Upskilling and exploring avenues for information specialists in collaboration with data scientists

"pattern matching" as a means of refining search results

NICE National Institute for Health and Care Excellence

NICE Information services

- NICE
  - UK non-departmental government body
  - Provides guidance & recommendations for health & care practitioners
  - Assesses new health technologies
- Information Services
  - Embedded in the organisation business plan
  - Activities include: intelligence gathering, scoping, literature searching

Coding team

- 2022 – Dedicated coding microteam
  - 6 Senior Information Specialists
  - 12 months
  - Posit (R studio)
  - Focus on IS service improvements
  - Output: Completed 5 projects, 3 additional projects were in progress

NICE

Projects completed

- Trials reformatting
- Trials to RIS format
- Topic selection reformat
- BNF manufacturer scraper
- RIS to DOI OVID strategy
- Cochrane to CRD translator
- NICE bibliographies

NICE

User uptake

- The team access the code via Posit Cloud (web based)
- Demonstrated the code once it was available
- Reminders in team meetings
- Time saved

NICE

Questions?

Any questions about the IS coding team project?

NICE National Institute for Health and Care Excellence

Custom classifier

- Continuation of the IS upskilling objective
- Two topics at NICE had created customer classifiers; COVID & breast cancer
- The aim of the project was:
  - to explore how custom classifiers work
  - to consider if classifiers are an area IS should explore further
  - how its use aligns with our current search practices
  - Assess if it should be incorporated to EPPI R5
- The concept of Humans and Animals were used to explore the project aims

NICE

Animals/non-human Vs Humans

- No animal/human limits for some databases
- Currently we have two different animal filters
- Humans/human studies are clearly defined…

NICE

Overview

- Supervised machine learning
- Requires training data – lesson: it needs to be clean
- Algorithm:
  - bag of words approach applied to training data
  - Words turned into vectors and given a number; 1 or 0
  - Stop words are not included

NICE

Results

We created and tested 3 versions of the custom classifier

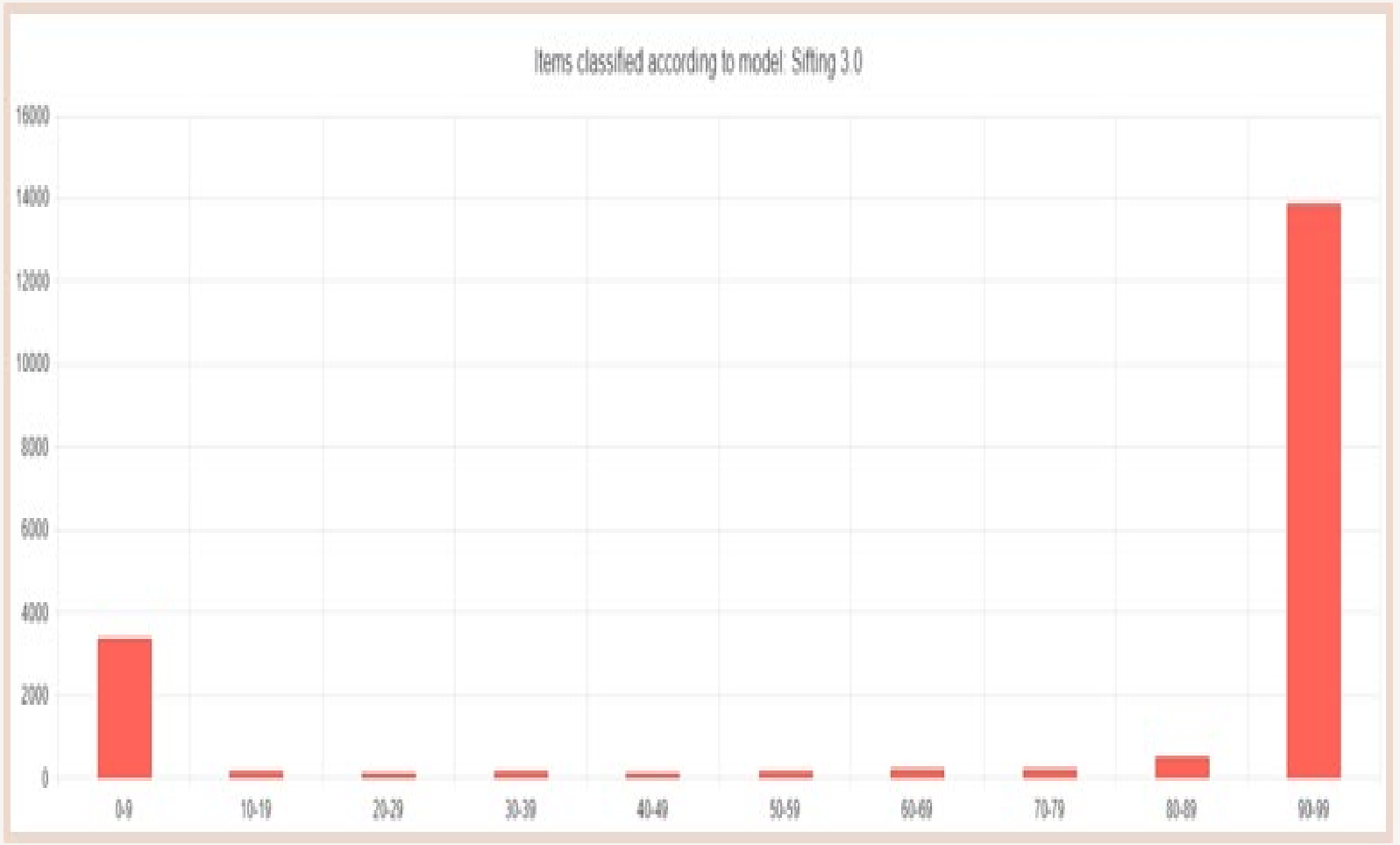
Version 3;

15757 human references

2963 animal references

Focused on the 0-9 interval – refinement still required

Project paused while IT incorporated the functionality into EPPI R5



NICE National Institute for Health and Care Excellence

Lessons learned

- Chose two defined concepts; be prepared to further refine those concepts
- Clean your data and then clean it again
- For the project team, we found the black box element conflicted with our need for transparency

NICE

Questions?

Any questions about custom classifiers?

NICE National Institute for Health and Care Excellence

Pattern Matching

Topics explored

COVID

Breast Cancer

Human/Animal

Work with a data scientist to:

Explore how pattern matching works

Explore if/how pattern matching could be utilised by IS

NICE

Compiling terms for concepts

- Animal terms:
  - adapted from a list on Github
  - generic animal terms e.g. minipig
  - latin animal terms e.g. drosophilia
  - 29 idioms
  - Lemmas
- 25 human terms

NICE

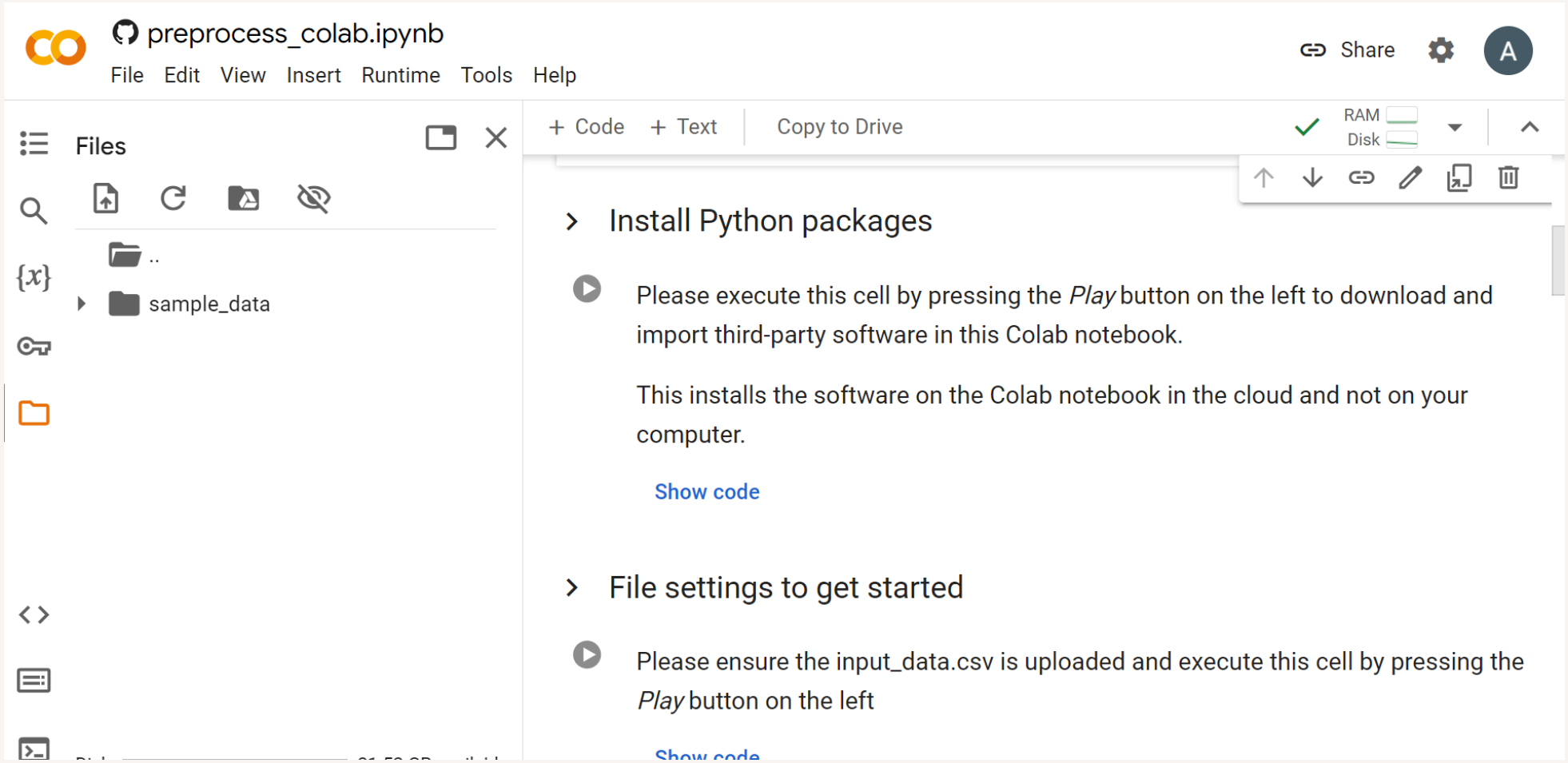
```
#@title Animal and Human terms { form-width: "20%" }
#Below animal list is adapted from https://gist.github.com/borlaym/585e2e09dd6abd9b0d0a
animals = [
    "animal",
    "bird",
    "fish",
    "Aardvark",
    "Albatross",
    "Alligator",
    "Alpaca",
    "Ant",
    "Anteater",
    "Antelope",
    "Ape",
    "Armadillo",
    "Donkey",
    "Baboon",
    "Badger",
    "Barracuda",
    "Bat",
    "Beaver",
    "Bee",
    "Bison",
    "Boar",
    "Buck",
    "Buffalo",
    "Bull",
    "Butterfly",
    "Camel",
```

NICE

```
lowercase_animals_specific = ["drosophila",
                              "arabidopsis",
                              "amblycephala",
                              "chrysodeixis", #Chrysodeixis chalcites
                              "daphnia" #Daphnia magna
                              "dicentrarchus", #Dicentrarchus labrax L
                              "lepidoptera",
                              "esper",
                              "megalobramaamblycephala",
                              "noctuidae",
                              "marmoset"
]

#Below are idioms using animal names within them, but should not be counted as animals
#any punctuations would be replaced by 'space' when the text is pre-processed, so the punctuations in the
#below idioms should reflect that.
lowercase_animal_idioms = ["sitting duck",
                           "fly on the wall",
                           "bee s knees",
                           "bees knees",
                           "two birds with one stone",
                           "chicken out",
                           "wild goose chase",
                           "horse around",
                           "until the cows come home",
                           "dark horse",
                           "hold your horses",
                           "straight from the horse s mouth",
                           "two shakes of a lamb s tail",
```

NICE


preprocess_colab.ipynb
File Edit View Insert Runtime Tools Help
Share
A
Files
+ Code + Text
Copy to Drive
RAM
Disk
..
sample_data
Install Python packages
Please execute this cell by pressing the *Play* button on the left to download and import third-party software in this Colab notebook.
This installs the software on the Colab notebook in the cloud and not on your computer.
Show code
File settings to get started
Please ensure the input_data.csv is uploaded and execute this cell by pressing the *Play* button on the left
Show code

NICE

What have we found?

- Transparency – no ‘black box’
- Language is complex - Zwei Fliegen mit einer Klappe schlagen
- Context
- Usability
- Coding skills

NICE

Collaborative working

- Data Scientist
- Different skill set
- Breaking down barriers

NICE

Questions

Any questions about pattern matching?

NICE National Institute for Health and Care Excellence

Thank you for listening

Contact information: Amy.Finnegan@nice.org.uk

For this project I have worked with several colleagues:

- IS coding team
- Nicola Walsh – Senior Information Manager
- Mariam Sood – Data Scientist
- Daniel Tuvey – Senior Information Specialist
- Paul Levay – Senior Information Specialist

NICE National Institute for Health and Care Excellence