Feb, 2023

Home Assignment

Data Science

Introduction

This is a home assignment for the role - **Data Science**. Please follow the instructions:

1. The main goal of this assignment is to **analyze/extract topics** within a huge corpus.
2. The primary language to use in this exercise is **Python** (and its auxiliary packages of your choice) on top of **Jupyter Notebook**.
3. Your code should be research-grade, as detailed below:
    a. The code should be reliable (think of the correctness and edge cases).
    b. The code should be readable and maintainable.
    c. Take (some) memory efficiency into consideration while solving the problem.
    d. The data visualization should be functional enough to support the answers.
4. The solution should be extensible:
    a. Work *iteratively* within the given time frame to get *something* before you deepen your research.
    b. Think of the step-by-step implementation.
    c. Being *rigorous* is important for the maintenance, correctness, and flexibility of the results, but try to be concise and elegant where possible.
5. We value your time. The whole assignment should take roughly **0.5-1 day**. If you think it will take longer, please let us know ahead of time

Good Luck!

Objective

Analyze a dataset of headlines. First - clean, explore, and tokenize the text. Then, apply common NLP methods used for topic modeling.

Part 1 - Data cleaning and exploration

1. Load the attached dataset using Pandas.
2. Clean and prepare the data for analysis by performing the following tasks:
   a. Remove missing values
   b. Handle any outliers present in the data
   c. Convert any non-numeric columns to numeric/date (when possible)
   d. Tokenize / clean to your understanding.
3. Perform an EDA of your choice to help you understand the data.

Part 2 - Topic modeling

Attempt at least one topic modeling method of your choice to cluster the headlines into topics.

1. Since we cannot determine the number of topics in advance, experiment sensibly to assess a reasonable amount of topics.
2. Devise a method (either computationally, graphically, or both) to decide about the number of topics, and then match headlines to topics
3. Try to estimate somehow the matching error.

Part 3 - Bonus

Compare the method with an additional topic modeling method.

Deliverables

Jupyter Notebook: A Jupyter Notebook containing all the code.