Deree
School of Graduate and Professional Education

THE AMERICAN COLLEGE OF GREECE
Pierce
Deree
Alba
1875

Module	**ITC 6110 – NATURAL LANGUAGE PROCESSING (NLP)**		
Term	**SPRING SEMESTER 2024**		
Assessment	**GROUP PROJECT**	**Weight**	**50%**
Duration	*You can use the duration of the Term to complete the project*		
Deliverables	*1. Report (document) submitted via TurnitIn* *2. Code on Blackboard or GitHub* *3. An oral presentation of your work*		
Method of Submission	*TurnitIn, Blackboard, GitHub*		
Deadline:	*13th Week* *US Grading scale*		

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*

General Instructions

Your project involves generating a series of experiments, extracting findings and observations from your experiments, and drawing conclusions. Essentially you will need to collect data (or they will be provided to you by the instructor) and use Python as the programming language along with any appropriate libraries to process the data, generate graphs, implement models, and test results. Tables, diagrams, and data visualizations are essential for presenting your findings.

Deliverables: a) Python code submitted on Blackboard or uploaded on GitHub, along with any optional instructions for running it b) a report of ~5,000±500 words that will present your findings, which will be submitted via TurnitIn. The report must be self-contained: all experiments performed along with all conclusions drawn should be reported. If you need to exceed the word limit, use an appendix. c) an oral presentation, where all members of the team need to present.

NOTE: There is zero tolerance on plagiarism, and ChatGPT is NOT allowed for the solutions and report.

Team size: 2-4 people (sent to the instructor by email before Week 3)

Grading: peer-review

Teams consist of 2-4 persons. Group members will be asked to rate the relative contribution of themselves and the other group members. The ratings provided by each member must add up to the number of persons the group consists of (see example below). These ratings will be taken into account in the final grading of the project for each individual.

		Person being rated		
		Person-1	**Person-2**	**Person-3**
Person doing the rating	**Person-1**	1.25	1	0.75
	Person-2	1.10	1.10	0.80
	Person-3	1	1	1
Average Rate		**1.12**	**1.03**	**0.85**
Individual score (project grade: 80%)		**89.6**	**82.4**	**68**

Example: In a group consisting of three members, each member provides a rating of all group members. As this is a three-member group, the ratings provided by each member add up to 3.00. A rating of 1.00 means that the person in question did exactly as much as expected of him/her. A rating that is less than 1.00 means that the person in question did less than expected, whereas a rating that is greater than 1.00 means that this person's contribution was greater than expected.

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*

Project guidelines

The objective of this project is to illustrate the practical usage of a range of techniques throughout the NLP pipeline. This project provides hands-on experience in NLP, covering all the various stages from data preprocessing to model development and testing. It also offers opportunities for further exploration and enhancement across various techniques (from simplistic to advanced).

Your solutions should adhere to the following steps and guidelines in order to be marked. Grading will also be relevant to the effort, level of difficulty and advancements made in each step of the process. All steps outlined are mandatory unless explicitly stated as optional or extra.

1. *Data Collection (10%)*
   Gather a dataset tailored to meet the requirements of the group project. The dataset needs to be able to address the model building exercises outlined below in Step 4.

   - The dataset can originate from any public or private source, and can feature any type of textual data (for example, it can cover movie reviews, ratings, news, articles, among others). You will need to describe your data in your report and presentation.

   - The data can be in any language of your choice (ideally in English). If the raw data contain multiple languages, extra pre-processing & data handling steps may need to be performed.

   - The dataset ideally needs to be labelled to conduct supervised learning tasks, specifically classification (such as sentiment analysis or topic classification; sentiment analysis is considered a subset of topic classification). If the data are unlabeled, you may need to either manually label them or find alternative ways to address this issue (extra points!). The latter is an active field of research so you are encouraged to investigate (time-permitting).

   - You may need to experiment with and adjust, if needed, the size of your data, while maintaining adequate historicity and diversity, to ensure it will run smoothly on your machines without any problems. Any adjustments must be explained in your report.

2. *Data Preprocessing and Normalization (10%)*

   a. As with any other ML pipeline, deal with missing and duplicate values, inconsistencies, input errors, text formats, outliers, etc.

   b. Normalize your input text data. Apply, where appropriate, the removal of stop words, punctuation, and special characters. Conduct spell checking and typo corrections. Use regular expressions (RegEx) or other techniques to fix any erroneous data, contractions, typos, and more. If in existence, handle appropriately any URLs, HTML tags, emojis, emoticons, etc. If the data contain personal, PII or sensitive information, it also needs to be removed, handled or anonymized. Deal with frequent and/or rare words if/where needed. Experiment with and decide upon using (or not) stemming or lemmatization. Finally, apply tokenization prior to converting the data into numerical format for the next step.

   c. Justify all your pre-processing steps in the report, code and presentation

3. *Feature Engineering (Embeddings) & Text Visualization (15%)*

   a. Extract features from the text data: this step involves converting the text data into numeric format. This can be achieved using various techniques like Bag of Words, TF-IDF, or word embeddings (Word2Vec, GloVe), Deep Learning techniques and more. Justify your choice of embedding(s) in your report and presentation. (Optional) experiment also with N-grams.

   b. Build a solution where you provide a word from your corpus that returns the N most similar words (5-10 words will suffice). There are multiple solutions and techniques to address this.

   c. Visualize the text embeddings in a scatterplot using techniques such as t-SNE (preferrable) and/or PCA. Colour the findings appropriately or add supporting text information or labels in the plots to detect any groups/clusters, patterns, trends, similarities, etc.

   d. (Extra points/Optional) Ideally, compare multiple embedding techniques; discuss their results, overall performance, similarities, dissimilarities, advantages and disadvantages.

   e. (Extra points/Optional) Feel free to use any other static or interactive visualization techniques of your choice to depict and discuss your data and findings (e.g. word frequencies)

4. *Model Building (40%)*

   a. UNSUPERVISED LEARNING (10%)

      i. Topic modelling: Apply (unsupervised learning) topic modelling using at least one algorithm of your choice to identify topic clusters, groups with similar words and hidden semantic patterns within the body of your text. The topic modelling algorithm can be as simplistic or as advanced as you wish (e.g. LDA, BERTopic).

      ii. Create supporting visualizations, extract key words per cluster (in a tabular format or visualization), and discuss your findings.

      iii. (Extra points/Optional) Ideally, you can perform a comparison between various topic modelling techniques and discuss their overall performance.

   b. SUPERVISED LEARNING (30%)

      i. Task 1 (10%): Text classification (topic classification or sentiment analysis):

         a. Conduct text classification (the label can be a topic such as spam/ham or drama/comedy or sentiment such as positive/negative) using at least (a) one traditional ML technique and (b) one DL model (can be a custom solution, a pre-trained model with transfer learning, etc.). Compare the models and discuss the results.

         b. Apply XAI techniques (e.g., LIME, SHAP) to explain your findings for one of your classification models on either a local or global basis.

ii. Task 2 (10%): Named Entity Recognition (NER): create one (or optionally more) model(s) for Named Entity extraction (you can extract e.g., persons, locations, organizations, monetary sums, time, dates, product names, and many more).

You will need to create your own entity labels depending on your input data; you can create as many labels as you like to address this problem but you will need to use at least 3-5 distinct labels (entities) to make it interesting (data dependent).

You may also need to create manually labelled training examples in formats appropriate for ingestion by your model. There are a lot of online articles on how to perform this step for NER.

You can use any NER model(s) of your choice, ranging from off-the-shelf easy-to-use solutions to powerful state-of-the-art pre-trained models and transfer learning. You will need to justify your choice and findings in your report and ppt.

iii. Task 3 (10%): pick one of the following tasks:

Machine Translation: Machine translation involves translating text from one language to another. Models in this category include sequence-to-sequence models, Transformer models like BERT, and statistical machine translation approaches. Feel free to use any approach of your choice and conduct translation of your text inputs to any language of your choice.

OR

Text summarization: generate a new text or sentence from each input sample in your data, which captures the key points of that particular input.

The summarization can range from simple key word extraction based on popularity or frequency to (ideally) more advanced techniques for extractive and/or abstractive summary generation. Feel free to use any strategy and algorithms of your choice. Justify the algorithms of your choice and findings.

Note: the grading will be relevant to how advanced the solution is.

- Across all tasks: Model Training, Testing, Tuning and Refinement
  As with all ML/DL models, remember to split the dataset into training and testing sets. Train the models on the training data and evaluate their performances on the test data.

  Use appropriate evaluation metrics such as accuracy, precision, recall, F1-score, ROC/AUC or any other NLP metrics as you best see fit. Justify your choice of metrics. Test the optimal model with test data to ensure robustness, reliability and generalization. Refine any models based on user feedback (you may need to act as the user / agent in this case) and performance metrics.

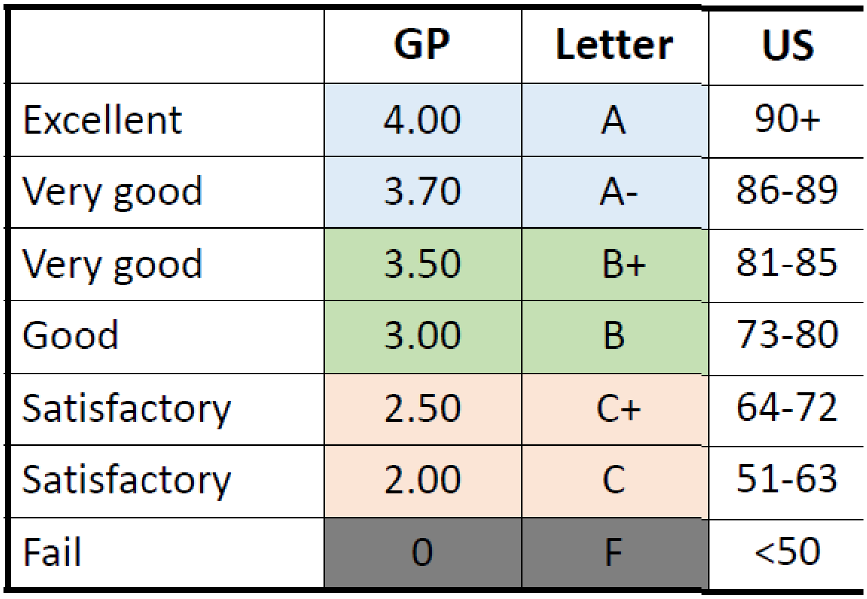
5. *Report quality (15%)*

   Document the entire project, including the dataset used, the challenges faced, preprocessing techniques investigated, model architecture, training process, deployment steps and performance metrics.

   The quality of report is based on many factors including: organization of the material, presentation of data, experiments, models, evaluation, drawing conclusion using various aids such as tables, diagrams, equations etc., and references to external sources and publications. Finally, discuss potential avenues for future work.

6. *Presentation (10%)*
   Prepare a presentation to showcase the project and its outcomes. During the presentation each group will present their work in a comprehensive manner and will be called to answer questions regarding their work. Every member in the team has to present a part of the analysis and findings.

Grading scale: US System



	GP	Letter	US
Excellent	4.00	A	90+
Very good	3.70	A-	86-89
Very good	3.50	B+	81-85
Good	3.00	B	73-80
Satisfactory	2.50	C+	64-72
Satisfactory	2.00	C	51-63
Fail	0	F	<50

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*