QUANTUM

Our Global **Data Science Practice**

Introduction – a problem in unstructured data

A typical company holds a huge amount of data as text – reports, reviews, communications, applications, notes. Finding a way to automatically extract information from text can be hugely valuable – it may allow you to listen to your customers at scale, to automate report-based decision-making or even to design a chatbot.

This challenge mirrors the lifecycle of a typical NLP (Natural-Language Processing) Data Science project in Aviva. You’ll need to understand the problem, clean and explore the data, solve the data problem and present a sensible solution.

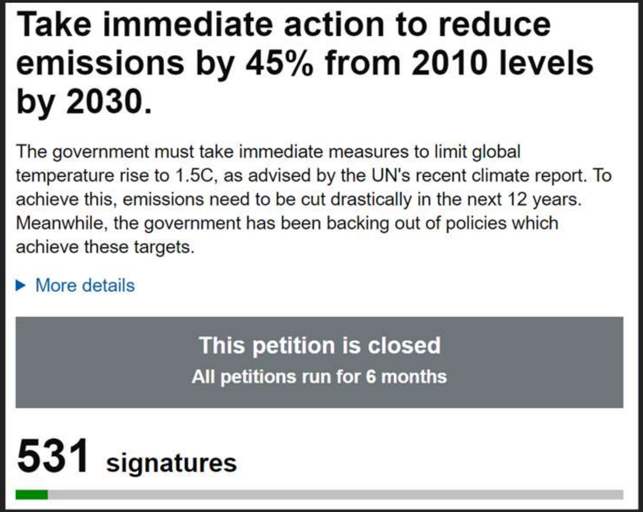
The Challenge – understanding public concerns

The challenge is to analyse the text of public petitions. The data comes in two parts:

- training_data.json - the text (title and abstract) of 12,387 petitions, plus the number of signatures gained by each one
- holdout_data.json – the text of a further 3,000 petitions

For **Part I** of the challenge, create a model to predict whether each petition in the holdout set surpassed 50 signatures. You'll need to submit your predictions and will be scored against the reality.

For **Part II**, design a method to automatically group together petitions of similar topics. You'll need to explain your method and show your results.



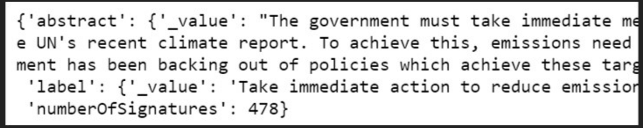
Take immediate action to reduce emissions by 45% from 2010 levels by 2030.

The government must take immediate measures to limit global temperature rise to 1.5C, as advised by the UN's recent climate report. To achieve this, emissions need to be cut drastically in the next 12 years. Meanwhile, the government has been backing out of policies which achieve these targets.

▸ More details

This petition is closed
All petitions run for 6 months

531 signatures



```
{'abstract': {'_value': "The government must take immediate me
e UN's recent climate report. To achieve this, emissions need
ment has been backing out of policies which achieve these targ
 'label': {'_value': 'Take immediate action to reduce emissio
 'numberOfSignatures': 478}
```

(Contains Parliamentary information licensed under the Open Parliament Licence v3.0. https://www.parliament.uk/site-information/copyright/open-parliament-licence/)

Rules, hints and everything else

- Use any coding language you like. You may wish to look at the provided Jupyter Notebook for some hints even if you don't use python.
- Do not use any data besides the data provided.
- For **Part I**, predict whether each petition in the holdout set has more than 50 signatures. The submission should be a CSV list of 3,000 Booleans in the same order as the holdout petitions and must be emailed across on the hack day.
- On the hack day, you'll need to present your approach and results for both parts of the challenge. We're interested in what you did and why you did it.
- The winner will need to submit their code so that we can verify the approach is reproducible.
- You'll be assigned points based on:
    - The success of your predictions in **Part I** (measured by F1 score)
    - The approaches you use (particularly think about best practices in Data Science, e.g. cross-validation)
    - The convincingness of your topic modelling in **Part II** (think about how you would convey your topics to a manager and how you would convince them that your topics are 'right')
    - Your overall communication (in the real world, you would need to convince both technical and business managers that your work is robust and commercially useful)

Email any questions to jonty.haberfield@aviva.com.

QUANTUM
Our Global Data Science Practice