INTERNATIONALE
HOCHSCHULE

PORTFOLIO

Assignments for the course: Project: Data Analysis (DLBDSEDA02)

TABLE OF CONTENTS

1. TOPICS AND TASKS

Imagine that you work for a municipality that wants to address citizens' worries and complaints better. In the first workshop with decision-makers, politicians, representatives of active associations, and NGOs, it has become apparent that the source of unsatisfaction of many citizens seems to be that past decisions did not address currently discussed topics. There is a system for citizens to submit complaints directly online to the local administration, and this system has gathered a considerable number of written complaints. The drawbacks of these data are that a few individuals seem to use this system extensively, making the data not representative of the entire population. Also, the data is unstructured and too vast in numbers for your colleagues to oversee systematically. From the workshop with local stakeholders, we know that messaging services are extensively used to utter complaints by many citizens who do not use the municipality's online complaint system. Accordingly, these data should also be considered in the decision-making processes in the near future.

Within the framework of this course, one of the following topics must be selected.

Note on copyright and plagiarism:
Please take note that IU Internationale Hochschule GmbH holds the copyright to the examination tasks. We expressly object to the publication of tasks on third-party platforms. In the event of a violation, IU Internationale Hochschule is entitled to injunctive relief. We would like to point out that every submitted written assignment is checked using a plagiarism software. We therefore suggest not to share solutions under any circumstances, as this may give rise to the suspicion of plagiarism.

1.1. Task 1: Use NLP techniques to analyze a collection of texts

In this task, you will use NLP techniques in Python to analyze texts comprising complaints regarding decisions made by the local municipality. The data are unstructured, not allowing for direct systematic analysis. In addition, the number of complaints makes overlooking the most pressing issues an intricate task. Your goal is to extract these most frequently addressed topics from the written texts, providing decision-makers with this information.

Task: Use NLP techniques to analyze a collection of texts

Conduct this analysis in the following 3 phases:

1.1.1. Conception phase

This phase represents the most important part of the process. Anything that is overlooked or forgotten in this phase has a negative effect on the implementation later and will lead, in the worst case, to useless results.

The first step is to create a written concept to describe everything that belongs to the data analysis workflow. **This step is perhaps the most important of the entire process**. It is crucial to take enough time for this phase **BEFORE** the next steps can be taken. **It is therefore essential to follow the sequence of the respective steps carefully.**

1. Choose an appropriate sample data source. Open datasets can be found, e.g., on www.kaggle.com or https://github.com/owid. Search for "complaints", "comments", "NLP", and similar terms. This project is about practicing NLP techniques in general. Accordingly, you should choose the sample data to match the described use case as close as possible, but you might also choose an appropriate dataset from other domains, e.g., customer complaints, online comments, product reviews, etc. Inspect the data and include its structure in your considerations for the following steps.

2. Describe how you plan to preprocess the data to achieve "clean texts".
3. The data comprise unstructured texts. Briefly describe at least two approaches to convert these data into numeric vectors that can be used as inputs for the following step.
4. Once the texts are converted to numerical vectors, you will extract the most prevalent topics. Briefly describe at least two techniques that you will apply to achieve this.
5. For each step, list the Python libraries that you plan to use.

A conceptual text (1/2 DIN A4 page) has to be prepared for the submission, explaining these thoughts and considerations. The concept will be submitted as **PDF file**. The text field inside the PebblePad submission form can be left empty.

Throughout the process, online meetings provide an opportunity to talk, share ideas and/or drafts, and obtain feedback. In the online meetings, exemplary work that has been previously submitted can be discussed with the tutor. Here, everyone has the opportunity to get involved and learn from each other's feedback. There are also other channels available for you to address questions to the tutor and/or to your fellow students, such as the CourseFeed. Through the latter, you will obtain feedback, tips, and advice before submitting your results of this phase. **It is recommended to make use of these channels to avoid errors and to make improvements.** You submit your work after making use of all available informative media. This will be followed by feedback from the tutor, and the work in the second phase can begin.

1.1.2. Development phase/reflection phase

In this phase, you will **conduct the data analysis** based on the concept from the conception phase with the help of the selected Python libraries. You will probably run into some issues at the beginning of this phase. This is intended and part of the learning cycle. To tackle these issues, consult the documentation of the used Python libraries you chose in the conception phase.
Here, in the development phase, the actual data analysis begins:

- Install all required dependencies, such as Python libraries. It is recommended to use virtual environments with "venv" or "conda" and export the dependencies as a TXT or YAML file.
- Set up an open GitHub repository to store the code in a version-controlled way.
- Load the sample text data obtained in the conception phase into your Python environment.
- Preprocess the data to achieve "clean text" to be further analyzed.
- Use at least two different techniques to vectorize the texts and briefly compare the results.
- Extract the most prevalent topics discussed in the texts using at least two different semantic analysis techniques.
- Finally, briefly discuss your results.

An explanation of the procedure is submitted as **text (1/2 DIN A4 page)**. The file should contain the link to the GitHub repository. Furthermore, the steps of this phase should be described briefly. The explanation will be submitted as **PDF file**. The text field inside the PebblePad submission form can be left empty.

Again, it is recommended to use the provided channels to avoid errors and improve your work. Once this is done, you can hand in your second phase results for evaluation. Following feedback from the tutor, your work on the final draft will continue in the third phase.

1.1.3. Finalization phase

In the finalization phase, the goal is to **optimize the data analysis workflow** after receiving feedback from the tutor and completing the task. Certain elements may have to be improved or changed again.

You create a **full abstract (2 DIN A4 page)** describing the solution of the task in terms of content and concept, presenting a short break-down of the technical approach in a clear and informative way. Here, you should also briefly discuss the quality of the data and the results of your analysis. This abstract should also include an outlook as to how the procedure can further be improved, as well as a personal reflection about the project: What did you learn on a personal level? What were the problems you ran into, and how did you solve these issues? What are your problem-solving strategies for similar upcoming projects?

In addition, you upload the **finished product** as a PDF file with a link to the GitHub repository containing the fully functional code, together with a short technical description how to use the code. A zip folder is not necessary in this course.

Also, in the finalization phase, **it is recommended to use these channels to avoid errors and improve your work.** After submitting the third portfolio phase, the tutor submits the final feedback, which includes evaluation and scoring within six weeks.

1.2. Task 2: Extract prevalent topics from Twitter messages

In this task, you will extract the most frequently discussed topics on Twitter concerning your city or region. You will use Python for this analysis by connecting to the Twitter API.

Task: Extract prevalent topics from Twitter messages

Implement the data system in the following 3 phases:

1.2.1. Conception phase

This phase represents the most important part of the process. Anything that is overlooked or forgotten in this phase has a negative effect on the implementation later and will lead, in the worst case, to useless results.

The first step is to create a written concept to describe everything that belongs to the data analysis workflow. **This step is perhaps the most important of the entire process**. It is crucial to take enough time for this phase **BEFORE** the next steps can be taken. **It is therefore essential to follow the sequence of the respective steps carefully.**

1. Create a developer account on Twitter and safely retrieve your credentials to connect to the Twitter APIs.
2. Set up an open GitHub repository to store the code in a version-controlled way.
3. Install all required dependencies, such as Python libraries. It is recommended to use virtual environments with "venv" or "conda" and export the dependencies as a TXT or YAML file.
4. Concisely describe the steps of your planned data analysis workflow. This should include:
    a. Retrieving relevant tweets for your city or region.
    b. Preprocessing of the data to obtain "clean text".
    c. Entity analysis steps regarding frequently used hashtags and most active users.
    d. Appropriate NLP techniques to extract the five most commonly discussed topics.
    e. A justification of your choice of methods.

A conceptual text (1/2 DIN A4 page) has to be prepared for the submission, explaining these thoughts and considerations. The concept will be submitted as **PDF file**. The text field inside the PebblePad submission form can be left empty.

Throughout the process, online meetings provide an opportunity to talk, share ideas and/or drafts, and obtain feedback. In the online meetings, exemplary work that has been previously submitted can be discussed with the tutor. Here, everyone has the opportunity to get involved and learn from each other's feedback. There are also other channels available for you to address questions to the tutor and/or to your fellow students, such as the CourseFeed. Through the latter, you will obtain feedback, tips, and advice before submitting your results of this phase. **It is recommended to make use of these channels to avoid errors and to make improvements.** You submit your work after making use of all available informative media. This will be followed by feedback from the tutor, and the work in the second phase can begin.

1.2.2. Development phase/reflection phase

In this phase, you will **conduct the data analysis** based on the concept from the conception phase with the help of the selected Python libraries. You will probably run into some issues at the beginning of this phase. This is intended and part of the learning cycle. To tackle these issues, consult the documentation of the used Python libraries you chose in the conception phase and of the Twitter APIs.
Here, in the development phase, the actual data analysis begins:

- Use your credentials to connect to the Twitter API using Python.
- Load an appropriate number of tweets relevant to your city or region.
- Preprocess the data according to the steps described in the conception phase.
- Conduct two entity analysis steps to extract the most prevalent hashtags and most active users. Briefly discuss your results and their implications for the further steps.
- Use appropriate NLP techniques described in the conception phase to extract the five most prevalent topics in the tweets.
- Finally, briefly discuss your results.

An explanation of the procedure is submitted as **text (1/2 DIN A4 page)**. The file should contain the link to the GitHub repository. Furthermore, the steps of this phase should be described briefly. The explanation will be submitted as **PDF file**. The text field inside the PebblePad submission form can be left empty.

Again, it is recommended to use the provided channels to avoid errors and improve your work. Once this is done, you can hand in your second phase results for evaluation. Following feedback from the tutor, your work on the final draft will continue in the third phase.

1.2.3. Finalization phase

In the finalization phase, the goal is to **optimize the data analysis** after receiving feedback from the tutor and completing the task. Certain elements may have to be improved or changed again.

You create a **full abstract (2 DIN A4 page)** describing the solution of the task in terms of content and concept, presenting a short break-down of the technical approach in a clear and informative way. This abstract should also include an outlook as to how the procedure can further be improved, as well as a personal reflection about the project: What did you learn on a personal level? What were the problems you ran into, and how did you solve these issues? What are your problem-solving strategies for similar upcoming projects?

In addition, you upload the **finished product** as a PDF file with a link to the GitHub repository containing the fully functional code, together with a short technical description how to use the code. A zip folder is not necessary in this course.

Also, in the finalization phase, **it is recommended to use these channels to avoid errors and improve your work.** After submitting the third portfolio phase, the tutor submits the final feedback, which includes evaluation and scoring within six weeks.

2. TUTORIAL SUPPORT

In principle, several channels are open to attain feedback for the portfolios. The respective use is the sole responsibility of the user. The independent development of a product and the work on the respective portfolio parts is part of the examination performance and is included in the overall assessment.

On the one hand, the tutorial support provides feedback loops on the portfolio parts to be submitted in the context of the conception phase as well as the development and reflection phase. The feedback takes place within the framework of a submission of the respective part of the portfolio. In addition, regular online tutorials are offered. These provide you with an opportunity to ask any questions regarding the processing of the portfolio and to discuss other issues with the tutor. The tutor is also available for technical consultations as well as for formal and general questions regarding the procedure for portfolio management.

Technical questions regarding the use of "PebblePad" should be directed to the exam office via mail.

3. EVALUATION

The following criteria are used to evaluate the portfolio with the percentage indicated in each case:

Evaluation criteria	Explanation	Weighting
Problem Solving Techniques	*Capturing the problem *Clear problem definition/objective *Understandable concept	10%
Methodology/Ideas/Procedure	*Appropriate transfer of theories/models *Clear information about the chosen Methodology/Idea/Procedure	20%
Quality of implementation	*Quality of implementation and documentation	40%
Creativity/Correctness	*Creativity of the solution approach *Solution implemented fulfils intended objective	20%
Formal requirements	* Compliance with formal requirements	10%

The design and construction of the portfolio should take into account the above evaluation criteria, including the following explanations:

Problem Solving Techniques: Correct reflection and implementation of data engineering concepts such as reliability, scalability, and maintainability. Demonstration of independently tackling technical issues. Following appropriate problem-solving strategies and independently coming up with solutions.

Methodology/Idea/Procedure: Correct usage of technical frameworks and reasonable argumentation for technical choices.

Quality of implementation: Fulfillment of the technical requirements. Reproducibility of the created product in the form of a functional code provided in a GitHub repository. Concise and comprehensive documentation.

Creativity/Correctness: Creative approach to the problems and correct task fulfillment by the conducted analysis.

Formal requirements: Compliance with the formal requirements for submission (see below).

4. FORMAL GUIDELINES AND SPECIFICATIONS FOR SUBMISSION

4.1. Components of the examination performance

The following is an overview of the examination performance portfolio with its individual phases, individual performances to be submitted, and feedback stages at one glance. A template in "PebblePad" is provided for the development of the portfolio parts within the scope of the examination performance. The presentation is part of this examination.

Stage	Intermediate result	Performance to be submitted
Conception phase	Portfolio part 1	• Concept presentation in written form, which concisely shows that you thought about the requirements listed in the task description (200 words; approx. 1/2 page as PDF-file)
		Feedback
Development phase/ reflection phase	Portfolio part 2	• Explanation of the implementation in written form, including a link to a Git Hub repository (200 words; approx. 1/2 page as PDF-file)
		Feedback
Finalization phase	Portfolio part 3	• The full abstract as an attached 2-page PDF file including the project documentation as well as a technical and personal project reflection. • The final product as a PDF file with a link to the GitHub repository and a short description about how to use the code • Result from phase 1 (can be revised in the meantime) • Result from phase 2 (can be revised in the meantime)
		Feedback + Grade

4.2. Format for Digital File Submission

Conception phase

Recommended tools/software for processing	Git Hub, IDE of choice (VS Code)
Permitted file formats	PDF
File size	-
Further formalities and parameters	Files must always be named according to the following pattern: **For the performance-relevant submissions on "PebblePad":** Name-FirstName_MatrNo_ Course _P(hase)-1_S(ubmission) Example: Mustermann-Max_12345678_ DataAnalysis_P1_S

Development/reflection phase

Recommended tools/software for processing	Git Hub, IDE of choice (VS Code)
Permitted file formats	PDF
File size	-
Further formalities and parameters	Files must always be named according to the following pattern: **For the performance-relevant submissions on "PebblePad":** Name-FirstName_MatrNo_ Course _P(hase)-2_S(ubmission) Example: Mustermann-Max_12345678_ DataAnalysis_P2_S

Finalization phase

Recommended tools/software for processing	Git Hub, IDE of choice (VS Code)
Permitted file formats	full abstract as PDF + final product as PDF (with link to GitHub)
File size	as small as possible
Further formalities and parameters	Files must always be named according to the following pattern: **For the performance-relevant submissions on "PebblePad":** Name-FirstName_MatrNo_ Course _P(hase)-3_S(ubmission) Example: Mustermann-Max_12345678_ DataAnalysis_P3_S

4.3. Format of Abstract

Length	2 pages of text
Paper size	DIN A4
Margins	Top and bottom 2cm; left 2cm; right 2cm
Font	General Text - Arial 11 pt.; Headings - 12 pt., Justify
Line Spacing	1,5
Sentences	Justified; hyphenation
Footnotes	Arial 10 pt., Justify
Paragraphs	According to mental structure - 6 pt. after line break
Affidavit	The affidavit shall be made in electronic form via "myCampus". No submission of the examination performance is possible before it. Please follow the instructions for submitting a portfolio on "myCampus".

If you have any questions regarding the submission of the portfolio, please contact the exam office via mail.

Please also note the instructions for using PebblePad & Atlas!

Good luck creating your portfolio!