**DS 344 Project**

**October 2023**

Jack Grimston and Hannah Quass

**Outline**

Section 1: Natural Language Processing

* 1. Problem and data description
  2. Data processing
  3. Modelling
  4. Summary

1.1 Problem and data description

In this section of the project our main aim is to explore data of house listings that we scraped online. The focus is doing sentiment analysis on the description of a house in R, and investigating the relationship between the sentiment of the description and the price the house is listed for. The data was scraped from Pam Golding’s website which can be accessed here: <https://www.pamgolding.co.za>

We decided to restrict our project to only look at houses located in the Boland area of the Western Cape. As well as the text description and price, we also scraped data about the number of bedrooms, bathrooms, garages, parking spaces, and erf size of each house. The data was scraped using Python’s Beautiful Soup package and is available in the excel file titled output.xlsx.

1.2 Data Processing

1.2.1 Data Preparation

To perform sentiment analysis, we first had to clean and tokenise the text data. This was done using R’s tidytext package. We made use of the “afinn” sentiment lexicon to get sentiment values ranging from -5 for very negative to 5 for very positive for each word. A score of 0 indicates a word with neutral sentiment.

1.2.2 Data Exploration

After some exploration such as plotting the most common positive and negative words, and observing individual usages of certain words, we found some interesting results that will guide the rest of the application:

* Many descriptions have no negative words, and most have far fewer negative than positive words.
* Some of the words labelled as negative do not carry a negative sentiment in their context. Notable examples are the negative word “die”, which is used to refer to Afrikaans place names such as “Die Laan” and “Die Boord”. Another example are the negative words “fire” and “alarm”, which actually form a bigram, and the context is that a house has a fire alarm!
* Some words that should be labelled positive are not included in the afinn lexicon and so have been assigned a sentiment of 0.

We manually adjusted the sentiments in the cases mentioned above, assigning the words “die”, “fire” and “alarm” neutral sentiments and assigning our own positive sentiment scores to some words not in the afinn lexicon such as "sublime", "exquisite", "captivating", etc. The subjectivity of these choices is a limitation to this step, but our own subjective scores still convey more information than a score of 0.

1.2.3 Sentiment Negation

Another important step in sentiment analysis is to deal with negations such as “not bad”, which clearly must not carry a negative sentiment. We dealt with sentiment negation by tokenising the data by bigrams, and then “flipping” (multiplying by -1) the sentiment of the second word in a bigram if the first word is a negation word.

An interesting case that we had to consider was the usage of the word “miss”. In the data exploration step we found that one of the most common negative words was the word “miss”, and on sampling some descriptions containing this word we found that the context it was most frequently used in was “don’t miss out on the opportunity to own this fantastic house, etc.”

1.3 Modelling

1.3.1 Developing features from the sentiments of house descriptions.

Next, we will develop a statistical application using the results of the sentiment analysis. We do this by developing features based on sentiment. These features are:

* The aggregate sentiment of the description weighted with the number of words in the description.
* number of positive words per description
* number of negative words per description
* proportion of positive words per description
* proportion of negative words per description
* description length

1.3.2 Investigating features in different price brackets

We looked at how the values of these features varied over different price ranges. The results did not show that the sentiment predictors vary greatly over different price ranges. The number of positive words increased slightly for more expensive houses, but so did the average description length, so the proportion of positive words remained roughly the same.

1.3.3 Linear regression using only sentiment features.

We performed a simple linear regression using all the above sentiment features to see whether they can accurately predict price. The model yielded a very low value of only 0.08.

1.3.4 Linear regression using sentiment and numeric features.

To try improving on the preceding result we used the other numeric features we scraped earlier and included them in the linear regression model. This gave a slightly better value of 0.67. The ranking of each feature based on its regression coefficient is as follows:

* proportion of negative words
* weighted sentiment
* proportion of positive words
* no. bathrooms
* no. garages
* no. bedrooms
* no. negative words
* no. positive words
* no. parkings
* description length
* erf size

We see a slightly strange result, which is that the variables with the largest regression coefficients are some of the sentiment predictors, which previously didn’t predict price well. There are several reasons for this behaviour, but the full exploration of them is beyond the scope of this report.

1.4 Summary

Our findings can be briefly summarised as follows:

* More expensive houses tend to have longer descriptions.
* The weighted sentiment over all houses does not vary m­uch with price.
* Estate agents very rarely use negative words when describing a house.
* The sentiment of a house’s description is not a good predictor for its price.
* Numeric features of the house such as the number of bedrooms, bathrooms, and garages are better predictors for price.

The most salient point of our findings is that text data of house descriptions is overwhelmingly skewed to use more positive than negative words, which makes sense given the context of trying to advertise the best features of a property. If accurate predictions of a house’s price are to be made, the sentiment of its description is not the best predictor to consider.

A possible extension of the application could be to encode such numerical features into the analysis of the text description, for example picking up bigrams like “has three bedrooms” or words giving hints as to the location and setting of the property which will all influence the price.