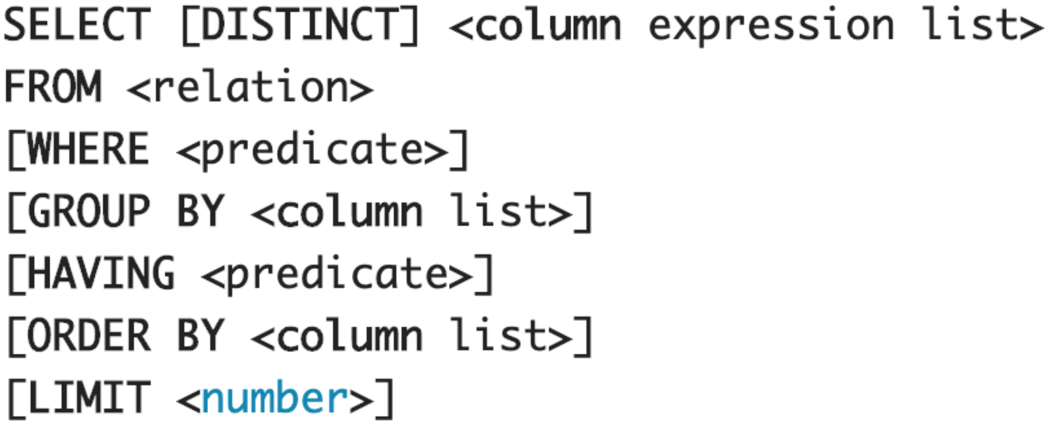
SQL is a language used to interact with relational databases (tables).

Why do we use SQL?

The way `pandas` stores tables is not efficient for large amounts of data. SQL allows us to get information from databases that handle large amounts of data well.

SQL Syntax

The main element of SQL is the `SELECT` statement.



```
SELECT [DISTINCT] <column expression list>
FROM <relation>
[WHERE <predicate>]
[GROUP BY <column list>]
[HAVING <predicate>]
[ORDER BY <column list>]
[LIMIT <number>]
```

Note: The parts in square brackets [ ] are optional.

SQL Order of Evaluation

Unfortunately, the SQL syntax does not match the order of evaluation. Here is the actual order of evaluation:

1. `FROM` : which table(s) are we considering?
2. `WHERE` : selects rows based on a predicate
3. `GROUP BY` : forms groups
4. `HAVING` : selects groups based on a predicate
5. `SELECT` : chooses which column(s) we want in the output

SQL Example

table name: courses

Professor	Course	Term
Sam	Data 100	Summer
DeNero	CS61A	Fall
Hug	CS61B	Spring
Hug	Data 100	Fall
Garcia	CS61A	Spring
Adhikari	Data 8	Spring
Hilfinger	CS61B	Fall
Wagner	Data 8	Fall

all courses taught in Spring

```
SELECT course
FROM courses
WHERE term = 'Spring';
```

↕ equivalent

```
courses[courses['term']=='Spring']['course']
```

Final Thoughts on SQL

- SQL also enables merging of tables (but in SQL it's called joins). See the textbook for how to join tables in SQL.
- SQL is really, really useful. It's worth learning well.
- If you're having trouble with SQL syntax, it might be helpful to either:
    i. Write what you want in `pandas` and then convert to SQL
    ii. Say what you want in English and then convert to SQL

Feedback

- If you have any feedback for me (about my teaching, slides, ot anything else), fill out my feedback form!
- If you have any questions, feel free to ask me in person or by email (rkunani@berkeley.edu).