# Natural Language Processing for Creatives: Project report

## Methods

The project explores different ways of analysing YouTube comments, using:

- Sentiment analysis
- Word clouds
- LDA
- Bag of words

## Description of code sources, data, its source, and collection

### Code sources

**Code Classes:**

**Intro to data science**
Fun API Labs: Exploring YouTube Data with YouTube API

**NLP for creatives**
Classification with topic modelling features (using BoW and LDA)
We Rate-dogs-sentiment-analysis

**Additional codes**

**Geeks for Geeks: YouTube analysis**
Sentiment Analysis of YouTube Comments
YouTube Data Scraping, Preprocessing and Analysis using Python (Word Cloud)
How To Create A .gitignore File To Hide Your API Keys

**Hiding API keys**
How to HIDE Your API Keys in Python Projects
How to Hide API Keys in Python: An Environment Variables Example

## What you learnt during working on the project

I learnt methods of hiding API keys and explored testing and training classifiers. I learnt how to create a word cloud with Python.

## Challenges and how you approached them,

I experienced issues retrieving data via a Google API. I received a 400 error, when tried to hide an API key and test whether the code still worked for YouTube data. I did not understand instructions from the initial pages I consulted, so review multiple sources and YouTube videos.

I received the below error. I researched the cause of this error and changed the code to “target_names=None” after which, the line of code worked.

```
report = classification_report(y_test, y_pred, target_names=class_names)
```

```
gave an error ValueError: Number of classes, 20, does not match size of
target_names, 33. Try specifying the labels parameter
```

## What went well in your project,

The project provided examples of different ways to interpret YouTube comments. The graphs are clear to read.

## What you would do differently if working on a similar challenge in the future

For the word cloud, I would explore filtering out words according to word class. For example, if adjectives. At the moment this graphic has limited usefulness. I would also change the colour scheme and have a key to indicate what the different colours represent.

I would provide more explanations for the results of the LDA and Bag of Words classification.

## Ethical statement:

### Your take on the data

YouTube comment data is variable - it will change as the most popular videos change and as comments are added to the videos. It is only representative of the time the data was retrieved.

### Sources and bias

The data is directly from YouTube who are not transparent about the algorithms determining what becomes the top videos and comments is subject to YouTube's algorithm(s), which they are not transparent about. Comments that can be viewed could be restricted by YouTube or individual channel guidelines, e.g. some negative comments may be filtered out if they are considered hateful.

In sentiment analysis and text classification, there could be bias in design of what is decided to be positive or negative and there may be misclassification. There could be false positives and false negatives. For example, when the sentiment analysis was run on another video, it interpreted a potential praise comment with "😭" emojis as negative, when the emojis and rhetorical question can have different interpretations in informal speech.

```
The comment with most positive sentiment: Ari is simply amazing. ❤❤❤❤❤❤ with
score 0.9849 and length 29
The comment with most negative sentiment: Ari are u even real😭😭😭😭💗💗💗💗
💗 with score -0.9081 and length 28
```

Written comments can miss tone, there can be difference in meaning between authors and meaning can also change over time, faster than libraries and packages are updated.

### Potential uses of your project,

Additional ways to build on the project include comparing the results with other videos in the same category, other categories and comparison between countries.

The data could be analysed further with different classifiers to compare results.

## LLM (Large Language Model) disclaimer

Not directly used to write the code.