kjr.com.au

Client: The Australian Government's lead science and technology agency for defence

PROJECT: Training & Implementation of a Natural Language Processing (NLP) Classification model to categorise text topics

AWS (AMAZON WEB SERVICES) BUILD AND TOOL CONSUMPTION

TOOLS & TECHNOLOGY

ABOUT KJR

Our services are purposefully designed to provide a cohesive experience for organisations embarking on digital transformation. Our business aptitude is **your advisory**, our technical skills **are your project delivery** and our training roots **enables your team** to build upon success

+61 1300 854 063

BACKGROUND

The Australian Government’s lead agency responsible for applying science and technology to safeguard Australia and its national interest, required assistance to design and implement a solution for classifying large volumes of scientific documentation to improve document visibility and availability for wider use. This task was to be achieved through the development of a Natural language Processing classification model, that classifies scientific text topics.

PROBLEM

Like all Machine Learning development activities, the first step is to obtain the training data, and in this project, the initial requirement was to ingest the input data. KJR was engaged to develop a framework for the automated ingestion of > 25 million JSON files contained within the clients Scopus dataset, into a MongoDB database for model training.

SOLUTION

Firstly, our team established the MongoDB database, in which the clients requested and where the training data was stored. There were several constraints to be considered when designing the infrastructure architecture to host the resulting database, primarily the size of the data. The input data consisted of 25 million JSON documents that amounted to close to 500 gigabytes. Secondly, the source data needed to be efficiently accessible to where the Machine Learning model was going to be trained.

After the evaluation of several potential architectures our team installed MongoDB onto an AWS EC2Linux instance and uploaded the data. To reduce the number of failure points, the team bulk uploaded the data into an S3 bucket by leveraging the AWS Snowball service. Employing the use of AWS Snowball helped to significantly reduce the amount of effort required to upload data. The team tested the ingestion process on subset of documents to confirm that the process was efficient and resilient. After several iterations, the resulting process was robust and multi-threaded allowing the database to be populated and the model training to begin.

Python and the PyMongo library were used to develop the ingestion script, and with the dataset approaching 25 million distinct records, MongoDB indexes were established to ensure the efficient retrieval of data when required.

DELIVERABLES

- Ingestion of 25 million XML files (the GFI Scopus dataset) into a MongoDB database and the extraction of a training dataset and a testing dataset.
- Porting of the MongoDB database to the client’s PROTECTED network (along with the associated documentation).
- Model analysis and advisory.

KEY OUTCOMES

- Utilised MongoDB to manage high volume of data and analyse test results as it supports dynamic query.
- A pipeline of data that is fed to train and validate the model.

Client: The Australian Government's lead science and technology agency for defence

PROJECT: Training & Implementation of a Natural Language Processing (NLP) Classification model to categorise text topics

MODEL CREATION, ACCURACY & TIME SAVED

TOOLS & TECHNOLOGY

BACKGROUND

The Australian Government’s lead agency responsible for applying science and technology to safeguard Australia and its national interest, required assistance to design and implement a solution for classifying large volumes of scientific documentation to improve document visibility and availability for wider use. This task was to be achieved through the development of a Natural language Processing classification model, that classifies scientific text topics.

PROBLEM

After ingestion of the Scopus dataset into the MongoDB database, the business required he design and training of two Natural Language Processing (NLP) classifiers. The data used for training and then testing consisted of various subsets from the Scopus dataset.

SOLUTION

The Scopus dataset is structured into three layers; The top-level split into four (4) general categories, the middle into twenty-six (26) specific categories and the bottom layer into three hundred and forty-four (344) detailed categories, each covering a different scientific discipline.

DELIVERABLES

- Development of an environment in AWS utilising AWS SageMaker Suite tools.
- Development of an architectural blueprint for 2 classification models.
- Code for generating the classification model.
- The development of an NLP model that utilises multi-tag classification built off the businesses system framework.

KEY OUTCOMES

- The client was able to feed their highly technical textual data into the model and obtain the predicted topics.
- The models provided can predict a given text’s topic within a few seconds, a precision rate of 73%. In comparison, human assessors would take several minutes each, and a team of multiple domain specialists would be required to achieve similar or better precision.
- The client was satisfied with the model’s performance against their technical documentation set, which contained a vast array of technical topics.
- Implemented tools to analyse classifier data that will improve the model’s accuracy ongoingly.

ABOUT KJR

Our services are purposefully designed to provide a cohesive experience for organisations embarking on digital transformation. Our business aptitude is **your advisory**, our technical skills **are your project delivery** and our training roots **enables your team** to build upon success

+61 1300 854 063

Client: The Australian Government's lead science and technology agency defence

PROJECT: Training & Implementation of a Natural Language Processing (NLP) Classification model to categorise text topics

DETERMINING SUCCESS WHERE NO GROUND TRUTH EXISTS- HUMAN IN THE LOOP

TOOLS & TECHNOLOGY

ABOUT KJR

Our services are purposefully designed to provide a cohesive experience for organisations embarking on digital transformation. Our business aptitude is **your advisory**, our technical skills **are your project delivery** and our training roots **enables your team** to build upon success

+61 1300 854 063

BACKGROUND

The Australian Government’s lead agency responsible for applying science and technology to safeguard Australia and its national interest, required assistance to design and implement a solution for classifying large volumes of scientific documentation to improve document visibility and availability for wider use. This task was to be achieved through the development of a Natural language Processing classification model, that classifies scientific text topics.

PROBLEM

The business required the implementation of an NLP classifier model that classifies text accurately and quickly to augment and improve their team’s capacity to classify an ever-increasing volume of scientific literature.

ENSURING SUCCESS

SPEED

Measuring the model's success started with calculating the total time saved per document review. The document summaries (abstracts) used in the model training were on average 175 words in length. The average person reads over 250 words per minute and with this in mind a benchmark of one (1) minute per review was established as the time it would take to read each abstract. In comparison, the models developed by the team consistently processed the abstracts in ~ 10 seconds.

ACCURACY

While the model’s reading speed, when compared to the average human readers, was orders of magnitude different, its accuracy in identifying scientific specialties consistently is where the model demonstrated most benefit.

The model achieved a 73% precision rate (the percentage of model predictions matched the actual result) with processing rate of less than 10 seconds per abstract. It is unlikely that a single individual would know, in detail, all 344 Scopus categories; Therefore, the chance of a person being able to accurately specify which scientific category a randomly assigned document belongs to, at a speedy rate would be low. To get sufficient range of expertise, it was estimated that 10 readers would be required to achieve similar accuracy.

In effect the human efficiency would be 10 minutes per prediction compared to the model’s 10 seconds per prediction. Of course, the model predictions serve to assist human readers in identifying items of interest in a given field. A domain expert would be able to correct mistakes the model makes and help it improve over time. To assist in this process, a contextual explanation of the model’s output was provided, so that human readers can understand the model’s reasons for assigning a specific category.

TOOLS & TECHNOLOGY

aws

mongoDB

DELIVERABLES

- Structured a label-based accuracy matrix where the client can input the test result and it will calculate the model’s accuracy, precision and recall score.
- Implemented a Lime explainer which visualised the testing results explaining how the model predicted labels on what vocabularies.
- Provided the client data analysis, visualisation and management scripts which can be used to improve the model’s performance in the future.

KEY OUTCOMES

- Provided an accurate, consistent, and easily maintainable NLP framework.
- Implemented a classifier model that automatically classifies text to reduce time taken per classification; 1 minute to <10 seconds.
- Implemented a classifier model that performs at 73% precision in text classification across 28 different topic categories.

ABOUT KJR

Our services are purposefully designed to provide a cohesive experience for organisations embarking on digital transformation. Our business aptitude is **your advisory**, our technical skills **are your project delivery** and our training roots **enables your team** to build upon success

+61 1300 854 063