**Data Science Term Project Report**

Yoon jiseob

Lee JongMin

Kim JungHoon

1. **Define.**

To begin with, a purpose of our term project was to extract highlight clips by analyzing Twitch chat logs. Since there were no proper criteria to evaluate our results, we used highlight clips uploaded in YouTube to check our model’s accuracy. To do so, we used Type I and Type II methods.

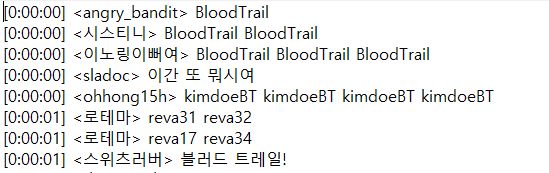
1. **Procedure**

First, we used Twitch chat downloader (tcd) to collect all the Chatlogs.



<fig 1.1 tcd usage>

Fig 1.1 shows example of usage of tcd library. which means collect chat logs from latest 100 video from streamer id ‘kimdoe’ into text file. The way tcd library store the chat logs into text file is presented as below.



<fig 1.2 example of tcd result>

First came up with time, user id, and chat logs. However, as we can see in above fig 1.2, within 1 second more than 5 chats are created by users. Which means within one text files, there exists millions of data.

Before processing any methods to analyze chat logs, due to the presence of either meaningless or needless words, data cleaning is needed. Hence, we print out which word came up most frequently in orderly manner to expect which word can be used or cannot be used for data analysis.



<fig 1.3 tore the word into list>



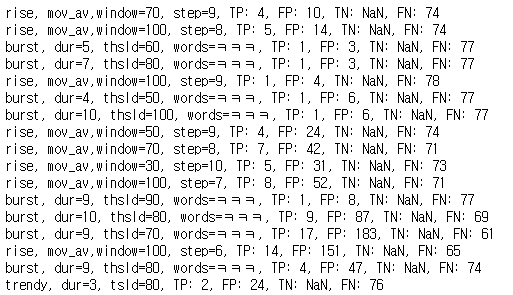
<fig 1.4print most frequent word in sorted manner>

By looking at fig 1.3and fig 1.4 we can notice that ‘?’ , ‘ㅗㅜㅑ’,. ‘ㅋㅋㅋ’ came up most. Which means, there is a possibility of the time where above 3 words came up can be part of highlight clips.

Second, we used 3 methods to support fulfilling our purpose.

1. Hot – select, if the chat per second incremented by given step.
2. Trendy – select, if word appears more than threshold within given range of time.
3. Burst – select if specific word appears more than threshold within given range of time.

Eventually, our model will return range of time that can be used for comparing with YouTube highlight clips.



<fig 1.5 example results>

Fig 1.5 shows the sequence of our model results. Algorithm, parameters, and error results. To evaluate our models, we used type I and type II methods. Simply if our predicted results (range of time) are present in streamer’s YouTube highlight clips, we score it as True Positive. On the other hand, if our predicted results are not shown in streamer’s YouTube Highlight clips, we score it as True Negative.

1. **Conclusion**

For the conclusion, as it mentioned in very start, we were having trouble to identify whether the predicted results by our model can be used for highlight clips. Because the highlight clips are subjective and dependent. Which means depends on category streamers are working on, the meaning of highlights can be vary.

Moreover, in fig 1.4, word “ㅋㅋㅋ” took 3rd place for frequency. However, the most results came up with False Positive. There can be many reasons, for example, “ㅋㅋㅋ” can be used either if streamer make hilarious situations or just followed by other words meaninglessly. So in our model, Specificity does not take big role. Since most of results are predicted as False Positive.