Applied Text Analytics &
Natural Language Processing

Dr. Mahdi Roozbahani (GT Faculty and founder of Filio)
Wafa Louhichi (GT Alumna and quantitative analytics specialist)

Class Introduction

GT

Learning Objectives

- Why Applied Natural Language Processing (NLP)?

- What you will learn in the class?

- What deliverables do we expect from you?

Why Applied NLP?

- Text and documents are everywhere!
- There are hundreds of languages in the world
- They are primary information artifacts in our lives
- There are large volumes of textual data
- Big and small companies are looking for this skill

GT

There are a lot of Texts and Written Information

- Thanks to the internet!

- WWW: webpages, Twitter, Facebook, Wikipedia, Blogs,...

- Digital libraries: Google books, ACM, IEEE,...

- Lyrics, closed caption,... (YouTube)

- Police case reports

- Legislation (law)

- Reviews (products, rotten tomatoes)

- Medical reports (EHR electronic health records)

- Job descriptions

Examples of Applications of NLP?

- Establish authorship, authenticity; plagiarism detection

- Classification of genres for narratives (e.g., books, articles)

- Tone classification; sentiment analysis (online reviews, twitter, social media)

- Code: syntax analysis (e.g., find common bugs from students' answers)

- Machine translation (e.g., Google Translate)

What Makes NLP Challenging?

- Interdisciplinary field that lies at the intersection of linguistics and machine learning

- Ambiguity at multiple levels in the human language:
    - Lexical (word level) ambiguity -- different meanings of words
    - Syntactic ambiguity -- different ways to parse the sentence
    - Interpreting partial information -- how to interpret pronouns
    - Contextual information -- context of the sentence may affect the meaning of that sentence

What will You Learn in the Class?

- Pre-processing:
    - How to clean texts and documents
    - Tokenization
    - Reducing the inflectional forms of a word
        - Stemming
        - Lemmatization
    - Normalization
    - …

GT

What Will You Learn in the Class?

- Text Representation:
    - One hot encoding
    - BoW (frequency counting)
    - TF-IDF
    - Embeddings

What Will You Learn in the Class?

- Overview of classification methods:
    - Naïve Bayes
    - Logistic Regression
    - SVM
    - Perceptron
    - Neural Network

What Will You Learn in the Class?

- Overview of Deep learning
    - Convolutional neural network
    - Recurrent neural network
    - Long short-term memory

What Will You Learn in the Class?

- Overview of topic modeling
  - Principal Componenet Analysis
  - Singular Value Decomposition
  - Latent Dirichlet Allocation

What Will You Learn in the Class?

- Overview of Transformer models
    - Bidirectional Encoder Representations from Transformers
    - Generative Pre-trained Transformer

What Deliverables Do We Expect from You?

- There are four homeworks
    - HW1: Text pre-processing, Classification introduction
    - HW2: Classification methods, dimensionality reduction, SVD
    - HW3: Deep learning
    - HW4: Transformers and unsupervised models

What Deliverables Do We Expect from You?

- There are 10 quizzes [%15]
    - Solidify your learning.
    - Quizzes measure your understanding of the topics and they will be mostly conceptual questions.
    - Each quiz will have multiple choice questions.
    - They will be available for a specific duration within a week
    - You will have a limited time to finish each quiz.
    - All quizzes are mandatory to be taken.

GT

Course Goals and Summary

- Demonstrate how to pre-process textual data

- Differentiate different text representation methods and techniques

- Explain different NLP tasks

- Develop and assess the performance of different NLP models using a variety of techniques