Project 2 : supervised learning

Deadline: 13 January 2025

Data exploration

You will use this dataset :
https://drive.google.com/file/d/1_kg5JzAzntzLI6eGM3_vmUSoeWk7f8ip/view?usp=sharing

You will undertake an initial exploration of the data, including data cleaning, visualization, and the production of initial conclusions. This step is crucial to establish a solid foundation for subsequent projects.

Supervised learning

Description

In the second phase, supervised learning and unsupervised learning will be utilized. You will need to create a supervised text processing model using NLP techniques. In addition to modeling, you will need to develop an interactive application where users can submit text in the chosen theme and receive a prediction, along with explanations for that prediction. This will allow you to apply your NLP skills in a practical manner.

Supervised Tasks Examples

1. Number of Stars
   - Predict the number of stars (e.g., 1-5) based on the review content.
   - This can be treated as a regression or classification problem.
2. Sentiment Analysis
   - Predict the sentiment of a review (positive, neutral, negative).
   - Label examples as training data for fine-tuning models.
   - You can use few shot or zero-shot models
3. Categories/Subjects Detection
   - Detect the primary subject of the review:
     - Pricing
     - Coverage
     - Enrollment
     - Customer Service
     - Claims Processing

        - Cancellation
        - or other categories
    - Use few-shot learning or zero-shot classification models (e.g., OpenAI, HuggingFace's pipeline) to assign categories to unlabeled reviews.

Streamlit Applications

Example 1: Prediction Application

- **Objective**: Predict the number of stars or the main subject of the review.
- **Features**:
    - User inputs a review.
    - Model predicts the star rating and/or category (e.g., "Pricing" or "Customer Service").
    - Display results in real-time.
- **Implementation Steps**:
    - Use a pre-trained text classification model for stars.
    - Implement a zero-shot or fine-tuned classifier for subjects.

Example 2: Insurer Analysis Application

- **Objective**: Provide insights into insurer performance based on reviews.
- **Features**:
    1. **Summary by Insurer**: Aggregate reviews and generate a summary using NLP techniques.
    2. **Review Search**: Enable search functionality for specific reviews using keywords or filters (e.g., star ratings, subjects).
    3. **Metrics**:
        - Average star rating by insurer.
        - Average star rating by subject for each insurer (e.g., "Pricing: 3.2/5").
- **Implementation Steps**:
    1. Preprocess reviews to structure them by insurer and subjects.
    2. Generate summaries using text summarization models (e.g., T5, GPT-3).
    3. Reuse code from **Project 1** for implementing the search functionality.
    4. Compute and visualize metrics using Streamlit widgets like bar charts and tables.

    -

Scoring

1. Data Cleaning: 2 points (negative points if not well-executed)
    - Highlighting frequent words (and n-grams)
    - Spelling correction: 2 points
2. Summary, Translation, and Generation: 2 points
    - Produce a clean file with multiple cleaned columns and corrected/translated texts
3. Topic Modeling and Lists of Topics: 2 points
4. Embedding to Identify Similar Words : 2 points (possible negative points)
    - Word2Vec Training: 2 points, GloVe: 1 points
    - Visualization of embeddings with Matplotlib and Tensorboard: 2 points
    - Implementation of Euclidean or cosine distance: 1 point
    - Semantic search : bonus 2 points
5. Supervised Learning, each model well-made and well-presented: 2 points (possible negative points). Sentiment analysis is a particular case of supervised learning.
    - TF-IDF and classical ML
    - Basic model with an embedding layer (embedding visualization with Tensorboard: additional 1 point)
    - Model with pre-trained embeddings (embedding visualization with Tensorboard: additional 1 point)
    - USE (Universal Sentence Embedding) or equivalents, RNN LSTM, CNN, BERT, or other models on Hugging Face
    - LLM
    - Comparison of different models
6. Results Interpretation (possible negative points)
    - Error analysis: 1 point
    - Sentiment detection: 2 points
    - Classical models with themes: 2 points
    - Deep learning models for words: 2 points
7. Creation of Streamlit applications
    - Prediction (2 points)
    - Summary (2 points)
    - Explanation (3 points)
    - Information Retrieval (3 points)
    - RAG (3 points)
    - QA (3 points)
8. Clarity of Presentation: 2 points (possible negative points)

You can use this template :
https://docs.google.com/presentation/d/1hyaVKY31U0wP4kensljOgIiudkRC5N1OxZMWqZ07Y5Q/edit?usp=sharing

Template en français
https://docs.google.com/presentation/d/1LGq58zA_5Usmqkz043iHYe3VqDrbQOARXUI_QWD_W3Y/edit?usp=sharing

Project 3: Theme Analysis and Evolution of NLP Techniques

For the third project, students are tasked with selecting one of the proposed themes, such as chatbots, text generation, sentence embedding, etc. Subsequently, they will conduct in-depth research, curating a collection of relevant scientific articles that exemplify the progression of techniques within their chosen theme. Students are required to present these articles in a structured manner and analyze the evolving landscape of NLP techniques as gleaned from their selected articles.

Please upload on DVO your presentation one weekend before.

Presentations will be conducted in groups of four students during the final session.

Themes :

- Chatbot Development and Evolution
- Text Generation Techniques (RNN, and GPT)
- Word and Sentence Embedding Approaches
- Text Summarization Strategies
- Text Translation Innovations
- Image Generation from Texts and Multimodal Models
- Detecting Bias and Ensuring Fairness in NLP
- Automatic Audio Summarization Advancements
- Question Answering Systems
- Information Extraction Methods
- Sentiment and Emotion Analysis in NLP
- Evaluation Metrics for NLP Techniques
- etc.

You can choose your theme from here :

🟩 Groups presentation of Theme Analysis and Evolution of NLP Techniques

You may also be able to add more themes.

Note: The themes provided serve as options for students to choose from for their research and thematic analysis. Each group of four students will focus on one theme, exploring the evolution of NLP techniques within that specific domain.

Example of structure

- Context (3 min)

- List of techniques (10 min)
    - description
    - comparison
    - pros and cons
    - examples
    - visualization
- References (papers, especially remarkable papers + articles + youtube videos) (2 min)
    - definitions
    - explanations
    - visualizations