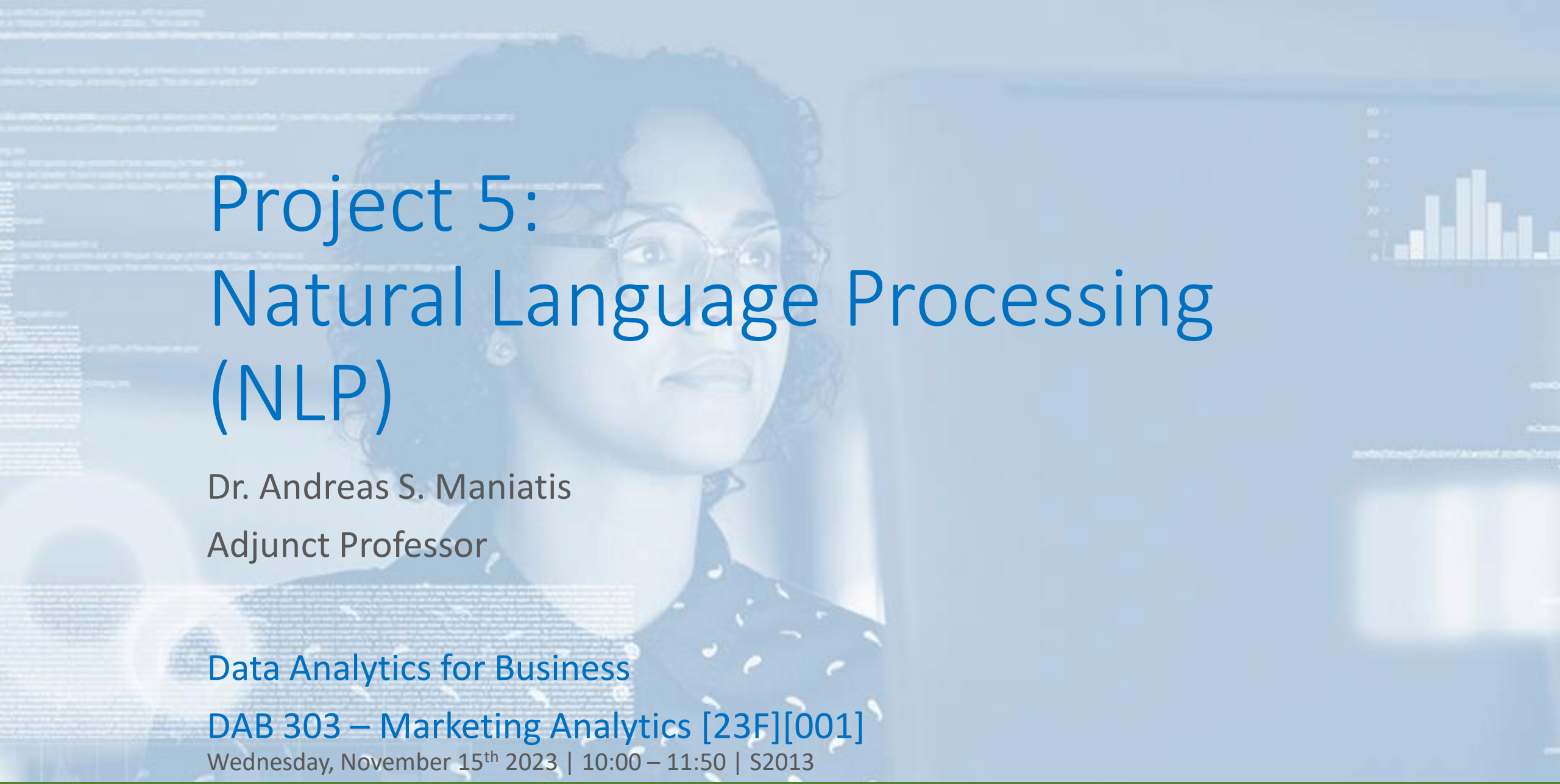

Project 5: Natural Language Processing (NLP)

Dr. Andreas S. Maniatis

Adjunct Professor

Data Analytics for Business

DAB 303 – Marketing Analytics [23F][001]

Wednesday, November 15th 2023 | 10:00 – 11:50 | S2013

ST. CLAIR COLLEGE

Agenda

- Business objective of the project
- About Natural Language Processing (NLP)
- Implementation methodology
- Submission details

Business Objective

Business Objective of the Project

- Understand and gain insights from a social media dataset, by performing various exploratory data analysis, data visualization, and data modelling tasks
- Understand the concepts of **Natural Language Processing** (**NLP**) principles
- Perform **Sentiment Analysis** to analyzing an opinion or feelings about something using tweets in form of text

Sentiment Analysis

Sentiment Analysis – Introduction

- Sentiment analysis refers to analyzing an opinion or feelings about something, using data like text or images, video or any unstructured data
- It helps companies in their decision – making process
- If public sentiment towards a product is not so good, a company may try to modify the product or stop the production altogether to avoid any losses (e.g. twitter feeds, product reviews, etc.)

Natural Language Processing (NLP)

- NLP is a subfield of linguistics, computer science, information engineering, and artificial intelligence
- It is concerned with the interactions between computers and human (natural) languages
- NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way
- NLP breaks down the process of understanding into small chunks
- Building an NLP pipeline includes 3 stages:
  - Text Processing
  - Feature Extraction
  - Modeling
- In a nutshell, NLP is applying machine learning models to text and language
- A crash course in NLP can be found here

The NLP Pipeline

- Text Processing
  - TP
- Feature Extraction
  - FE
- Modeling
  - It includes designing a statistical or machine learning model, fitting its parameters to training data, using an optimization procedure, and then using it to make predictions about unseen data.
  - The nice thing about working with numerical features is that it allows you to choose from all machine learning models or even a combination of them.

Text to Numeric

- Statistical algorithms require data in mathematical form to train the machine learning models.
- Text data needs to be converted into some numeric form to be used in training the machine learning models
- Two main approaches to convert text to numbers:
  - Bag of words
  - TF-IDF (Term Frequency and Inverse Document Frequency)

NLP Pipeline

- The nice thing about working with numerical features is that it allows you to choose from all machine learning models or even a combination of them. Treats each document as an unordered collection or bag of words
- To obtain a bag of words from a piece of text apply text processing steps such as cleaning, identifying stop words, normalizing, stemming, lemmatization etc.
- Turn each document into a vector of numbers representing how many times each word occurs in the document
- A set of documents is called a corpus
- This corpus gives the context for the vectors to be calculated
- Collect all the unique words present in the corpus to build the vocabulary
- Arrange them in some order and create a table like structure where each word is a column, and each document is a row
- Count the number of occurrences of each word in each document called Document Term Matrix
- This approach treats every word as equally important

TD-IDF

- TF-IDF is a combination of two terms:
  - Term Frequency (TF)
  - Inverse Document Frequency (IDF)
- TF = (Frequency of a word in the document)/(Total words in the document)
- IDF = Log((Total number of docs)/(Number of docs containing the word))
- The idea behind the TF-IDF approach is that the words that occur less in all the documents and more in individual document contribute more towards classification
- In this approach we assign weights to words that signify their relevance in documents

Methodology

Methodology (I)

- The project is spread over 2 weeks and is completed in 1 part
- Description of the various steps will be presented (Jupyter Notebook). You need to:
  - Review and reverse-engineer the provided code,
  - Run the code,
  - Secure that the final code is error free,
  - Explain the code with commenting, and
  - Include all code output on the Jupyter Notebook
- Reporting/presentation must include insights (through visualizations), and recommendations

Generic Methodology (II)

1. Data Import
2. Data Overview
3. Data Cleansing (Missing Values, duplicates, etc.)
4. Exploratory Data Analysis (EDA)
5. Statistical Analysis
6. Create various visuals using Python Packages
7. Variable distribution
8. Variable Summary
9. Correlation Matrix
10. Data Pre-Processing for Model Building
11. Model Building

Methodology (III)

- Prepare a final report document (~ 10 pages):
  - Record your observations with respect to the analysis done,
  - Use your findings to identify significant NLP patterns, and
  - Devise a high-level marketing strategy to entice these individuals to continue using the service.

Methodology (IV)

- Prepare a final Powerpoint presentation (~ 5 slides, without covers):
  - Simplify your findings,
  - Keep in mind that you are addressing managers/stakeholders, and
  - If needed to refer to technical matters, keep it simple, use business terminology that managers comprehend, rather that technical jargon.

Submission

Submission

Submission will be done via Blackboard, FOLLOWING PEER REVIEW, and it will be a group submission, including:

- One file per group (in .zip format):
  - Jupyter Notebook (Including extended code commenting and analytical block code description):
    - Lab file (.ipynb)
  - Report (.pdf): Include the major steps and finding of your analysis, and
  - Presentation (.pptx): 4 – 5 slides (excluding covers and introduction), for presenting your findings to the management

Suggestions

Keep in mind and please comply with the following suggestions:

- Keep focused on time and resource distribution:
  - Work as a group for the whole two weeks period
  - Complete your work early. Don’t work on the day of peer-review
- Deliverables:
  - To be uploaded after peer review
  - Late presentations / submissions will not be graded
- Peer review:
  - It’s suggested that you start from the presentation and answer review questions that may lead you to code
  - Rehearse your presentation
  - All members need to be present and to present!
  - 10 MINUTES MAX per Group Presentation!

Thank you!

Questions?

Dr. Andreas S. Maniatis
Adjunct Professor
AManiatis@StClairCollege.ca
http://www.linkedin.com/in/andreasmaniatis

ST. CLAIR COLLEGE