به نام خدا

سوالات درس تجزیه و تحلیل داده های حجیم

امتحان میان ترم، ترم اول ۱۴۰۱

لطفا موارد زیر را در پاسخ و ارسال، دقیق رعایت نمایید.

در برگه های پاسخ اسم دانشجو نوشته شود. اگر جواب سوالی از برگه دو یا چند نفر مشابه و یا مثل هم باشد نمره برای آنها منظور نمی شود. یا اینکه احساس شود جواب سوالی از جایی کپی و یا کمک گرفته شده باشد که دانشجو بلد نبوده علاوه بر در نظر نگرفتن نمره، در جلسه حضوری زمان ارائه پروژه آن سوالات به طور شفاهی پرسیده می شود. مطالبی که می نویسید باید دقیق یاد گرفته باشید.

جوابها، تایپی باشد و همه جوابها **در یک فایل** با فرمت فشرده با شرایط زیر ارسال شود. کد برنامه ها ی نوشته شده همراه با توضیح آنها، شرح الگوریتمها و نحوی پیاده سازی/کتابخانه ها پیوست شود که مورد بررسی و تست قرار گیرد.

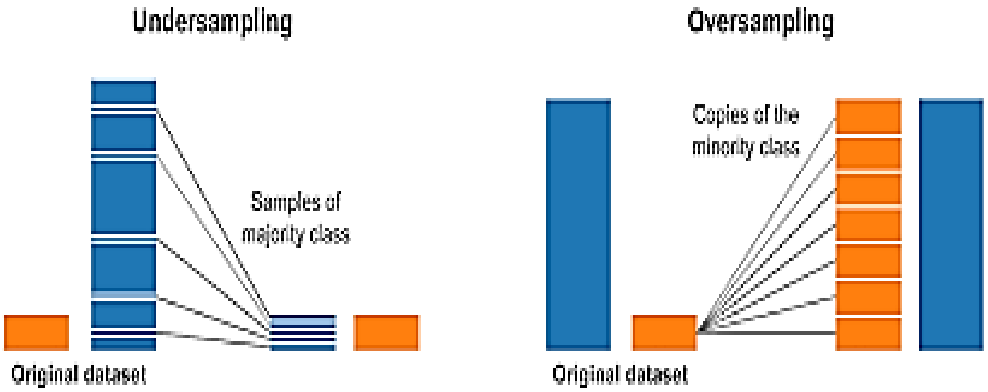
اسم فایل به اسم دانشجو - تاریخ مثل: Ali_Taghavi_28_8_1401.zip

فایل فوق، **فقط یک بار** برای TA کلاس ارسال شود. یا از طریق ایمیل یا واتس آپ.

اسم: علی تقوی درس: تجزیه و تحلیل داده های حجیم: ۱۴۰۱/۸/۲۸ فایل ارسال می گردد.

Release Date: Saturday 28/Aban, Due Date: Wednesday 2/Azar Time:12:00PM
Grade: 6 Points. Our answer will be released on Thursday 3/Azar.

1- Which concepts are shown in this following figure? And where these concepts and techniques are used?



2- What is CAP theorem and why it is important?

3- Answer the following questions related to data cleaning with considering the attached dataset using Python programming language.

The dataset which we have provided is for a social media and it is raw data. That means that data is not cleaned. We first should clean it. As you know there are many techniques for data cleaning which some of them have been taught in the class. A good reference (book) for this purpose has also been uploaded in the university portal and you can use it.

Some necessary steps for mentioned dataset that should be applied for data cleaning are mentioned here, from (a) to (l).

1) Write programs with proper data cleaning technique for each step. from (a) to (l).

   a) Remove newlines and Tabs

   b) Remove Punctuation/ Unicode characters/ Special Characters

   c) Hashtag removes

   d) Tokenization

   e) Stop words

f) Remove URLs

g) Remove HTML tags

h) Repeated characters reduction, for example: Hellllo → Hello

i) Remove capitalization/ Case normalization

j) Remove Whitespaces, for instance, He llo → Hello

k) Typo Correction/ Misspelled words: big “dada” → big “data”;

l) Stemming or lemmatization

2) Are there any other data cleaning steps that have not been mentioned from (a) to (l), write them? Why do you think these steps are necessary? Clarify your reasons with examples and references.

3) Please note that for (k) part, are there any tools/libraries/ and etc to simplify/ to increase accuracy of typo correction? The goal is increasing the accuracy and performing with an optimal way. Good Luck