Project workplan

Text summarizer for chat messages in a group chat

Andrea Auletta Davide Baggio Marco Brigo
Marco Bernardi Sebastiano Sanson

May 11, 2024

Contents

1 Short summary

Our Natural Language Processing (NLP) project aims to automate the summarization of group chat messages, providing users with concise summaries of conversations without the need to read through all messages. Imagine a scenario where you're part of bustling team chat or a lively online community. Messages fly back and forth very quickly about various topics and discussions. Obviously chat logs accumulate rapidly and it might happen that you miss some important points, requiring significant time and effort to read through the entirety of the convesation history. Our aim is to save users time and cognifitive effort when facing situations like this. By leveraging state-of-the-art natural language processing techniques our system will analyze the content of group chat messages, summarizing the most relevant information, the key points, main topics, and notable highlights of the discussion, presenting users with a condensed version of the conversation that is easy to digest and comprehend.
The idea is to try several existing approaches that can be useful to reach an optimal goal. Then we will try to create our own version by using the techniques that we found to be the most effective.

2 Tasks

Number	Task	Mandatory	Participants
1	Study of the state-of-the-art of the topic	Yes	Auletta, Baggio, Bernardi, Brigo, Sanson
2	Dataset collection	Yes	Auletta, Baggio, Bernardi
3	Data preprocessing	Yes	Brigo, Sanson
4	Summarization of text messages	Yes	Auletta, Baggio, Bernardi, Brigo, Sanson
5	Summarization of images	No	Auletta, Brigo, Sanson
6	Summarization of audio	No	Baggio, Bernardi
7	Test and evaluation of the model	Yes	Auletta, Baggio, Bernardi, Brigo, Sanson
8	Documentation	Yes	Auletta, Baggio, Bernardi, Brigo, Sanson

2.1 Time schedule

Task	Auletta	Baggio	Bernardi	Brigo	Sanson	Total
1	5	5	5	5	5	25
2	5	5	5	0	0	15
3	0	0	0	5	5	10
4	20	20	20	20	20	100
5	5	0	0	5	5	15
6	0	5	5	0	0	10
7	10	10	10	10	10	50
8	5	5	5	5	5	25
Total	50	50	50	50	50	250

3 Deliverables

1. **Study of the state-of-the-art of the topic**:
   - **Deliverable**: Report of the state-of-the-art;
   - **Description of the deliverable**: Each member of the group will be assigned a certain number of papers that they will have to read and summarize. At the end of this process, a final report condensing all the information will be produced, showcasing the state-of-the-art in the field;
   - **Measurable objectives**: We will report about at least 15 papers in our state of the art analysis (3 for each of us);
2. **Dataset collection**:
   - **Deliverable**: Dataset;
   - **Description of the deliverable**: A dataset of chat messages will be collected or found;
3. **Data preprocessing**:
   - **Deliverable**: Preprocessed and analyzed dataset;
   - **Description of the deliverable**: The dataset will be preprocessed and analyzed to be used in the model. A notebook file will be produced (.ipynb);
4. **Summarization of text messages**:
   - **Deliverable**: Model (text messages);
   - **Description of the deliverable**: An implementation of the model that can summarize text messages will be produced;
5. **Summarization of images**:
   - **Deliverable**: Model (images);
   - **Description of the deliverable**: An implementation of the model that can translate images in text and then summarize them with the text;

6. **Summarization of audio**:

    - **Deliverable**: Model (audio);
    - **Description of the deliverable**: An implementation of the model that can write down the audio and then summarize them with the text;

7. **Test and evaluation of the model**:

    - **Deliverable**: Evaluation report;
    - **Description of the deliverable**: The model will be tested and evaluated;
    - **Measurable objectives**: The model will be evaluated in different ways:
        - Manual evaluation: We will check several properties about the sentences produced in the summaries such as: readibility, structure and coherence, grammaticality, content coverage, referential clarity, conciseness and focus and non-redundancy;
        - Automatic evaluation: we will use some metrics to evaluate the model, for example ROUGE-1, ROUGE-2 and F-Measure. We will try to achieve a ROUGE > 0.4, and F-Measure > 0.5;

8. **Documentation**:

    - **Deliverable**: Documentation;
    - **Description of the deliverable**: The necessary documentation will be produced to understand the project;