RWTH AACHEN UNIVERSITY

The Second Assignment of IDS 2018-2019

Introduction

This assignment guides you through the analysis of three different data sets using the techniques and tools provided in the course. This assignment tests the understanding of the material in lectures 10-19. It is necessary to follow the assignment in the given order since the result of certain questions might depend on answers to previous steps.

Pre Processing

Dataset: The file "Online Retail.xlsx" contains a real-life data set which contains transactions which occurred in an online retail store. The store offers all-occasion gifts. Many customers of the company are wholesalers. The data set has the following attributes:

- `InvoiceNo: a 6-digit integral number uniquely assigned to each transaction.`
- `StockCode: a 5-digit integral number uniquely assigned to each distinct product.`
- `Description: Product name`
- `Quantity: The quantities of each product per transaction`
- `InvoiceDate: The day and time when each transaction was generated`
- `UnitPrice: Product price per item in sterling`
- `CustomerID: a 5-digit integral number uniquely assigned to each customer`
- `Country: the name of the country where each customer resides`

1. Invoices with a `InvoiceNo` starting with the letter 'c' are order cancellations. Take a look at the following exercises (clustering and association rules): based on these questions, would you recommend keeping the order cancellation in your data set? Give a reason for your answer and continue working with a data set that reflects your answer.

2. The attributes `Description` and `CustomerID` contain empty values. The `Country` attribute contains an "unspecified" value. For each of the three attributes reason how you would handle these values and why. Modify the data set according to your reasoning.

3. Explore into the attributes `Quantity` and `UnitPrice` by plotting each attribute visually. Do these attributes contain noise and/or outliers? If so, reason how you would handle them and modify your data set accordingly.

Please note, that this is a real-life data set and it is very likely that you come across even more data quality issues, while you answer the previous questions. Although we usually aim for a data set with the highest quality, for the scope of this assignment is sufficient to only modify what was asked in question 1-3.

Visualization

4. Create a stream graph that visualizes the number of purchases (invoices) per country over time.

   a. Modify the data set to only contain purchases made in the countries *Belgium, Ireland (EIRE), France, Germany, the Netherlands, Norway, Portugal, Spain and Switzerland*.
   b. Modify the data set such that it shows per month for each country how many purchases were made (i.e. how many invoices were created).
   c. Use the modified data to create a stream graph.
   d. Use this graph to compare the purchases made by each country. Can you find interesting differences between the countries? Are there certain times were the sales were particularly high/ particularly low?

5. Create a heat map that visualizes how much (in sterling) each country purchases per month.

   a. Modify the data set to only contain purchases made in countries *Belgium, Ireland (EIRE), France, Germany, the Netherlands, Norway, Portugal, Spain and Switzerland*. (Or use the version of the data set that you created for question 4 a).
   b. Modify the data set such that it shows per month how much money (in sterling) was spent in the shop per country.
   c. Use the modified data to create a heat map.
   d. Compare the amount of the purchases over time and between each country. Are there times where purchases are particularly high/ particularly low?

6. Compare the results obtained from the stream graph and the heat map. Is there a relation between the number of purchases and the amount purchased in sterling?

Clustering

7. Presume that the business analyst would like to cluster transactions with similar types of products into the same group (here don't consider the quantity of the products). For each product, only use its 'StockCode' to represent it. All the results here should be based on the preprocessed data set obtained from question 1 to 3 of this assignment. Presume that this obtained data set from question 1 to 3 has a variable name 'cluster_dataset' and is expressed by Pandas DataFrame in your code.

    a. Calculate and show the number of occurrences of each product in data set 'cluster_dataset'. For example, if a product appears in a transaction, then its occurrence number will be increased by 1 (do not consider the quantity of this product here). Preserve the 100 most frequent products and remove all the other products in 'cluster_dataset'. For example, if a row in 'cluster_dataset' contains unqualified product, then remove this row from 'cluster_dataset'. Show the new 'cluster_dataset' in your result.
    b. Based on question a, please reorganize the data from 'cluster_dataset' and generate a new data set 'cluster_dataset_new' which has a suitable format (for k-means) for solving the transaction clustering problem mentioned above. Show the data from 'cluster_dataset_new' by using Pandas DataFrame in your result, where the index should be consistent with the values of 'InvoiceNo', the column name should be consistent with the values of 'StockCode' and each element in this DataFrame should have a value 0 or 1.
    c. Try values 2, 3, 4, 5 for parameter 'n_clusters' for the k-means function from Scikit-Learn over the data set 'cluster_dataset_new' generated in question b. Show the 'within cluster variation' (also called 'sum of squared distances') of the generated clusters for each different setting for 'n_clusters' in your result. Also write down the value that you have tried for setting 'n_clusters' which can help generate the best clustering results and explain how you make this decision.

Frequent Itemsets and Association Rules

8. For the clusters output by k-means function with the best 'n_clusters' from question 7, the business analyst now would like to research on the frequent purchase behaviours and specific purchase rules for each cluster.

    a. Set the minimum support for finding the frequent purchase behaviours to 0.2. Please provide the business analyst with the qualified purchase behaviours. For each product, only use its 'StockCode' to represent it. Also show the data set prepared for each cluster for mining the frequent behaviours by using Pandas DataFrame in your result, the data set for the cluster k should have the variable name 'fpb_data_k' in your code.
    b. Furthermore, the business analyst would like to analyze the purchase behaviour of the citizens from 'United Kingdom' for each cluster. Specifically speaking, he wants to discover if there exist some rules which indicate that the citizens from 'United Kingdom' tend to buy some specific products for each cluster. Set the minimum support to 0.2,

minimum confidence to 0.7. Please discover and show such rules (only show the rules with 'United Kingdom' appearing in antecedents in the rules) for each cluster for the business analyst. Also show the data sets prepared for each cluster for mining the relevant rules by using Pandas DataFrame in you result, the data set for cluster k should have the variable name 'r_data_k' in your code.

Text Mining

The task at hand is to perform text classification on a given corpus. Specifically, the kind of classification you are going to perform is *stylometry*: automatically identifying the author from a collection of excerpts of their books.

Dataset: the files `pg_train.csv` and `pg_test.csv` are corpora extracted from the Project Gutenberg book archive. They contain, in CSV format, fragments of English text labelled with the author. Notice that the data files have *just two columns*: the target, and the corresponding text. They are separated by a '#'.

12. Obtain a *binary* document-term matrix and then train a classification model. This matrix should be obtained from the corpus after some preprocessing steps:
    a. All text lowercase
    b. No punctuation
    c. Tokenization
    d. Stemming
    e. Stopword removal (use the default nltk stoplist, see template)
    f. After these steps, use the obtained matrix to train a logistic regression classifier.
13. Obtain a document-term matrix of *counts*, using exactly the same preprocessing steps of the question 12. Train a logistic regression classifier (same hyperparameters) based on this matrix of counts.
14. Obtain a *tf-idf* document-term matrix using exactly the same preprocessing steps of the question 12. Train a logistic regression classifier (same hyperparameters) based on this matrix of tf-idf scores.
15. Use the training data documents to train a doc2vec embedding in order to reduce the dimension of the document vector to 300 - again, after the same preprocessing steps of question 12. Use the doc2vec model you just trained to convert the training set to a set of document vectors; then, use this set of labelled vectors to train a logistic regression classifier (same hyperparameters).
16. Assess the results of your models:
    a. Use the four models to obtain classifications for the test set. **Be careful:** the test data has to be preprocessed and converted in the same way as the training data!
    b. Obtain confusion matrices for the four different models.

c. Obtain accuracy and f1 score for the four different models.
d. Briefly comment on the quality of the predictions for the four models.

The choice of hyperparameters for the models (epochs, embedding space dimension, etc) is yours, and it should be adequate to the task. Notice that, as specified, the classifier should have the same hyperparameters throughout the four tests in order to compare the representations. You are allowed to use any functionality of the Python packages `scikit-learn`, `nltk` and `gensim`. **Hint:** refer always to the official documentation.

Process Mining

For this part, refer to the online docs of pm4py. You will find particularly of interest the documentation on filtering (https://pm4py.github.io/filtering.html, or on the new website http://pm4py.pads.rwth-aachen.de/documentation/filtering-logs/).

important: if you did not do it in the instruction, you should make sure to have the latest pm4py version: to get it is sufficient to type `pip install pm4py --upgrade` from any terminal emulator on Windows (command prompt, PowerShell, etc) or any terminal on *nix systems. For the details, refer to the study guide and the Process Mining instruction.

Dataset: The file "event_log.xes" is the event log which is provided for this section to use with the pm4py library.

17. Trace Frequency
    a. Use the provided event log and identify the least frequent traces and the most frequent traces.
18. Process Discovery and Conformance Checking using first filtered event log
    a. Remove the two least frequent traces and create a new event log out of the original event log without the two least frequent traces.
    b. Use Inductive miner algorithm to discover the process model based on you new event log (the filtered log without two least frequent traces).
    c. Do the token replay conformance checking using your discovered model and the original event log. Does your process model fit?
        i. Calculate the fitness of your model.
        ii. Are there any deviations between the process model and the event log?
19. Process Discovery and Conformance Checking using second filtered event log
    a. Now use the original event log and remove the two most frequent traces, and discover the model based on your new event log(the filtered log without two most frequent traces).
    b. Do the token replay conformance checking using your newly discovered model and the original event log. Does your process model fit?
        i. Calculate the fitness of your model?
        ii. Is there any deviation inside the process model?
20. Process Discovery using complete log

a. Use the complete event log (original event log) and discover your process model using inductive miner.
b. Do the token replay conformance checking using your newly discovered model and the original event log. Does your process model fit?
c. How are these three discovered process models different from each other? Which model is the best fitting to the original log? Why?

Deliverables

The deadline for the assignment is Sunday 20/01/2019 23:59 . You will need to hand in your submission via moodle . Note that there is no extension for the deadline and also late submissions will not be considered. Your submission should include a **jupyter notebook** , which presents your results and also contains the python code used to obtain the results.

Report requirements:

- Use the provided jupyter notebook

Grading

Participation in the assignment is one of the prerequisite for taking the written exam. The two assignments and the exam form a whole and it is not possible to retake parts of the course, i.e., the results of the assignments expire after the exam. Furthermore, assignments can only be redone in the next academic year.

The grade of this assignment counts 20% towards the final grade. There are six main sections in your assignment and one related to your report style:

1. The preprocessing section counts 15 percent towards your assignment grade.
2. The visualization part counts 15 percent.
3. The clustering section counts 15 percent towards your assignment grade.
4. The frequent itemset and association rule section counts 15 percent towards your assignment grade.
5. The text mining section counts 15 percent towards your assignment grade.
6. The process mining section counts 15 percent towards your assignment grade.
7. As a data scientist showing results is as important as what you have done; therefore, the report style determines 10 percent of your assignment grade

- please note that correctness of your code, its result and also accuracy of your explanation are important.