ALMA Downtime & Jira Text Analysis & Classification Wiki

In this Dataiku project, we analyze and classify downtime types and Jira tickets related to the ALMA observatory. The project is divided into several flow zones, each focusing on a different aspect of the analysis.

Flow Zones Overview

1. Jira EDA/Topic Modeling/NGrams
2. Downtimes EDA/Topic Modeling/NGrams
3. BERT w/ Downtimes
4. BERT w/ Jira Tickets - All Text
5. BERT w/ Jira Tickets - Summary Text Only
6. OpenAI API Summarization + BERT

Jira EDA/Topic Modeling/NGrams

In this flow zone, we conduct an exploratory data analysis of ICT and PRTSIR Jira tickets. We create basic visualizations to understand the distribution of data between the two types of tickets. We also generate trigrams, the top occurring three-word key phrases, for both ICT and PRTSIR tickets. The kernel used for this analysis is 'mypython'. Topic modeling for Jira tickets with the tool pyLDAvis can been seen in the project 'HaleyNLP' in the 'Jira' flow.

Downtimes EDA/Topic Modeling/NGrams

This flow zone focuses on the exploratory data analysis of downtime typres (weather, technical, scheduling) from the Shiftlog dataset. Using the 'mypython' kernel, we extract the year and create a new year column, visualize the data distribution over time and across downtime types, and clean the text by removing stopwords and punctuation. We then extract the top unigrams, bigrams, and trigrams for each downtime type and create word clouds to visualize the results. Topic modeling is also completed and visualized with pyLDAVis, and interactive LDA tool.

BERT w/ Downtimes

Here, we employ a BERT model to classify downtimes by subject and comments using multi-class classification. The kernel used is 'env mypython', and text cleaning and preparation is done before this step. We refer to resources such as Multi-label Text Classification using BERT and TensorFlow for guidance.

BERT w/ Jira Tickets - All Text

In this flow zone, we prepare the datasets for running the BERT classification on hardware and software tickets. We concatenate all text columns into a single column and label ICT tickets as 0 and PRTSIR tickets as 1. We then remove all unnecessary columns and clean the text using the Dataiku plugin. Numbers are not removed, as they may provide valuable insights into software vs. hardware problems.

BERT w/ Jira Tickets - Summary Text Only

In this flow zone, we perform the same preparation as in the previous zone, but only keep the summary column. We use the same text cleaning methods. For both groups, we run a BERT model to classify Jira tickets as either software or hardware. Since the training data is heavily imbalanced, we use balanced evaluation metrics to assess the results.

OpenAI API Summarization + BERT

In this flow zone, we utilized the OpenAI API, specifically the 'davinci-002' model, to summarize the text data from the Jira tickets. By taking advantage of the free token credits provided by OpenAI, we were able to generate summarized text for our analysis.

Once the text was summarized, we fed the processed data into the BERT model to classify Jira tickets as either software or hardware problems. This step allowed us to leverage the power of both the OpenAI API and the BERT model for more effective classification and analysis of the Jira ticket dataset.

This experiment was limited, and only to show how the concept can work in the future. More tokens are necessary for better summarization and classification.