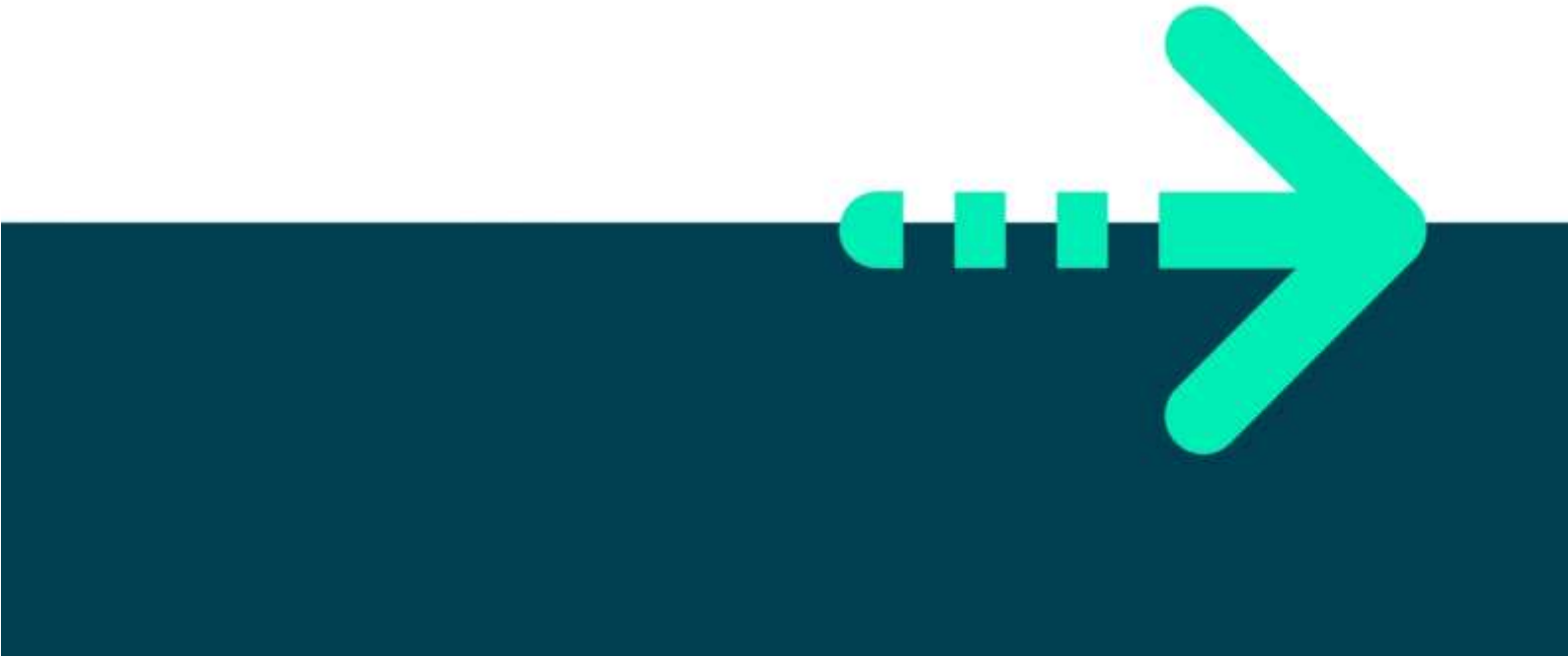
Lloyds Banking Group

Data Analyst Incubation

Sprint 2 : Natural Language Processing and Sentiment Analysis



Contents

Introduction

As Quay banking has conducted a historical analysis of its customer reviews and has often depended on branches to report in, the Customer Services Director has requested that the business utilise the merits of big data to become more data-driven, more responsive to customer feedback, better able to filter and draw insights and to utilise feedback from social media.

Business need

Sentiment analysis helps organisations understand customer feelings towards their products, identifying what they appreciate and what they dislike. This level of responsiveness to customer feedback is crucial for continuous improvement in customer services and financial products.

Analysing Sentiments can guide the development of new offerings that are designed according to customer appreciation. This provides a valuable repository of terms and preferred language that can guide our customer service training and marketing campaigns.

Email received (colleague handover)

From: RoryC@quaybanking.com
CC: LauraB@quaybanking.com
Subject: Handover - Tougher than I expected!

Long time, howve you beeeeeen?! I've only got 3days left on my notice, you coming for my leaving drinks? XD

Ive been trying to process this survey dataset, I've had about a day on it in Excel and managed to manually remove some stopwords like "it", "was", "when", filtered it to 1 and 5 stars (as the objectives stated) and tried my hand at stemming but i think im a little out of my depth here, haha! Anyway, Laura asked me to handover the project to you – so thanks truly, really appreciate your help on this.

I did manage to update the original file with what i did, no need to thank me! ;-)

Cheeeers!

Rory

Objectives

Continuing the work your colleague Rory started, you should process and analyse the qualitative survey data from 2015-2017, compiled from our Mortgage product and Online Services feedback forms.

- Analyse the frequency of significant words in the survey dataset. Our sales training and customer service training teams seek to understand the preferred language of our customers who felt strongly about a service, filtered to those that gave either 5 stars or 1 star on their feedback forms. The two areas of the business have different questions in this area – outlined below:

  1. Online Services feedback:
     - What were some of the issues that were common among customers trying to access our services via the customer PIN?
     - Are customers generally positive or negative when describing the interface and overall ease-of-use of our Online Services?

  2. Mortgage product feedback:
     - What was the overall picture of customer feedback regarding our interest rates?
     - Which branch received the most positive ratings and which branches received the most negative ratings?
     - Were there any branches that under reported?

- Use a pre-built Lexicon model, like Vader or Wordnet: to analyse the data and thus enrich the dataset with new metrics for each review around any of polarity, subjectivity, or neutrality.

- Build a Sentiment Analysis model, with the explicit purpose of rapidly processing and analysing reviews streamed via API from social media posts that include our bank's hashtag, separating ones that mention online services from those around mortgages.

  - Consider this question : would it be wiser to build two separate models based on the two surveys, or is there enough shared characteristics across the entire dataset to warrant a single robust model?

Supporting resources

If you haven't already done so, review the materials around Naïve Bayes and Amazon Food Reviews shared with you in

the github repository, within the Sprint_preparation folder. This includes a slide deck and notebook with data source:

1. Amazon Food Review Analytics.ipynb
2. naivebayes_sprintprep.pdf

Assignment detail

Task 1 – Customer access to Online Banking Service

Evaluate the negative reviews received from clients that concern access and ease of use issues. What features/words commonly come up in this context for Online Banking?

Task 2 – Mortgage Interest Rates and branch analysis

Evaluate mentions of Interest Rates in both positive and negative reviews. Evaluate this according to bank branches. What patterns can you observe in both?

Task 3 – Term Frequency

What words appear the most frequently in each of the following categories? Identify potential issues and counterintuitive results. Discuss this in your Summary Report.

- Mortgage Positive
- Mortgage Negative
- Online Positive
- Online Negative

Note: Present these results in an engaging and comparative manner.

Task 4 – Enriching the data

Using a Lexicon model from an NLP package, generate an additional column to enrich the data with a measure for either polarity, subjectivity, or neutrality

Task 5 – Sentiment Analysis Modelling

Build a Sentiment Analysis model that will be used for future predictive modelling. This will be used to analyse Social Media posts in real time to gauge the sentiment of people posting online about Quay Banking or even banking in general.

Analyse and discuss the *Most informative features(example below)* of your model in your Summary Report.

```
Most Informative Features
                   yuck = True              neg : pos    =     75.1 : 1.0
                    yum = True              pos : neg    =     42.5 : 1.0
              disgusting = True              neg : pos    =     28.1 : 1.0
                  awful = True              neg : pos    =     27.7 : 1.0
                   spit = True              neg : pos    =     27.5 : 1.0
                  trash = True              neg : pos    =     24.7 : 1.0
                 burned = True              neg : pos    =     24.5 : 1.0
              soothing = True               pos : neg    =     23.3 : 1.0
                 ruined = True              neg : pos    =     22.6 : 1.0
               inedible = True              neg : pos    =     22.0 : 1.0
                 rotten = True              neg : pos    =     21.6 : 1.0
                  worst = True              neg : pos    =     21.6 : 1.0
              satisfies = True              pos : neg    =     21.5 : 1.0
               terrible = True              neg : pos    =     21.5 : 1.0
                bedtime = True              pos : neg    =     20.6 : 1.0
```

Fig1. Most Informative Features for Amazon Food Reviews

Be sure to provide an appropriate evaluation of your model.

One working model and analysis would suffice for this project given the time. If you can manage it during the sprint, try building two models for a comparative analysis, as this would be an excellent opportunity for a deeper understanding of our customer survey data.

Deliverables

1. NLP and Sentiment Analysis Notebook(s)
2. Summary report, which includes:
    - Analytical Insights as required for each Task
    - All models and plots included in the analysis
    - Future Recommendations based on your findings
    - Data Source documentation
    - Challenges experienced and opportunities for future investigations

Self-assessment

You will now self-assess your performance of the task against the below points schema:

- Task 1 - 4 points
- Task 2 - 4 points
- Task 3 – 3 points
- Task 4 – 3 points
- Task 5 – 6 points

Potential total mark is 20

Delivery

When you have completed this scenario, share your notebook, summary report and score with your trainer. If you have scored over 17/20, be prepared to showcase your work briefly.

Regardless of how many tasks you completed, everyone should bring their findings and discussion points ready to share at the Round Table at the end of this sprint.