Lloyds Banking Group
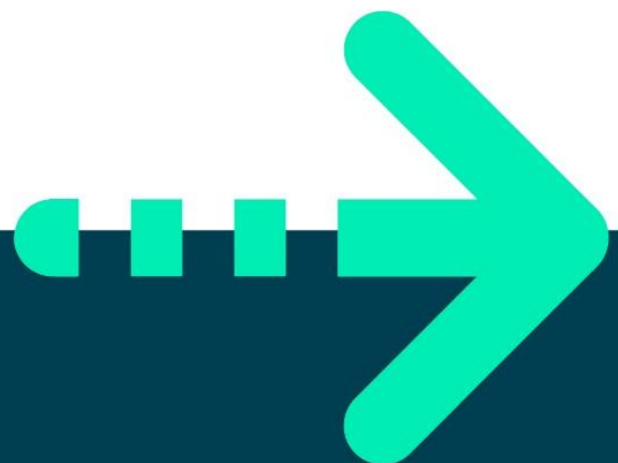
Data Analyst Incubation

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 **Customer Sentiments** and word preference 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 how we communicate the financial services our bank offers.

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:
    o What were some of the most common words and present an analysis of possible issues voiced by customers frustrated by accessing their online account?
    o What were the most informative words that customers used in their positive feedback; compare that to informative words used in negative when describing our Online Services?
      ▪ Advanced – include the use of TF-IDF (Term Frequency – Inverse Document Frequency)

  2. Mortgage enquiries feedback:
    o The most common words used in these reviews appear to be neutral, such as; service, rate, financial and staff. Select two and determine if the ratio they are used positive to negative reviews.
    o Gather context on these neutral words by extracting 3 words on either side of the neutral word, from each review. What are the most common words or phrases used to contextualise the neutral word?
    o Which branch received the most positive ratings proportionally to negative ratings?

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

- Build a Sentiment Analysis model, a proof-of concept model that can rapidly process and analyse the sentiment of social media posts that include our bank's hashtag. Build two separate models, one for Online Services, the other for Mortgage enquiries. Evaluate each with:
    o Confusion Matrix
    o Performance Metrics

o The most informative features discovered from the training data – (optional)

```
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

o Implement a series of test Social media posts for each model and the model's probability measure for the user's sentiment. For example:

  ▪ "this service was terrible, it used to be much more reliable" - (1):97% (5): 3%

o **Advanced** – Build a function that filters social posts that mention online services from those that refer to mortgages, then sends the post to the appropriate model. Evaluate how effective this was.

There will be a round table discussion at the end of the sprint to discuss your findings within the data, the effectiveness of your models and what could be improved for future data collection and analysis.

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

Roundtable Discussion of Analysis

1. Customer access to Online Banking Service

What features/words commonly came up in the context of access in Online Banking?

2. Mortgage Enquiries – context for neutral words

Discuss the insights derived from the included contexts of neutral words in positive and negative reviews. Does this provide a sufficient basis for determining bank branch performance in the regards to customer satisfaction?

3. Enriching the data

Evaluate whether this was a reliable method to enriching the data. Even if you had not managed to do so, is it a useful endeavour to include polarity and subjectivity measures?

4. Sentiment Analysis Modelling

What were the challenges, costs and risks to implementing such a model. What were the clear benefits of its operational success?

Deliverables

1. NLP and Sentiment Analysis Notebook(s)

2. Summary report, which includes:

   - Analytical Insights based on your notebook and discussions
   - Future Recommendations for customer surveys
   - Future recommendations for sentiment analysis modelling