Xi'an Jiaotong-Liverpool University
西交利物浦大學

XJTLU Entrepreneur College (Taicang) Cover Sheet

Module code and Title	**DTS301TC Data Mining**
School Title	**School of AI and Advanced Computing**
Assignment Title	**Individual Project**
Submission Deadline	**Sunday, October 30th 23:59 (Beijing Time), 2022**
Final Word Count	**N/A**
If you agree to let the university use your work anonymously for teaching and learning purposes, please type **"yes"** here.	

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online). With reference to this policy I certify that:

- My work does not contain any instances of plagiarism and/or collusion.
  My work does not contain any fabricated data.

By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.

Scoring – For Tutor Use	
Student ID	

Stage of Marking	**Marker Code**	**Learning Outcomes Achieved (F/P/M/D) (please modify as appropriate)**			**Final Score**
		A	**B**	**C**	
1st Marker – red pen					
Moderation – green pen	**IM Initials**	The original mark has been accepted by the moderator (please circle as appropriate):			Y / N
		Data entry and score calculation have been checked by another tutor (please circle):			Y
2nd Marker if needed – green pen					

For Academic Office Use			**Possible Academic Infringement (please tick as appropriate)**	
Date Received	**Days late**	**Late Penalty**	☐ **Category A**	Total Academic Infringement Penalty (A,B, C, D, E, Please modify where necessary) _____
			☐ **Category B**	
			☐ **Category C**	
			☐ **Category D**	
			☐ **Category E**	

Xi'an Jiaotong-Liverpool University
西交利物浦大學

DTS301TC Data Mining

Individual Project

Deadline: Sunday, October 30th 23:59 (Beijing Time), 2022.

Percentage in final mark: 60%

Learning outcomes assessed: D, E

Late policy: 5% of the total marks available for the assessment shall be deducted from the assessment mark for each working day after the submission date, up to a maximum of five working days

Risks:

- Please read the coursework instructions and requirements carefully. Not following these instructions and requirements may result in loss of marks.
- The formal procedure for submitting coursework at XJTLU is strictly followed. Submission link on Learning Mall will be provided in due course. The submission timestamp on Learning Mall will be used to check late submission.
- Policies on Academic Integrity are strictly followed.

Overview

The objective of this project is to apply data mining techniques in a real-world dataset to gain a better understanding of real-world data mining applications. In this project, you need to identify one appropriate data mining problem from a twitter dataset and apply data mining algorithms to extract useful information from the dataset using R or Python. According to the learning outcome E, you are expected to do some independent study and research in this individual project.

Dataset

The project uses a Twitter dataset, which contains 204,820 short messages (tweets) collected from Twitter (https://twitter.com), during the period of 14th - 16th, April 2016, from various locations in the United States. The dataset contains many different topics, e.g., weather, leisure, sports, traffic, etc.

The dataset is stored in an EXCEL file and needs to be processed with your R or Python program. Each record (row) contains information about a tweet. The columns are explained as follows.

- Tweet Id – the ID of a tweet
- Date – the date on which a tweet is published
- Hour – the time when a tweet is published

- Username – name of a user
- Nickname – nickname of a user
- Tweet content – the actual message
- Favs – number of users who like the tweet
- RTs – number of users who re-tweet the tweet (republish)
- Latitude – latitude of the location where a tweet is published
- Longitude – longitude of the location where a tweet is published
- Followers – number of followers of a user (the values are the same for the same user)

Requirements and Tasks

You are allowed to use existing R or Python libraries to solve the following tasks.

T1 Statistic Analysis and Data Visualization:

T1-1: Find the top 10 tweets. Tweets should be ranked based on the sum of (1) number of users who like the tweet (Favs), and (2) number of users who re-tweet the tweet (RTs).

T1-2: Find the top 10 users. Users should be ranked based on the number of followers that they have.

T1-3: Draw a figure to show the number of tweets posted at different times of the day (i.e., 24 hours).

T1-4: Draw a figure to show the number of tweets posted from different locations (or states) in the US.

T2 Data Cleaning and Pre-processing:

T2-1: Raw tweets are highly unstructured and often contain redundant and problematic information. For instance, the links, emojis and symbols (e.g., #, @) in a tweet may not be necessary for the text mining tasks. Use R or Python to clean and pre-process raw tweets.

T2-2: Apply necessary text mining preprocessing techniques, e.g., tokenization, stemming, stop word removal, etc.

You can further pre-process the dataset based on the topic you choose in T3.

T3 Data Processing and Analysis:

Identify one data mining problem and use data mining algorithm(s) to extract useful information from the given dataset. You can choose your own topic. Please make sure your topic is appropriate and have some research value. Some potential topics are listed for your reference.

- Identifying trending topics on twitter
- Extracting tweets related to specific domains, e.g., traffic, weather, sports, etc.
- Social event detection from tweets
- Spatial and temporal analysis of tweets
- Topic modeling of twitter data
- etc.

Report

You need to write a report to show all the contents for this project. In general, the report must be in English and should include the following contents:

1. XJTLU Cover Sheet
2. Source code and results for **T1 Statistic Analysis and Data Visualization**. You can add one or two paragraphs to explain anything that is not obvious.
3. Source code for **T2 Data cleaning and preprocessing**. You need to give some examples to show the tweet content before and after data pre-processing. You can also add one or two paragraphs to explain anything that is not obvious.
4. For **T3 Data processing and Analysis** you should include the following contents:
    a. **Introduction**: State clearly what is the topic, why you chose the topic, how it is related to data mining, and discuss if there are some existing studies related to the topic.
    b. **Methodology**: State what data mining algorithm(s) you use to solve the problem, explain how to use it and identify the novelty of your method (if any).
    c. **Experiments**: Include your code and some brief explanation.
    d. **Evaluation**: Show all the results (e.g., tables, figures, etc.) you get from your method and give the corresponding explanation. You can also discuss the pros and cons of different models if you implemented multiple models for your topic.
    e. **Conclusion**: Summary of the results, list some current limitations and future directions.
    f. **Reference**

If you refer to any work from other sources, the original work must be cited.

Maximum 2500 words for the report. (Clarity and brevity are valued over length).

Submission

Electronic submission on Learning Mall is mandatory. You need to submit a zip file (named IDnumber_YourName_DTS301TC_FinalProject.zip (e.g.: *1900000_ZhangSan_DTS301TC_FinalProject.zip*)) containing all your source code in R or Python and your report in pdf format.