Group Assignment: Text Analytics (IB9CW0)

Dr Nikolaos Korfiatis
(n.korfiatis@warwick.ac.uk)
NKResearch Analytics consultants.
Associate Professor of Business Analytics
Warwick Business School, MSc in Business Analytics

Instructions

Please read the instructions carefully and discuss with your colleagues and provide an outline of your approach. Clarifications and/or any questions will be provided/answered **explicitly** through the module forum at **my.WBS**.

1. Overview and Pedagogical Goal

The goal of this assignment is to familiarize you with the complete process of *extracting*, *refining* and *delivering* insights of particular business value that are extracted from unstructured data of non-conventional size. You are going to work in a group of four (4) in order to extract insights from text sources in a particular business scenario – that of analyzing customer feedback. The assignment maps to level 7 qualification level and aims to establish you ability to handle *the development of in-depth and original solutions to unpredictable problems and situations*.

The assignment is structured in three (3) parts. The first part (Part A) covers your ability to construct and demonstrate the handling of text data. It aims to familiarize you with the principles of text mining, the bag-of-words model and the development of metrics that can be used to analyse structural elements of text, such as readability and keyword selection. The core of this assignment involves the translation of these insights to actionable features. The second and third parts (Part B and Part C) also involves the generation of features and in particular (a) polarity – whether the text under consideration is positive or negative, (b) sentiment – the extraction of affective states from the text and (c) the evaluation and extraction of important topics that are covered and elaborated by the authors of these texts and the insights that can be provided upon (Part C).

2. Marking Criteria and Weights

The marking criteria for all parts of the assignment are as follows:

- Part A: 20% - Completeness of the solution, efficiency of the code, interpretation of the results.
- Part B: 35% - Completeness of the solution, efficiency of the code, interpretation of the results.
- Part C: 35% - Completeness of the solution, efficiency of the code, interpretation of the results.

Peer assessment will count for 10% of the group grade.

3. Feedback

Feedback will be provided in individual sessions upon request with points for further improvement.

4. Submission Instructions

The assignment solutions should be submitted as one-file **pdf document** containing both the narrative for each part as well as the code in the form of a compiled R markdown notebook in html. The students should combine all files with a zip file with the following naming format:

group_number_X.zip

Where X is your given group number.

No other files are going to be considered. It is your group's responsibility to comply with the requirements of the submission, otherwise this will have repercussions for marking.

Part A: Construction of Corpus –Airbnb reviews

Airbnb provides data for its listings across various cities in the world. These can be downloaded and access from

`http://insideairbnb.com/get-the-data.html`

For each city (Example city: EXPLD) the following data files are provided:

Date Compiled	Country/City	File Name	Description
dd/mm/yyyy	EXPLD	listings.csv.gz	Detailed Listings data for EXPLD
dd/mm/yyyy	EXPLD	calendar.csv.gz	Detailed Calendar Data for listings in EXPLD
dd/mm/yyyy	EXPLD	reviews.csv.gz	Detailed Review Data for listings in EXPLD
dd/mm/yyyy	EXPLD	listings.csv	Summary information and metrics for listings in EXPLD (good for visualisations).
dd/mm/yyyy	EXPLD	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
dd/mm/yyyy	EXPLD	neighbourhoods.csv	Neighborhood list for geo filter. Sourced from city or open source GIS files.
dd/mm/yyyy	EXPLD	neighbourhoods.geojson	GeoJSON file of neighborhoods of the city.

Your group should download the **relevant data files for one city**. Using joins, you need to create a tidytext dataframe which can provide insights for the review text. Considering the current conditions of the pandemic you can choose to exclude the period after February 2020.

Your goal on this part is to generate as many variables / features as possible that can be related either with the rating score or the price that the property is listed for rent. For the text columns (reviews as well as renter description) you are requested to provide a bag-of-words analysis answering to questions such as:

a. What are the dominant words per aggregation category (neighborhood, access to public transport etc.)?
b. What are the most common word combinations used to describe a property listing?
c. What variables can be extracted from the text that can be related with the rating score ?

a. Is mentioning the name of the owner important?

d. Using the textual description of the property supplied by the owner, how does this relate with the price that the property is listed for rent?

Note: Considering that some of the reviews are not provided in English, you are requested to come with an efficient way to filter them based on a language identification procedure. It is recommded that

Part B: Sentiment association with Prices

Using polarity and sentiment you are asked to demonstrate how text derived features connect with the price and review satisfaction. You can use different aspects of sentiment such as affection categorization as well as the use of syntactical features such as the use of exclamation marks and/or capital letters. For both cases a regression model should be used to evaluate the predictability of individual ratings and listing prices against the variables obtained from structured parts of the dataset as well as those from the unstructured part of the dataset.

Part C: Topic Modelling and latent Dirichlet allocation

Using a topic model across different levels of aggregation you are requested to provide an analysis of the topics that become dominant across the different levels. Your analysis should focus on finding which topics become important for satisfied and dissatisfied customers as well as how particular characteristics of properties are more likely to influence a dominance of different themes in a customer review.

Your topic solution should evaluate among others:

a. The number of topics that need to be created for the particular corpus.

b. The additive predictability that some topics add on estimating the renting price and rating score.

All solutions need to be properly described and articulated by providing the relevant fragments of code.