Assignment # 03

Data Science Tools & Techniques (MS DS)

Deadline: 16-Nov-2025 @ 11:59

Total Marks = 100

Instructions:

- Combine all your work in one folder. The folder must contain only deliverable files (no binaries, no exe files, etc.).
- Rename the folder as ROLL-NUM_SECTION (e.g., 22i-0001_A) and compress the folder as a zip file. (e.g. 23i-0001_A.zip). Do not submit a .rar file.
- All the submissions will be done on Google Classroom within the deadline. Submissions other than Google Classroom (e.g., email, etc.) will not be accepted.
- The student is solely responsible for checking the final zip files for issues like corrupt files, viruses in the file, and mistakenly executed sent. If we cannot download the file from Google Classroom for any reason, it will lead to zero marks in the assignment.
- Displayed output should be well-mannered and well-presented. Use appropriate comments and indentation in your source code.
- **Be prepared for a viva or anything else after the submission of the assignment.**
- If there is a syntax error in the code, zero marks will be awarded in that part of the assignment.
- Understanding the assignment is also part of the assignment.
- **Zero marks will be awarded to the students involved in plagiarism. (Copying from the internet is the easiest way to get caught.)**
- **Late Submission** will **not** be entertained, and **no request retake** will be accepted as per the course policy.

Tip: For timely completion of the assignment, start as early as possible.

Note: Follow the given instruction; failing to do so will result in a zero.

Q#1: Large-scale Web Crawl NLP + EDA Assignment (100 Marks)

Perform the following tasks. A Common Crawl-like dataset has been provided for experiments. Records: 50,000. File: dataset.csv

Task 1: Preprocessing & Tokenization (20 Marks)

- ✓ Load the provided CSV file and inspect top 100 records for structure and anomalies.
- ✓ Create a stopword dictionary tailored to the dataset (you will be provided with stopwords.txt). Explain choices for any added tokens.
- ✓ Tokenize the content field, normalize case, remove punctuation, and remove stopwords. Show example output for 10 pages.

Task 2: Stemming and Lemmatization (15 Marks)

- ✓ Using the tokenized output, perform stemming (Porter or Snowball) and lemmatization (WordNet or spaCy).
- ✓ Compare and contrast the outputs for 10 sample documents and explain which is more appropriate for web-crawl text and why.

Task 3: Bag-of-Words (15 Marks)

- ✓ Build a document-term matrix (Bag-of-Words) using the preprocessed tokens. Use vocabulary pruning to remove very rare and very frequent tokens (e.g., min_df=5, max_df=0.85).
- ✓ Report top 50 terms by document frequency and show their counts.
- ✓ Deliverable: DTM matrix, table of top 50 terms.

Task 4: TF-IDF for Top K words (20 Marks)

- ✓ Compute TF-IDF across the corpus. Select TOP K words by aggregate TF-IDF score where K is a variable (set K=50, 100, 200 in experiments).
- ✓ Save the TF-IDF vectors for the chosen K and provide a brief analysis of what kinds of words are ranked highest.
- ✓ Deliverable: TF-IDF results, saved vectors (as .npy or .csv)

Task 5: Visualization (30 Marks)

- ✓ Using the TF-IDF vectors for your chosen K (suggest K=100 for visual clarity):
- ✓ Reduce dimensionality using PCA and plot the top components, coloring points by language or http_status.
- ✓ Create a cluster map (hierarchical clustering heatmap) of the top 50 TF-IDF terms vs a sample of 500 documents.
- ✓ Bar chart of the top 30 terms by aggregate TF-IDF score.
- ✓ Heatmap of term co-occurrence (or term correlation) for top 50 terms.
- ✓ Deliverable: Plots included in the notebook, and a short interpretation for each plot (≈400 words total).

Report Writing (10 Marks)

- ✓ Submit a brief report (~800–1000 words) covering:
- ✓ Overview of preprocessing steps.
- ✓ Comparison of stemming vs lemmatization.
- ✓ Insights from BoW and TF-IDF analyses.

- ✓ Visualizations and key findings.
- ✓ Deliverable: Report (Word or PDF).

Deliverables

1. dataset.csv (dataset)

2. stopwords.txt (stopword list used)

3. Jupyter Notebook with code, outputs, and plots (clearly organized).

4. Report (Word/PDF).

5. Any saved vectors or intermediate files (TF-IDF vectors etc.).

Notes:

- ✓ Use appropriate comments and indentation in your source code.
- ✓ Clearly mention which libraries (and versions) you used.
- ✓ Be prepared for viva; understand the choices and results.
- ✓ Plagiarism will be penalized.