CENG442 Natural Language Processing

Midterm Exam

Due date: 06/05/2024 23:59

You may form groups of up to 2 or 3 members.

Sentiment Analysis of Azerbaijani Turkish Texts

To develop and evaluate models for sentiment analysis on Azerbaijani Turkish texts using various NLP techniques, with a focus on the impact of different tokenizations and word embeddings.
The dataset provided includes the following features: content (text), score (sentiment rating), and upvotes (popularity measure).

1. Project Introduction

Purpose: Explain the importance and applications of sentiment analysis.
Scope: Outline the methodologies and tools to be used in the project.
Deliverable: A brief document outlining the project scope and objectives.

2. Data Exploration and Preprocessing

Tasks:
Load and inspect the dataset.
Perform basic data cleaning (e.g., removing null values, filtering out non-Turkish texts).
Visualize data distributions (e.g., sentiment scores, upvote distribution).
Deliverable: Jupyter notebook with exploration code and initial findings.

3. Data Preparation

Tasks:
Split the data into training (80%) and testing (20%) sets.
Justify the splitting ratio.
Prepare a preprocessing pipeline to clean text data (e.g., removing special characters, converting to lowercase).
Deliverable: Documented Python script for data splitting and preprocessing.

4. Tokenization

Tasks:
Implement tokenization using Polyglot, NLTK, and spaCy.
Compare the results of each method.
Deliverable: Python notebook with implementation and comparison of tokenization methods.

5. Word Embeddings and Model Building

Tasks:
Create a baseline model using custom word embeddings built into an RNN with GRU layers.
Implement models using pre-trained embeddings: GloVe, Word2Vec, SVD, Polyglot, FastText.
Experiment with different activation functions (ReLU, sigmoid, tanh).
Deliverable: Python scripts for each model with a detailed explanation of the architecture.

6. Model Evaluation

Tasks:
Evaluate each model using appropriate metrics (accuracy, loss function).
Analyze the performance and discuss the results.
Deliverable: Detailed evaluation report with graphs showing performance metrics.

7. Visualization

Use Matplotlib and TSNE to visualize:
The distribution of sentiment scores.
The word embeddings space to interpret semantic similarities.
Model accuracy and loss over epochs.
Deliverable: Visualizations in a Python notebook.

8. Conclusion and Discussion

Tasks:
Summarize key findings.
Discuss challenges encountered and potential improvements.
Deliverable: Final report summarizing methodology, findings, challenges, and future work.

9. Presentation

Tasks:
Prepare a presentation to summarize the project.
Include discussion on methodology, results, visualizations, and learnings.
Deliverable: Slide deck for class presentation.

If you miss the presentation, your midterm score will be graded as 0.

END of THE DOCUMENT