THE UNIVERSITY *of* EDINBURGH

Text Technologies for Data Science

INFR11145

Coursework #2

Instructors:

Björn Ross & Youssef Al Hariri

22-Nov-2023

Scope

- Coursework 2 has three parts:
  - IR evaluation → 30%
  - Text analysis → 35%
  - Text classification → 30%

CW2 depends on

- Lectures:
  - Lecture 9/10: Evaluation
  - Lecture 15/16: Comparing Corpora
  - Lecture 17/18: Text Classification
- Labs:
  - Lab 6: Comparing Corpora
  - Lab 7: Text Classification

- Big portion of this CW is not supported by labs.
- PLEASE complete lab 7 ON TIME

IR Evaluation

- Implement a simple IR evaluation tool
- Input → Two files
  - Results file
  - Relevance assessments (qrels)
- Output → Results table with the following measures
  - Precision at k
  - Recall at k
  - R-Precision
  - AP/MAP
  - nDCG at k
- Measure significance test between systems

Text Analysis

- 3 “corpora”
  - Quran
  - Bible: Old Testament
  - Bible: New Testament
- Find which words are most distinctive for each
- Run topic models to discover topics relevant for each
- Present your analysis in the report
  - What does your analysis tell you about the corpora?
  - Please aim for high-quality formatting and presentation

Text Classification

- Improve classifier developed in Lab 7 and apply to new dataset (different one from Text Analysis part)
- Apply Text Classification Evaluation
  - Precision, recall, accuracy, F1
- Discuss results and improvements over baseline

Deliverables

- File containing IR evaluation scores
  - *Use provided script to check your format!*
- File containing classification results
  - *Use provided script to check your format!*
- Code
  - 1 well-documented file
- Report
  - 70% of all CW2 marks based on report

Allowed / Not Allowed

- Allowed:
  - Use ready code for optimisation
  - Use existing implementations for topic models and text classifiers
  - Discuss some functions with your friends
  - Use Piazza to ask question on implementation

- Not allowed:
  - Use any ready libraries or implementation of IR measures
  - Sharing any indication about results
  - Asking questions that can reveal the answer (ask privately in this case)
  - Sharing code

Penalties

- When sharing any indication of results publically
- Submitting results in incorrect format

Extensions policy

- Can be found in Assessment page on Learn
  - TTDS follows **Rule 1** this year
- Penalty of 5% per day coursework is submitted late unless
  - You have an Extension (3 days)
    - Must apply for extension before the deadline
  - OR You have Extra Time Adjustments
    - Must be registered with Student Disability Service
    - **Can** be combined with regular extension
- Maximum late submission depends on your situation (see web page). If you exceed the maximum late submission, a mark of zero will be given!
- Unaffected by extensions policy: Special circumstances

Timeline

- 10 Nov 2023
  Full instructions for CW2 published
- **1 Dec 2023, noon** UK time
  CW2 submission deadline

- 29 Nov 2023
  CW3 announced
- 15 Mar 2024, **noon** UK time
  CW3 submission deadline