Quantitative Text Analysis

Meeting 5

Petro Tolochko

Dictionaries and Concepts

Dictionaries

Dictionaries

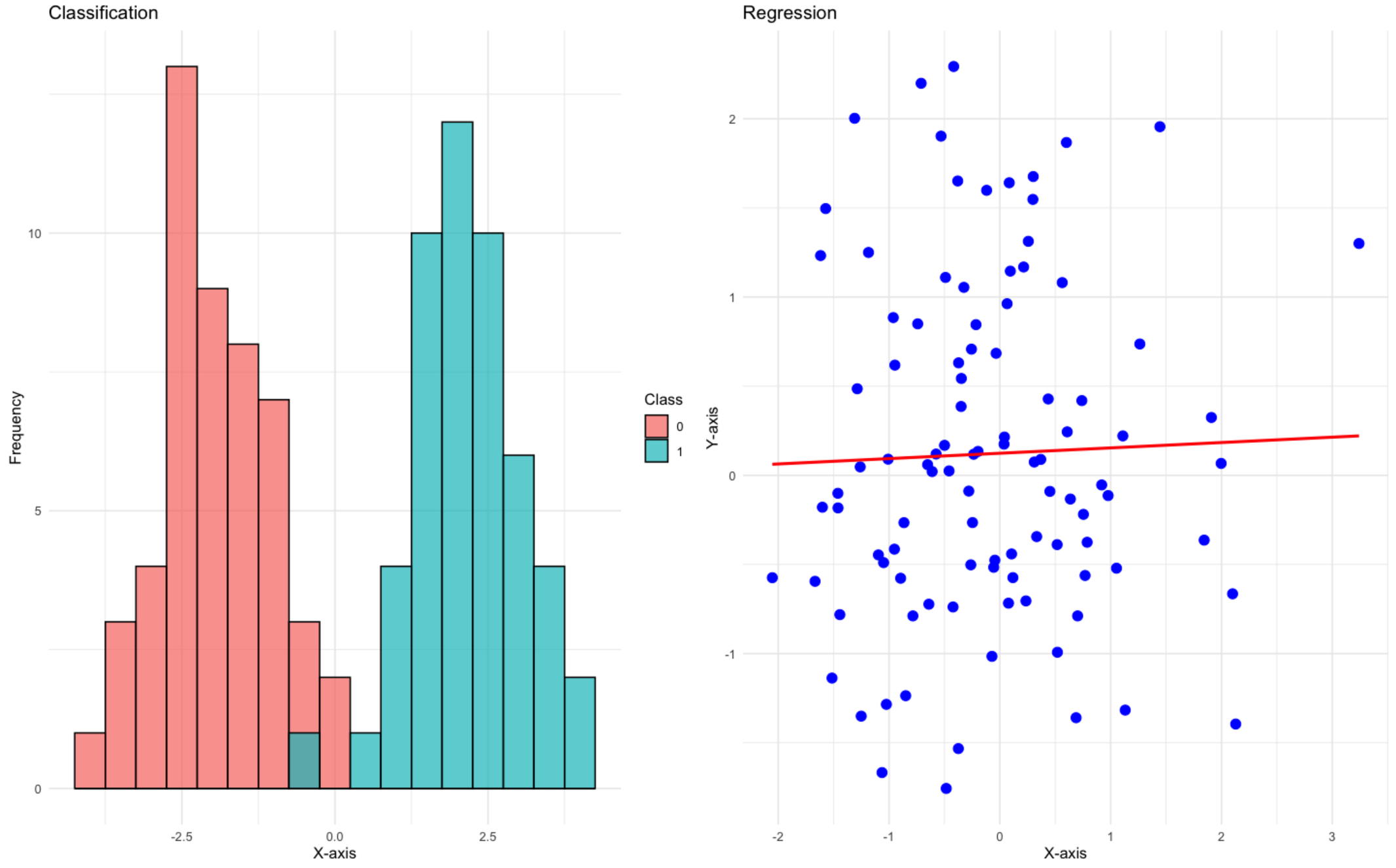
- Rule-based method
- List of words (or phrases) that indicate a category
- Create your own or use/edit existing dictionaries

Use cases for dictionaries

Use cases for dictionaries

- Classification
- Regression
- Search string (form of classification)

Classification vs. Regression


Classification
Frequency
10
5
0
-2.5
0.0
2.5
X-axis
Class
0
1
Regression
Y-axis
2
1
0
-1
-2
-1
0
1
2
3
X-axis

Classification Examples

- Classify texts into:
  - positive/negative
  - populist/non-populist
  - related/unrelated to a certain category

Regression Examples

- “Score” texts on some dimension:
  - Positive/negative
  - Emotional content (anger, sadness, etc.)

Existing vs. Own Dictionaries

Existing vs. Own Dictionaries

- Many existing and validated dictionaries:

Meteor
Media Texts Open Registry
v1.2.2

Search everything...

Login Register Resources About

Name	Type	Country
ConText Diesner, J et al. (2020)	Tool	
DDR Garten, Justin et al. (2017)	Tool	
DICTION Roderick P. Hart (1996)	Tool	
LIWC Pennebaker, J. W. et al. (1999)	Tool	
NLTK NLTK Team (2001)	Tool	
Netlytic Gruzd, A. (2016)	Tool	
T-LAB T-LAB di Lancia Franco	Tool	
WordStat Provalis Research	Tool	
corpustools Welbers K et al. (2018)	Tool	
iLCM Andreas Niekler et al. (2018)	Tool	
popdictR Gründl, Johann (2020)	Tool	
quanteda Benoit, Kenneth et al. (2018)	Tool	
tidytext De Queiroz, Gabriela et al. (2016)	Tool	
tm Feinerer, Ingo et al. (2008)	Tool	

Query

Total results: 14

Free text search

Entity Type

Tool ×

Countries ○ and ○ or

Channel

Languages ○ and ○ or

Used For ○ and ○ or

Dictionary Analysis ×

Concept Variables ○ and ○ or

Programming Languages ○ and ○ or

Existing vs. Own Dictionaries

- Many existing and validated dictionaries:
- Many instances of creating ad-hoc dictionaries:

Heidenreich et al., 2020; Lind et al., 2020

Table 3 Boolean search strings used for retrieval of migration-related news articles

Country	Language	Search string
Spain	Spanish	asilo* OR inmigra* OR refugiad* OR migrante* OR migratori* OR "sin papeles" OR "campo de desplazados" OR patera* OR emigra* OR "libre circulación" OR "fuga de cerebros"
UK	English	asyl* OR immigrant* OR immigrat* OR migrant* OR migrat* OR refugee* OR foreigner* OR "undocumented worker*" OR "guest worker*" OR "foreign worker*" OR emigrat* OR "freedom of movement" OR "free movement"
Germany	German	asyl* OR immigrant* OR immigriert* OR immigrat* OR migrant* OR migrat* OR flüchtling* OR ausländer* OR zuwander* OR zugewander* OR einwander* OR eingewander* OR gastarbeiter* OR "ausländische arbeitnehmer*" OR emigr* OR auswander* OR ausgewander* OR personenfreizügigkeit* OR arbeitnehmerfreizügigkeit* OR "freier personenverkehr*"

Validation in Dictionary Analysis

Validation in Dictionary Analysis

- Source and data selection determines results and conclusions
- Are the selected data sources and selected data points representative for your target concept or discourse?
  - Relevant?
  - Representative?

Relevance of search string validation

- Sampling based on search strings popular (Stryker et al. 2016) and recommended (Barberá et al., 2021)
- Reviews of search string validation procedures
    - out of 83 content analyses, 39% stated the search terms they used, and only 6% discussed their validity (Stryker et al. 2016)
    - out of 105 content analysis studies, 73.3% stated the search terms they used, only 12.4% reported validity metrics (Mahl et al., 2022)
- Careless application of non-validated search terms may lead to noisy inferences (Mahl et al., 2022)

Key validation approach

- How close is an automated measurement to a more trusted measurement:
    - Human understanding of text

Dictionary validation with manually created baseline

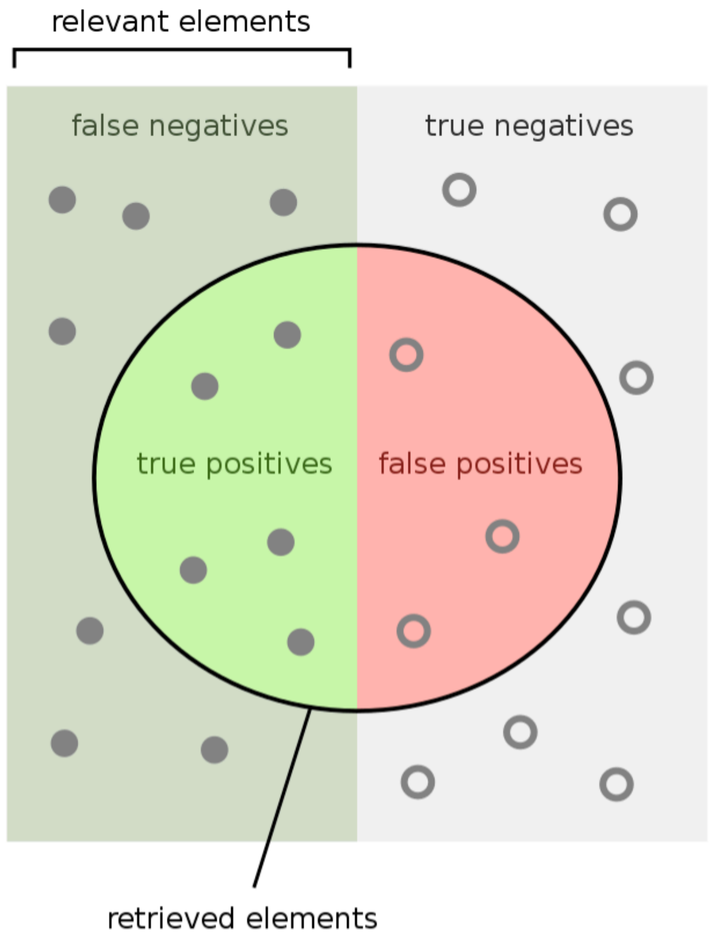
- Code a subset manually (consider intercoder reliability)
- Compare manual decisions with automated classification decisions (via recall, precision, F1)
- Iterative dictionary improvement
- Ideally: manual coding and dictionary development is performed by different persons

Creation of a manual baseline

- Codebook creation
- Who codes manually?
    - Expert coders: Coder recruitment and training sessions
    - Crowdcoders: test questions, majority choice
- Quality assessment: e.g., Inter-coder reliability of involved coders, majority vote
    - How reliable? Consider valid disagreement (Baden et al., 2023)
- Documents selected for baseline should be representative for target discourse (e.g., random selection or artificial week)

Recall, precision, F1

- Metrics frequently used to express the validity of a search string & more generally also of automated classification methods
- Precision (P)
- Recall (R)
- F1 = 2*(P * R)/(P + R)



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{retrieved elements}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

Dictionary pros vs cons

- Pros
    - Often needed to select data (search strings)
    - High reliability and control
    - High transparency and reproducibility
- Cons
    - Difficulty increases with the latency of the construct
    - Language nuances

Questions?

Regular Expressions (regex)

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/
- /ɹɛ.gɛks/
- /ɹɛ.dʒɛks/

Regular Expressions (regex)

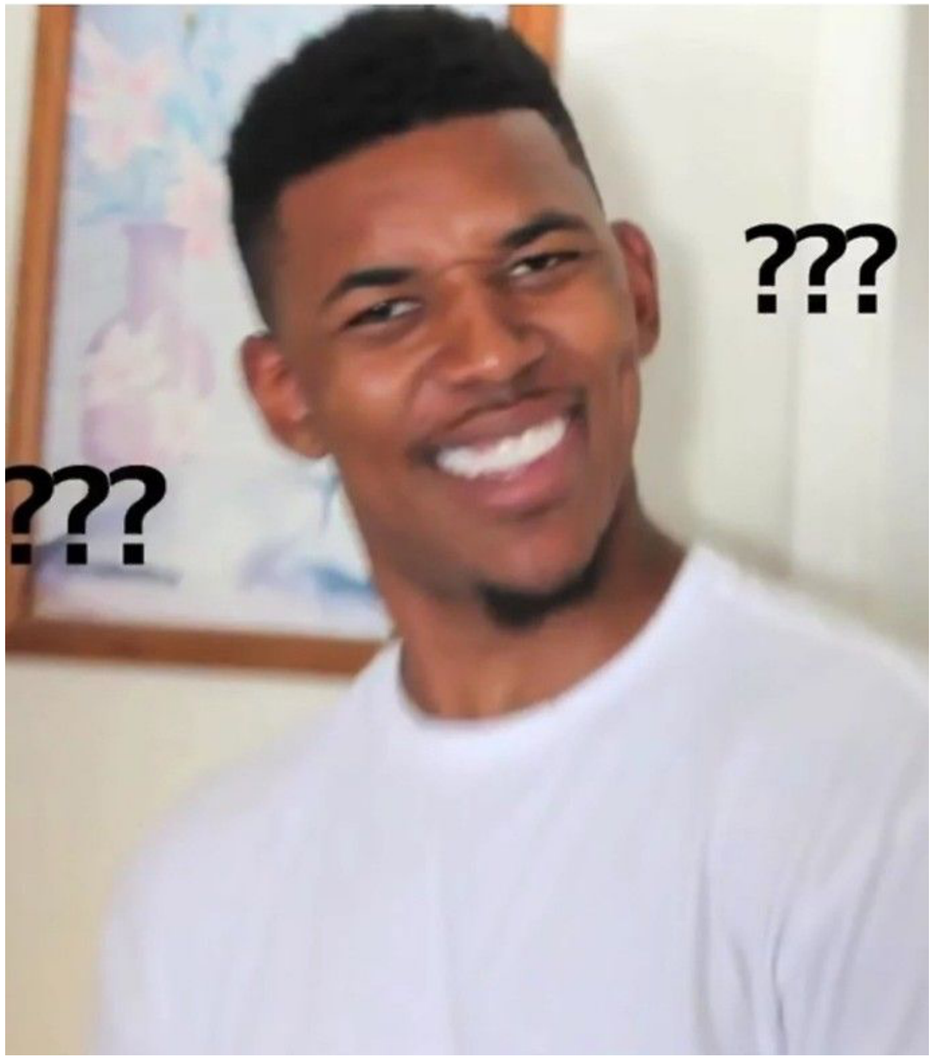
- How to pronounce?
- **/gɪf/**
- /dʒɪf/
- **/ɹɛ.gɛks/**
- /ɹɛ.dʒɛks/

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- **/dʒɪf/**
- /ɹɛ.gɛks/
- **/ɹɛ.dʒɛks/**

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- **/dʒɪf/**
- /ɹɛ.gɛks/
- **/ɹɛ.dʒɛks/**



regex

- Formal language to specify search strings

regex

- Formal language to specify search strings
- Insanely difficult

regex

- Formal language to specify search strings
- ***Insanely*** difficult

regex

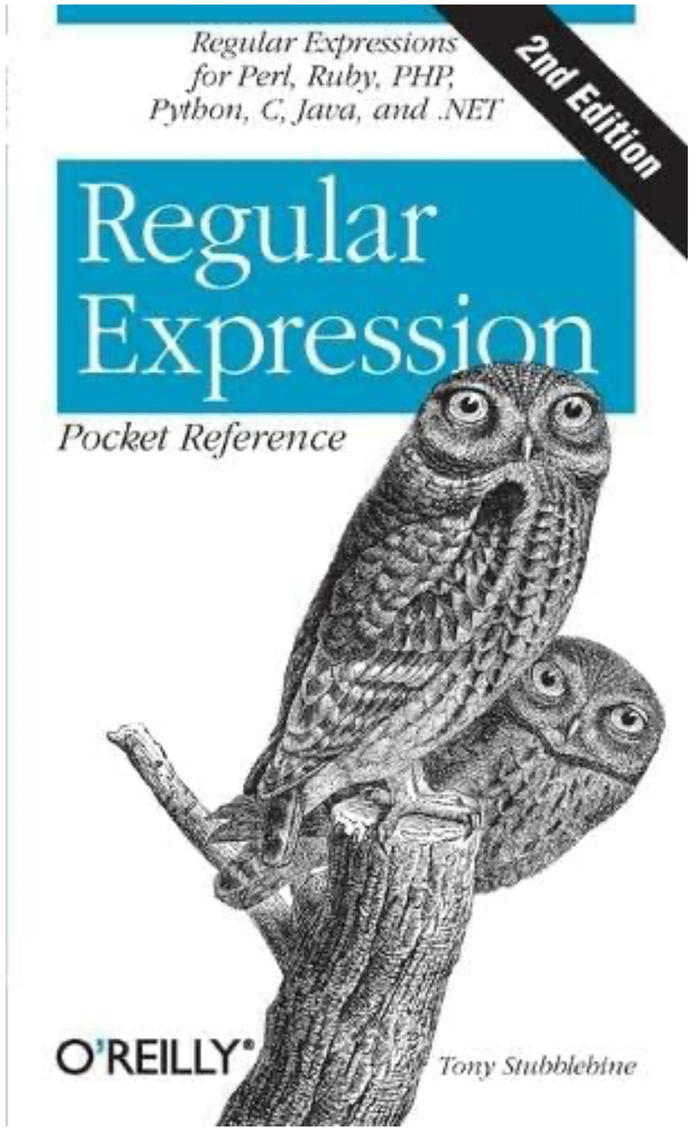
- Formal language to specify search strings
- ***Insanely*** difficult
- Nobody can remember anything

regex

- Formal language to specify search strings
- ***Insanely*** difficult
- Nobody can remember anything
- Different ***flavours***

regex

- Formal language to specify search strings
- ***Insanely*** difficult
- Nobody can remember anything
- Different ***flavours***
- ***“Some people, when confronted with a problem, think “I know, I’ll use regular expressions.” Now they have two problems.” Jamie Zawinski***


Regular Expressions
for Perl, Ruby, PHP,
Python, C, Java, and .NET
2nd Edition
Regular
Expression
Pocket Reference
O'REILLY
Tony Stubblebine

Grimmer / Jurafsky Cheat-sheet

- Disjunctions

RE	Match	Example Patterns Matched
`[mM]oney`	Money or money	"Money"
`[abc]`	'a', 'b', *or* 'c'	"Investing in Iran"
		"is dangerous business"
`[1234567890]`	any digit	"sitting on $7.5 billion dollars"
		"2005 and 2006, more than "
		"$150 million dollars"
`[\.]`	A period	" 'Run!', he screamed."

- Ranges

RE	Match	Example Patterns Matched
`[A-Z]`	an upper case letter	"Rep. Anthony Weiner
		(D-Brooklyn & Queens)"
`[a-z]`	a lower case letter	"ACORN's"
`[0-9]`	a single digit	"(9th CD) "

Grimmer / Jurafsky Cheat-sheet

- Negations

RE	Match	Example Patterns Matched
`[^A-Z]`	not an upper case letter	"ACORN**'s**"
`[^Ss]`	neither 'S' nor 's'	"**ACORN'**s"
`[^\.]`	not a period	" 'Run!', he screamed."

- Optional Characters: `?`, `*`, `+`

RE	Match	Example Patterns Matched
`colou?r`	Words with `u` 0 or 1 times	"color" or "colour "
`oo*h!`	Words with `o` 0 or more times	"oh!" or "ooh!" or "oooh!"
`o+h!`	Words with `o` 1 or more times	"oh!" or "ooh!" or "oooooh!" or

Grimmer / Jurafsky Cheat-sheet

- Start of the line anchor ^, end of the line anchor **$**

RE	Match	Example Patterns Matched
`^[A-Z]`	Upper case start of line	"Palo Alto" "the town of **P**alo Alto"
`^[^A-Z]`	Not upper case start of line	"the town of Palo Alto" "**P**alo Alto"
`^.`	Start of line	"Palo Alto" "the town of Palo Alto"
`.$`	Identify character that ends a line	"Wait**!**" "This is the end**.**"

- "Or"| statements, Useful short hand

RE	Match	Example Patterns Matched
yours|mine	Matches"yours" or "mine"	"it's either yours or mine"
\ d	Any digit	"1-Mississippi"
\ D	Any non-digit	"1-Mississippi"
\ s	Any whitespace character	"1, 2"
\ S	Any non-whitespace character	"1, 2"
\ w	Any alpha-numeric	"1-Mississippi "
\ W	Any non-alpha numeric	"1-Mississippi"

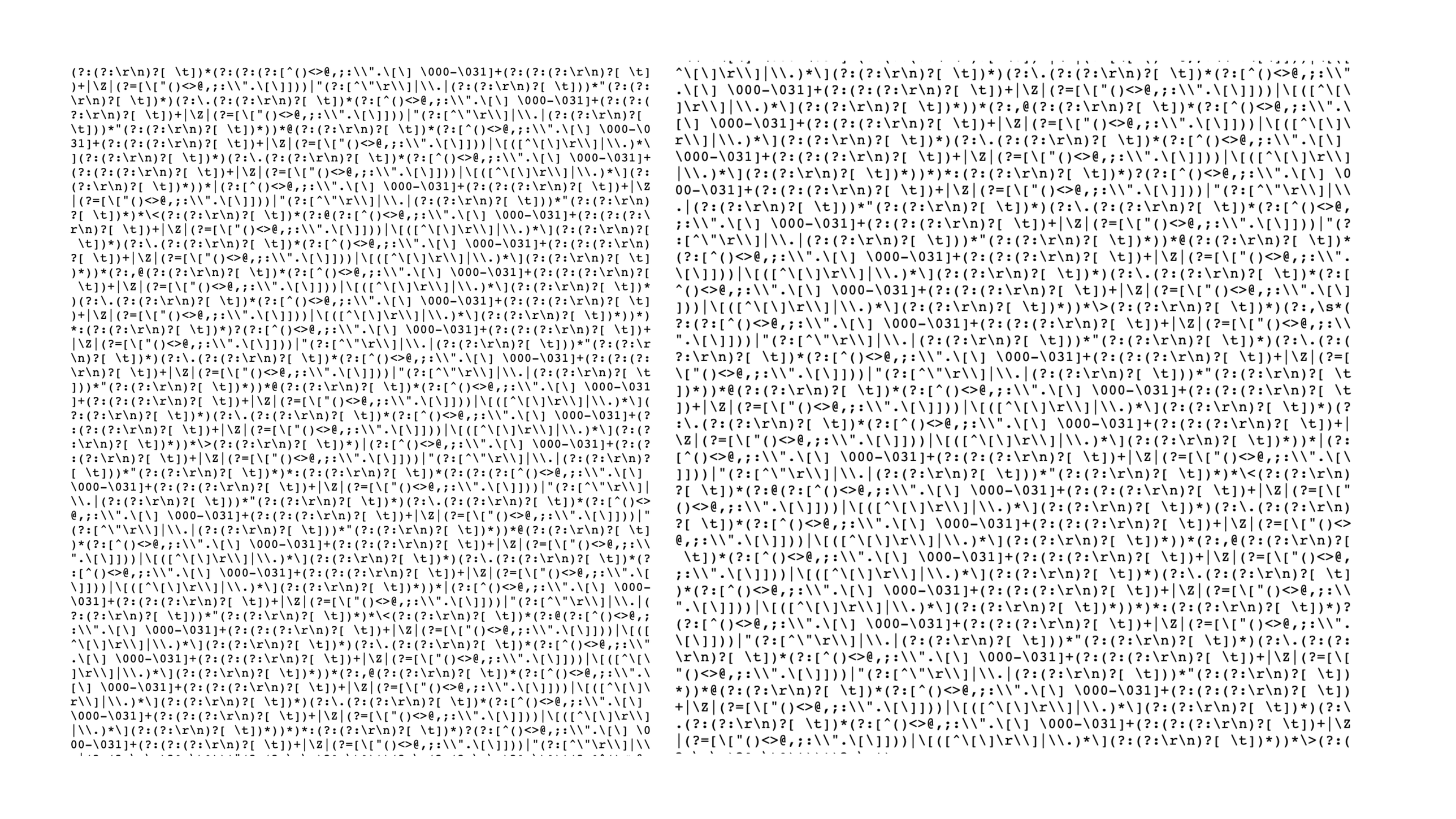
How difficult to regex an email?

How difficult to regex an email?

- Rather…

How difficult to regex an email?

(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:(2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|[1-9]?[0-9]))\.){3}(?:(2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|[1-9]?[0-9])|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])



```
(?:(?:\r\n)?[ \t])*(?:(?:(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t]
)+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:
\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(
?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[
\t]))*"(?:(?:\r\n)?[ \t])*))*@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\0
31]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\
](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+
(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:
(?:\r\n)?[ \t])*))*|(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z
|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)
?[ \t])*)*\<(?:(?:\r\n)?[ \t])*(?:@(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\
r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[
 \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)
?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t]
)*))*(?:,@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[
 \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*
)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t]
)+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*)
*:(?:(?:\r\n)?[ \t])*)?(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+
|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r
\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:
\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t
]))*"(?:(?:\r\n)?[ \t])*))*@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031
]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](
?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?
:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?
:\r\n)?[ \t])*))*\>(?:(?:\r\n)?[ \t])*)|(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?
:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?
[ \t]))*"(?:(?:\r\n)?[ \t])*)*:(?:(?:\r\n)?[ \t])*(?:(?:(?:[^()<>@,;:\\".\[\]
\000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|
\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>
@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"
(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*))*@(?:(?:\r\n)?[ \t]
)*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\
".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?
:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[
\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*|(?:[^()<>@,;:\\".\[\] \000-
\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(
?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*)*\<(?:(?:\r\n)?[ \t])*(?:@(?:[^()<>@,;
:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([
^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\"
.\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\
]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*(?:,@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\
[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\
r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\]
\000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]
|\\.)*\](?:(?:\r\n)?[ \t])*))*:(?:(?:\r\n)?[ \t])*)?(?:[^()<>@,;:\\".\[\] \0
00-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\
```

Concepts and Data

Concepts and Data

- What do we actually want to measure?
- And where do we find it?

Discovery

- Idea from qualitative research
- A “method” to discover new concepts through descriptive analysis
- Grimmer et at., 2022:
    - I. Context Relevance
    - II. No Ground Truth
    - III. Concept vs. Method
    - IV. Data Separation

Principles of Discovery

- Principle 1: ***Context relevance.***
- Text as data models complement theory and substantive knowledge. Contextual knowledge amplifies our ability to make computational discoveries

Principles of Discovery

- Principle 2: ***No ground truth.***
- There is no ground truth conceptualization; only after a concept is fixed can we talk meaningfully about it being right or wrong

Principles of Discovery

- Principle 3: ***Judge the concept, not the method.***
- The method you used to arrive at a conceptualization does not matter for assessing the concept's value – its utility does

Principles of Discovery

- Principle 4: ***Separate data is best.***
- Ideally after data is used for discovery it should be discarded in favor of new data for confirming/testing discoveries.

Codebook

- The codebook is the tool you use to code your content
- It is a kind of questionnaire that you use to inquiry the examined texts/photos/videos
- The codebook should be detailed enough so that
    - you can apply it again in the same way after some time (intracoder reliability)
    - other people (with a little training) can also use it in the same way as you (intercoder reliability)

Codebook

- In automated text analysis:
    - Validation
    - Documentation

Codebook Development

- Iterative process:
  - First draft based on the variables identified in the research question(s)/ hypothesis(s).
  - Tip: take other studies as a model
  - Code material examples using the draft codebook
  - Then refine/edit the codebook

Codebook Function

- Used for dimension reduction (e.g., Egami et al., 2022)

Codebook Function

- Used for dimension reduction (e.g., Egami et al., 2022)