# McDonalds Review Sentiment (NLP)

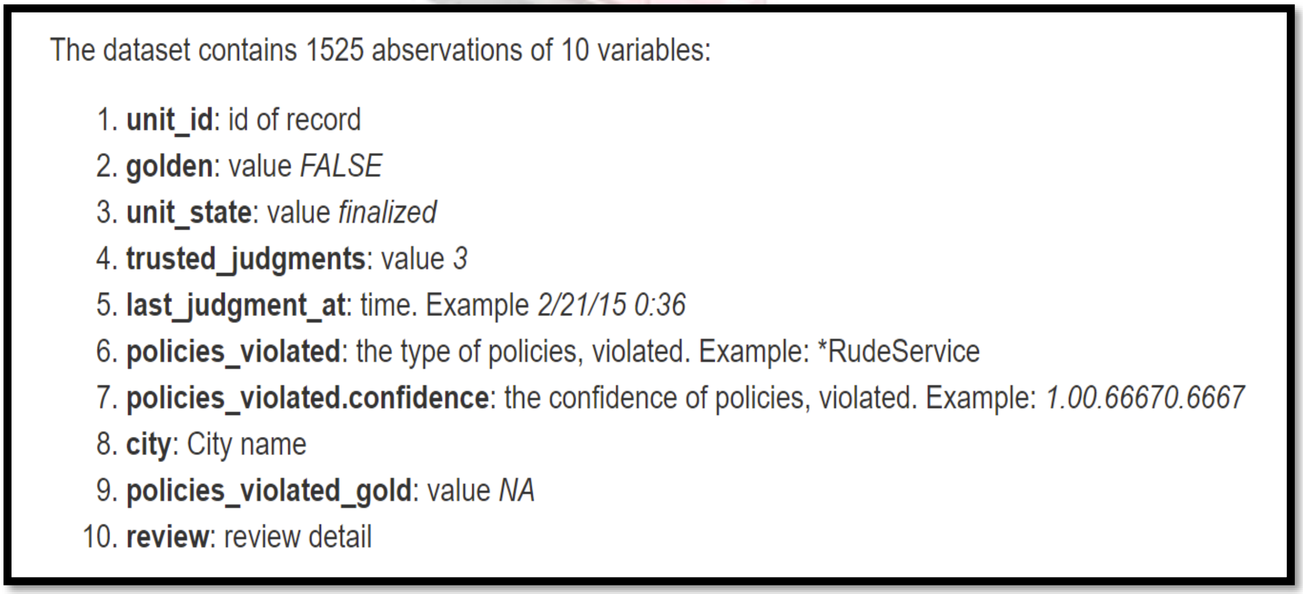
## Abstract:

Executive summary McDonald's is the **world's largest chain of fast-food restaurants**, serving around 68 million customers daily in 119 countries across more than 36,000 outlets. Founded in the United States in 1940, A McDonald's restaurant is operated by a franchisee, an affiliate, or the corporation itself. In this project we have done a sentiment analysis of negative McDonald's reviews. Contributors were given reviews culled from low-rated McDonald's from random metro areas and asked to classify why the locations received low reviews.

## Design:

This project is one of the T5 Data Science Boot Camp requirements. Data provided by data. World. In this module we will be laying the foundation for our analysis by processing and exploring a large amount of data and preprocessing it by using text processing techniques and apply NLP technique on it. The dataset contains negative McDonald's reviews. *Get the data* here.

## Understanding the dataset:



The dataset contains 1525 observations of 10 variables:

1. **unit_id**: id of record
2. **golden**: value *FALSE*
3. **unit_state**: value *finalized*
4. **trusted_judgments**: value *3*
5. **last_judgment_at**: time. Example *2/21/15 0:36*
6. **policies_violated**: the type of policies, violated. Example: *RudeService
7. **policies_violated.confidence**: the confidence of policies, violated. Example: *1.00.66670.6667*
8. **city**: City name
9. **policies_violated_gold**: value *NA*
10. **review**: review detail

**Algorithms**

- Data Collection
- Data Preprocessing (Very Important Step)
- Data Exploration and Visualization
- Model building (Of course the interesting part!!)
- Model Evaluation

## Tools:

**NLTK:** Natural Language Toolkit, one of the leading tools for NLP, renders a whole set of programs and libraries to execute statistical and symbolic analysis in Python. This tool helps in separating a piece of text into smaller units (tokenization). Through this tool, you can recognize named entities and can tag some text. It is the leading tool of NLP and is easy to use.

**SpaCy:** This tool is a successor of NLTK. It comes with pre-trained statistical models and word vectors. It is a library created for use in Python and Cython. It supports tokenization for 49+ languages.it enables to break the text into semantic segments like articles, words, punctuation. It can be used for named entity recognition (NER) with pre-trained classes, recognizing dependencies in sentences. It provides the fastest and most accurate syntactic analysis than any NLP library.

**Text Blob:** This tool was designed based on NLTK. For the probationer, it is the best option to understand the complexities of NLP and designing prototypes for their projects. The tool enables sentiment analysis, tokenization, translation, phrase extraction, part-of-speech tagging, lemmatization, classification, spelling correction, etc.

**GenSim:** This service is designed for information extraction and natural language processing. It has many algorithms that can be deployed irrespective of the size of the collection of linguistic data. As it is dependent on NumPy and SciPy (Python packages for scientific computing), the user needs to install these two packages before installing Genism. The tool is extremely structured, and it has top-notch memory optimization and processing speed. It enables operating large text files even without loading the whole file in memory. Genism doesn’t require costly annotations or hand tagging of documents because it uses unsupervised models.

**CoreNLP** can be used to create linguistic annotations for text, such as:

- Token and sentence boundaries.
- Parts of speech.
- Named entities.
- Numeric and temporal values; dependency and constituency parser.
- Sentiment.
- Quotation attributions.
- Relations between words

**Communication:** The slides will be provided here, feel free to any pull requests besides details are provided at the readme of the project.