August 2023

Detecting gibberish in open-ended survey questions

Tilburg University

Demian van Gils

Markteffect

Quick introduction

Demian van Gils

Bsc . Psychology
Tilburg University

Msc. Economic Psychology
Tilburg University

Msc. Data Science & Society
Tilburg University

Data Specialist
@ Markteffect

✉ **d.vangils@markteffect.nl**

in **Demian van Gils**

About Markteffect

Markteffect is a full-service market research company

Eindhoven, since 2007

+/- 75 FTE in Eindhoven and +/- 30 FTE in Amsterdam

Needs assessment
Campaign pre-test
Campaign effect
Customer Journey research
Image- & Brand-awareness
... & more!

my JEWELLERY

LIESHOUT SINDS 1719 Bavaria

AUTO DROP

PSV EST 1913

Een actie voor KNGF geleidehond.nl

Fontys

Sport	FMCG	Education	Finance	Leisure
NIKE	Red Bull	Universiteit Utrecht	Kempen	NEDERLANDS OPENLUCHT MUSEUM
umbro	HARIBO	Fontys	Rabobank	GLOW EINDHOVEN
AJAX AMSTERDAM PSV EST 1913	Unilever	Tilburg University	NN	efteling

☎ 040 239 22 90 ✉ sollicitatie@markteffect.nl

Online survey studies

Markteffect

Size & Scope

Around 90% of our studies involve online surveys

Over 3 million responses in 2022

In 33 different languages

Ensuring data quality

Layers of defense

1. In the survey itself
   - Bot detection
   - Survey design

2. In the data quality tool
   - Speeders
   - Patterns (straightliners & outliers)
   - Gibberish

3. By the researcher
   - Sanity check

Open ended questions

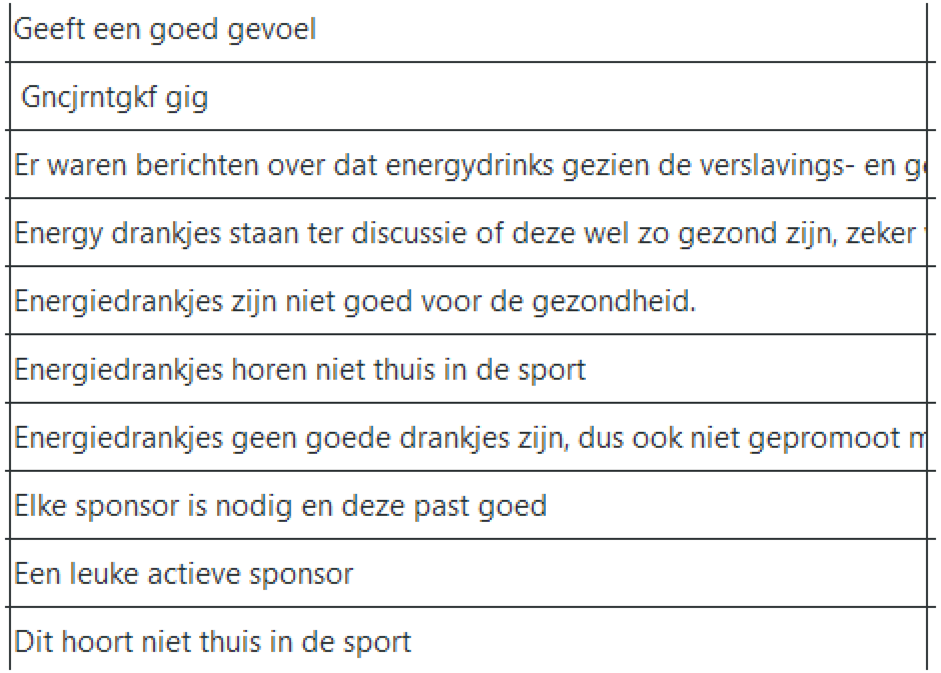
Different types of open ended questions

- *Real* open-ended questions
- Elaboration
- Multiple text
- Escape options

Open ended questions

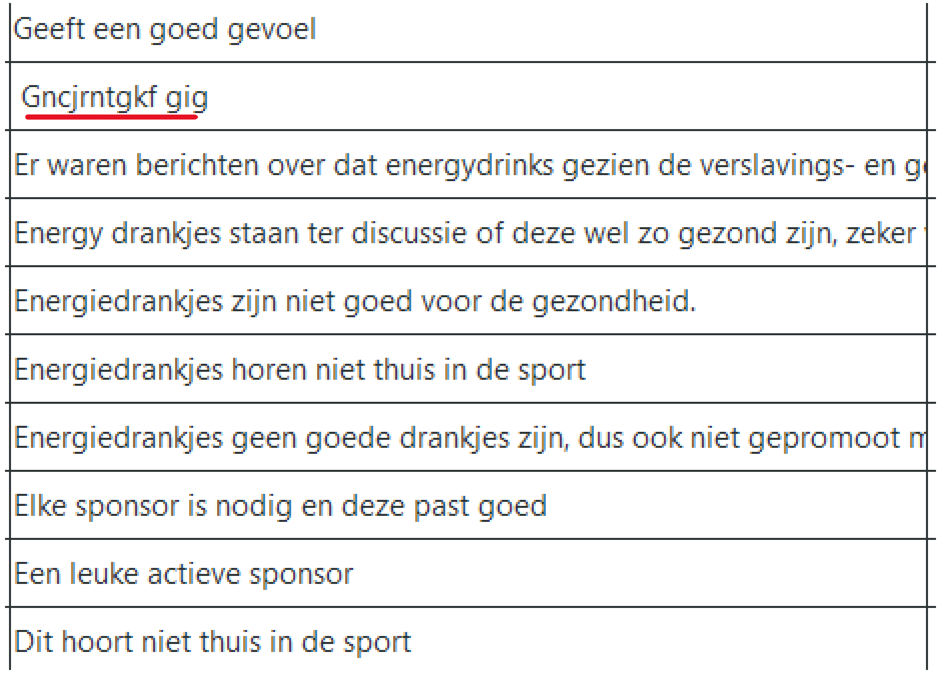
Different types of open ended questions

- Escape options
- Multiple text
- Elaboration
- *Real* open-ended



Geeft een goed gevoel
Gncjrntgkf gig
Er waren berichten over dat energydrinks gezien de verslavings- en g
Energy drankjes staan ter discussie of deze wel zo gezond zijn, zeker
Energiedrankjes zijn niet goed voor de gezondheid.
Energiedrankjes horen niet thuis in de sport
Energiedrankjes geen goede drankjes zijn, dus ook niet gepromoot m
Elke sponsor is nodig en deze past goed
Een leuke actieve sponsor
Dit hoort niet thuis in de sport

Image generated with Deep Floyd



Geeft een goed gevoel
Gncjrntgkf gig
Er waren berichten over dat energydrinks gezien de verslavings- en g
Energy drankjes staan ter discussie of deze wel zo gezond zijn, zeker
Energiedrankjes zijn niet goed voor de gezondheid.
Energiedrankjes horen niet thuis in de sport
Energiedrankjes geen goede drankjes zijn, dus ook niet gepromoot m
Elke sponsor is nodig en deze past goed
Een leuke actieve sponsor
Dit hoort niet thuis in de sport

Project layout

Markteffect

Project layout

What do we need?

- An automated “first layer of defense” against nonsensical text input
- Filters + machine learning

What should we keep in mind?

- Computational constraints
- Limited information
- Multilingual data
- Consequences for respondents

Project layout: Desk research

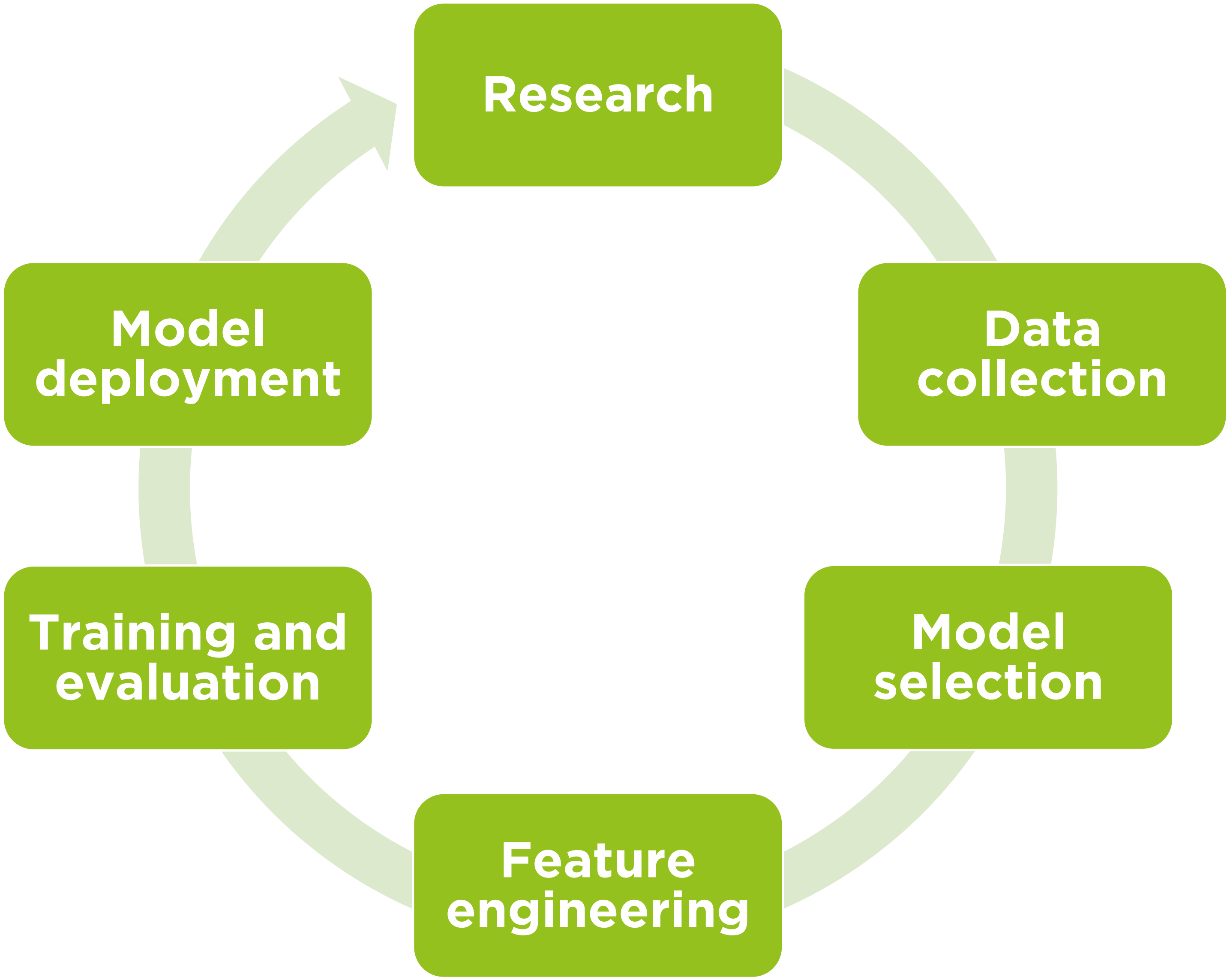
Define key concepts

Read research papers

Look at existing solutions

Project layout

Define our steps



Research

Data collection

Model selection

Feature engineering

Training and evaluation

Model deployment

Data collection

Data collection

Collecting good responses

- Different topics
- Different types
- Different languages

Collecting bad responses

- Is gibberish really ‘random’?

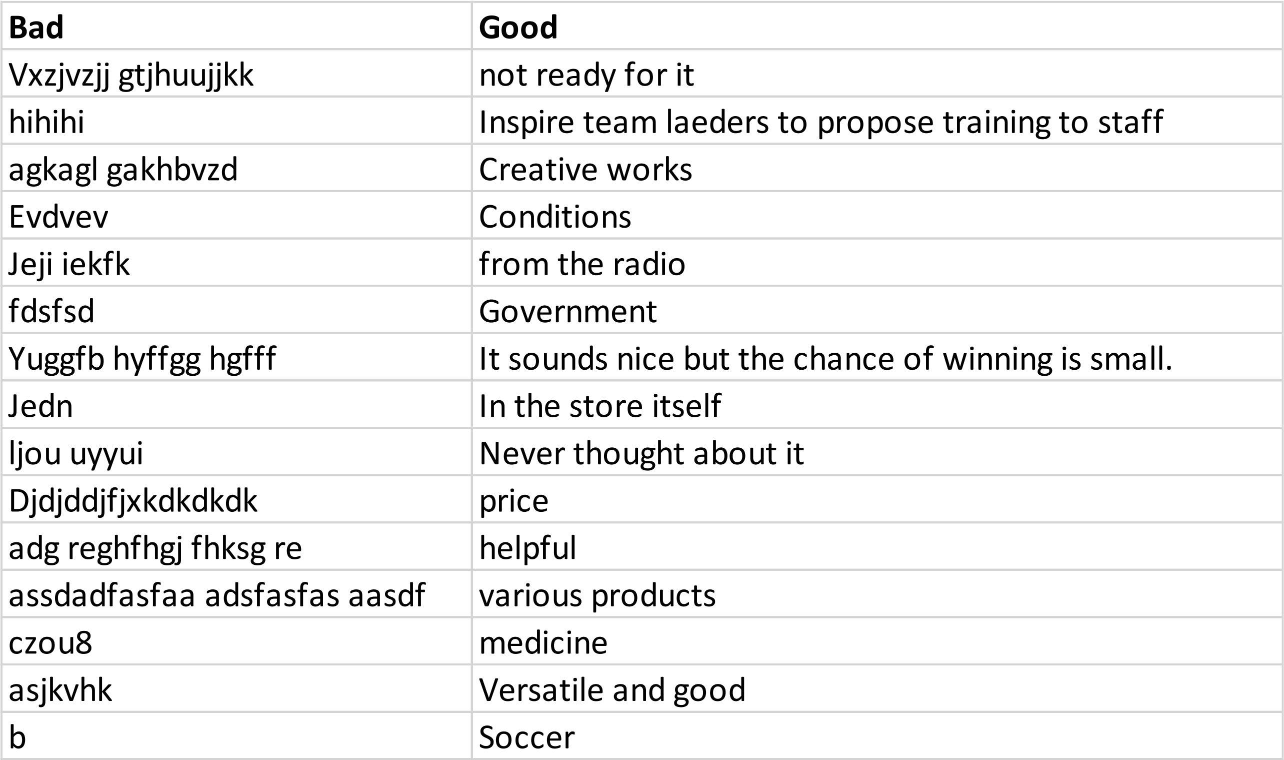
Data collection

Collecting good responses

- Different topics
- Different types
- Different languages

Collecting bad responses

- Is gibberish really ‘random’?



Bad	Good
Vxzjvzjj gtjhuujjkk	not ready for it
hihihi	Inspire team laeders to propose training to staff
agkagl gakhbvzd	Creative works
Evdvev	Conditions
Jeji iekfk	from the radio
fdsfsd	Government
Yuggfb hyffgg hgfff	It sounds nice but the chance of winning is small.
Jedn	In the store itself
ljou uyyui	Never thought about it
Djdjddjfjxkdkdkdk	price
adg reghfhgj fhksg re	helpful
assdadfasfaa adsfasfas aasdf	various products
czou8	medicine
asjkvhk	Versatile and good
b	Soccer

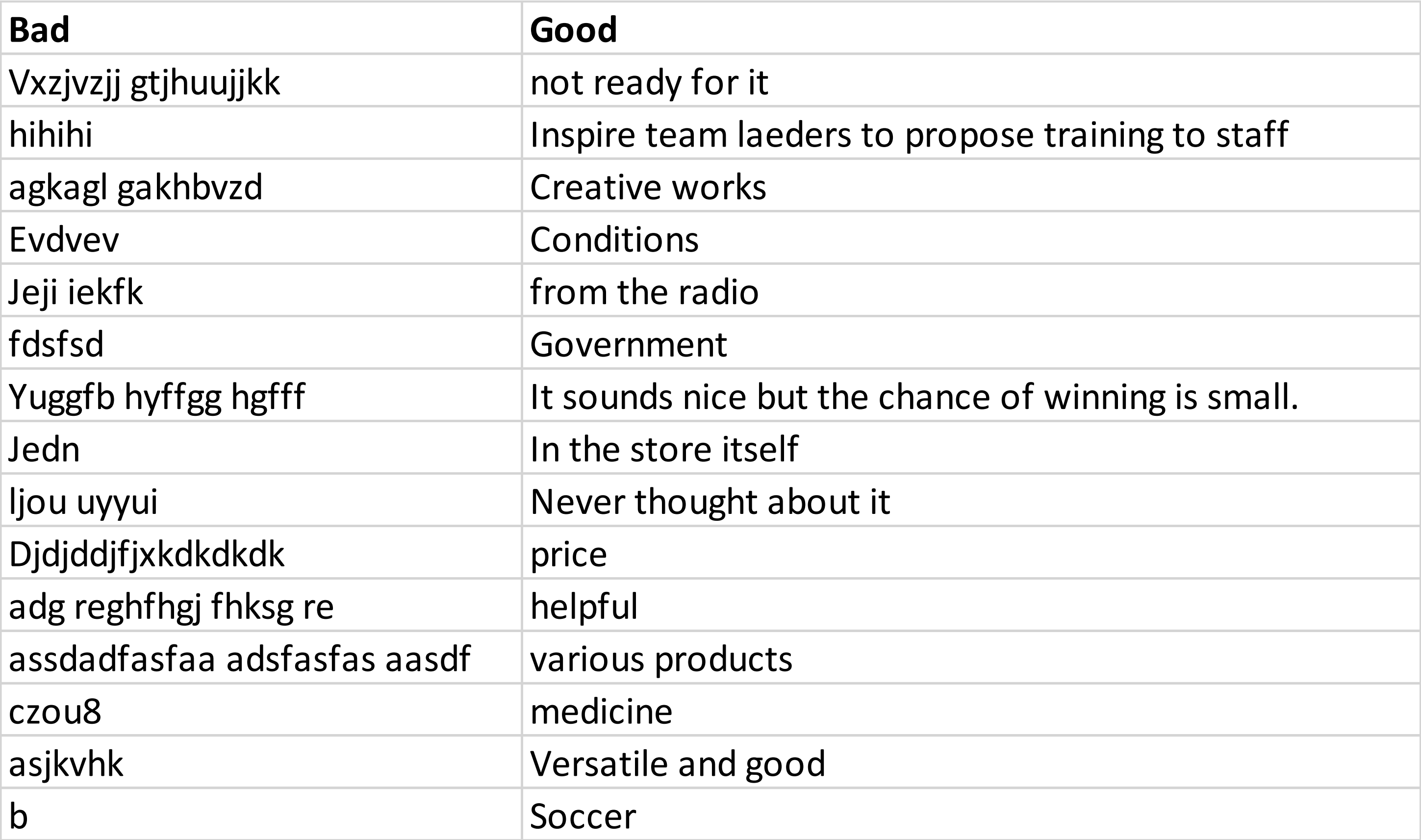
Feature Engineering

Feature engineering

Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling.

Feature engineering

What characteristics can you come up with?



Bad	Good
Vxzjvzjj gtjhuujjkk	not ready for it
hihihi	Inspire team laeders to propose training to staff
agkagl gakhbvzd	Creative works
Evdvev	Conditions
Jeji iekfk	from the radio
fdsfsd	Government
Yuggfb hyffgg hgfff	It sounds nice but the chance of winning is small.
Jedn	In the store itself
ljou uyyui	Never thought about it
Djdjddjfjxkdkdkdk	price
adg reghfhgj fhksg re	helpful
assdadfasfaa adsfasfas aasdf	various products
czou8	medicine
asjkvhk	Versatile and good
b	Soccer

Feature engineering: Proportion of vowels

We can use regex to calculate the proportion of vowels in a string

```
vowels = re.findall("[aeiouáéíóúàèìòùäëïöü]", input_string, re.IGNORECASE)
```

Why do you think the proportion of vowels would be a good predictor for detecting gibberish?

Feature engineering: Proportion of non-alphabetic characters

We can use regex to calculate the proportion of non-alphabetic characters

```
vowels = re.findall("[^a-zA-Z]", input_string)
```

Note that we ignore accented characters in this example.
You could use the unicode package in python to normalize accents.

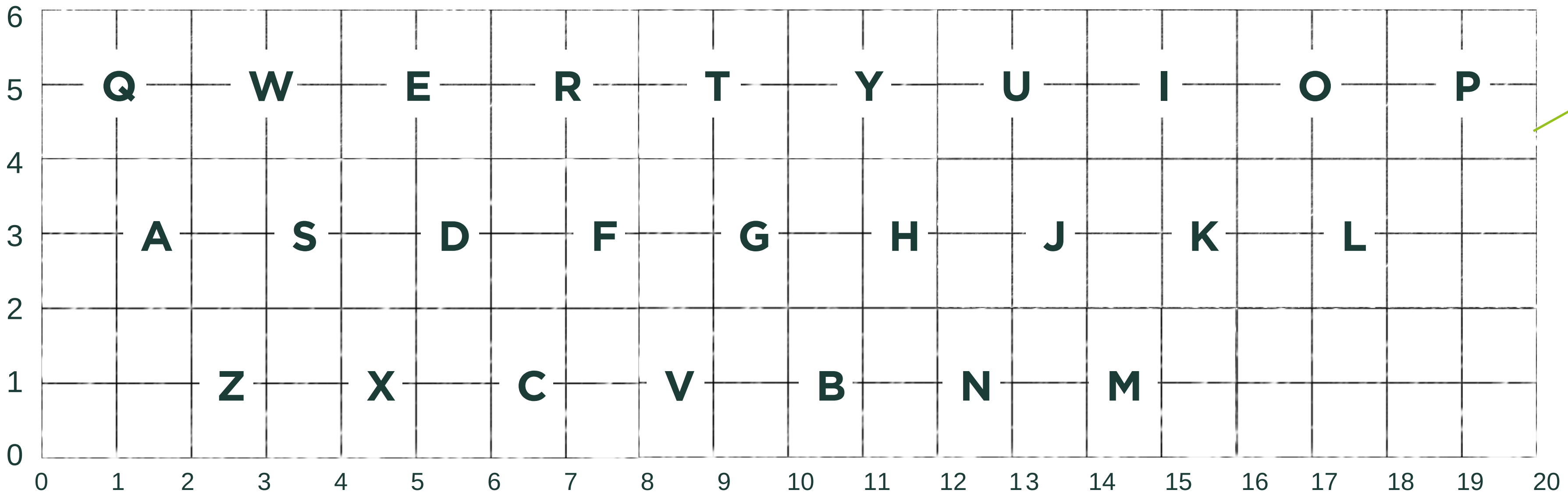
Also note that the pattern `"[^a-zA-Z]"` is not the same as `"[^A-z]"`!

Feature engineering: Keystroke distance

The (average) distance that was traveled on the keyboard to generate the response

Feature engineering: Keystroke distance

We map the keyboard layout on an x-y plane



	X	Y
Q	1	5
W	3	5
E	5	5
R	7	5
T	9	5
Y	11	5
U	13	5
I	15	5
O	17	5
P	19	5
A	1.5	3
S	3.5	3
D	5.5	3
F	7.5	3
G	9.5	3
H	11.5	3
J	13.5	3
K	15.5	3
L	17.5	3
Z	2.5	1
X	4.5	1
C	6.5	1
V	8.5	1
B	10.5	1
N	12.5	1
M	14.5	1

Feature engineering: Keystroke distance

We can now use Euclidean Distance to calculate the distance between each consecutive character in the input string

$$\boldsymbol{Euclidean\ Distance\ (d)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

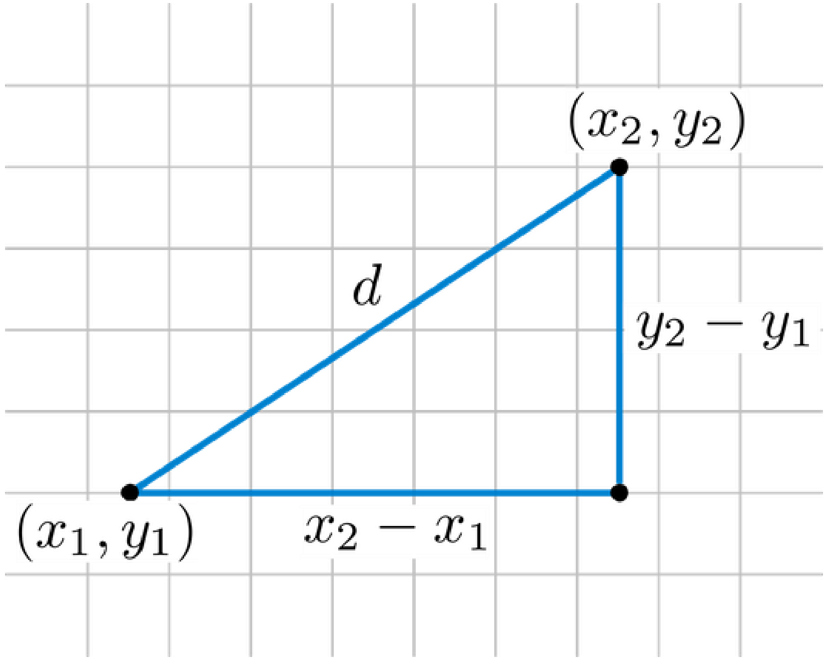


Image from: https://rosalind.info/glossary/euclidean-distance/

Feature engineering: Keystroke distance

We can then calculate the keystroke distance like so:

The keystroke distance is given by: $$\frac{1}{N-1}\sum\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

Where N is the length of the input string

Example “leuk”:

$$KD_{LEUK} = \frac{D_{LE}+D_{EU}+D_{UK}}{3}$$

$$D_{LE} = \sqrt{(17.5-5)^2+\ (3-5)^2} \approx 12.66$$

$$D_{EU} = \sqrt{(5-13)^2+\ (5-5)^2} = 8$$

$$D_{UK} = \sqrt{(13-15.5)^2+\ (5-3)^2}\ \approx 3.2$$

$$KD_{LEUK} \approx \frac{12.66+8+\ 3.2}{3}$$

$$KD_{LEUK} \approx \mathbf{7.95}$$

Feature engineering: Entropy

At its core, entropy is a measure of disorder or uncertainty in a system

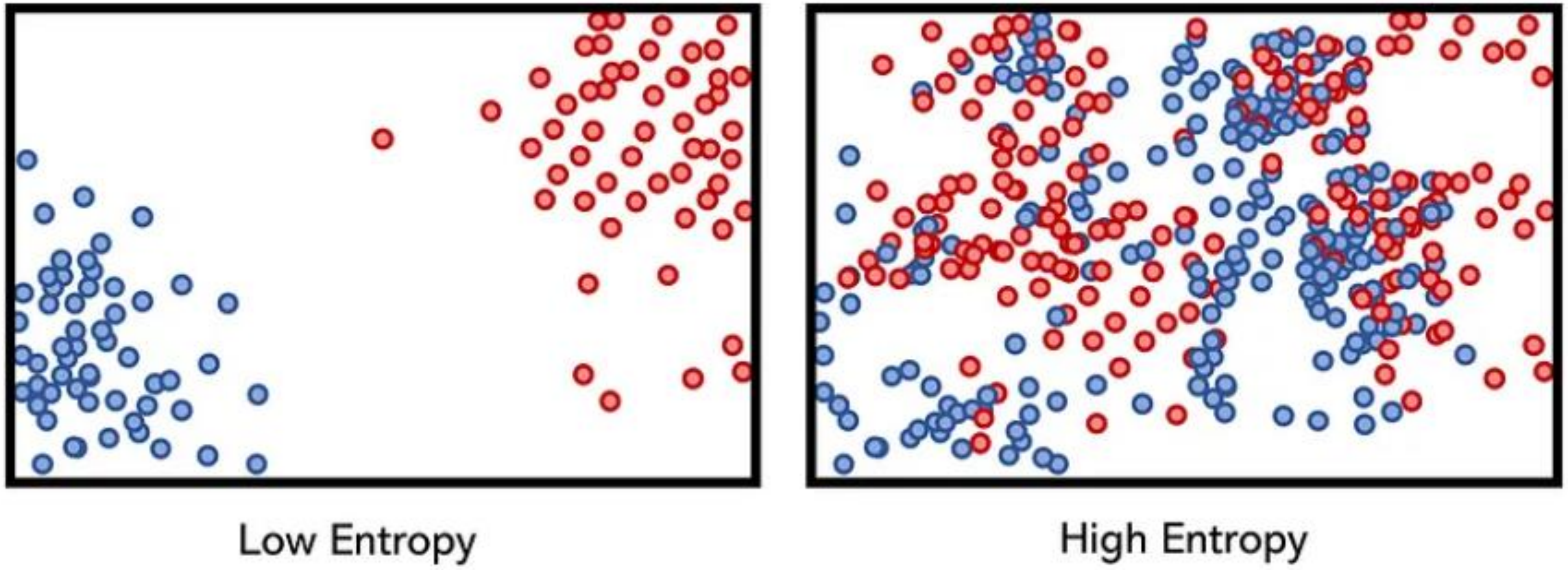


Image from: https://towardsdatascience.com/understanding-entropy-the-golden-measurement-of-machine-learning

Feature engineering: Entropy

A coin toss, using a fair coin, will have high entropy; we cannot accurately predict the next coin toss. Even if we have observed the following: [tails, heads, tails, heads].

A coin toss, using a weighted coin, where we have observed [heads, heads, heads, heads] has low entropy. We can be quite certain that the next coin toss will yield heads.

Feature engineering: Entropy

The (binary) entropy of a string

$$Entropy(S) = -\sum_i p(i) * log2(p(i))$$

Example “leuk”:

Entropy(leuk) = -¼ * log2(¼) - ¼ * log2(¼) - ¼ * log2(¼) - ¼ * log2(¼)

= -0.25 * -2 - 0.25 * -2 - 0.25 * -2 - 0.25 * -2

= 2

Notebook

Notebook

You can find the notebook and datasets at:

https://github.com/markteffect/guestlecture-uvt

Markteffect