**Readme file for Team 5**

**Team members: Sadaf Waqas, Kristine Bagrova, Joss Miller-Todd, Andrew Grassick, Richard S Marchesi, Albert Pasto Jorquera, Balint Juhasz**

The notebook (Team\_5\_Final\_notebook\_v2.ipynb) consists of the below 4 modules:

**Module 1: Reading in and processing Word documents (Focus Group data)**

We defined a class FocusGroup to read in (with the textract package) the word focuments and store the important attributes of the transcripts (corresponding talker – paragraph pairs). We defined a text processing pipeline which includes the below data processing steps. We leveraged the WordNetLemmatizer() class from the nltk library for lemmatization and extended the stopwords dictionary by looking at the most frequent words and excluding filler words (for example, okay) and words which are the consequence of the automatic transcription (for example, inaudible). Steps of text processing:

1. Making the string lowercase
2. Removal of punctuation
3. Tokenization
4. Removal of text with less than min\_token\_count tokens
5. Removing stopwords
6. Lemmatization
7. Removing stopwords (also after the lemmatization)

The output of the text processing pipeline is a list with two elements, the first element is the processed, tokenized text and the second element is the original text with the purpose to help with the intepretation of the results.

**Module 2: Running the gensim LDA method iteratively to identify additional stopwords and the application of the Bertopic Dynamic topic modelling after filtering out the stopwords**

We fitted the LDA model from the gensim package for each of the four types of Word documents (Gaming / Low PIU / Media / Social) iteratively with 5 topics and excluded the words which were included in at least 3 of these 5 topics. We iterated until there were only a few, 3 or 4 common words among the top 10 words per topic. The rationale for this data processing step is to avoid the situation when all of the topics consist of the same most frequent words, just with a different order. We used the dynamic topic modelling feature of the BERTopic class to identify topics and track their change over time. It turned out later that the Word documents were created within a single month time window, but the pipeline can be leveraged for documents spanning a wider time range in the future.

**Module 3: Analysis of Crisis Logger data with WordCloud**

First we used the KeyBERT class of the keybert library for keyword extraction and then used the wordcloud library to visulize the wordcloud. The main themes gathered are parents working from home, teenagers have been challenging during this time along with challenging work environments.

**Module 4: Analysis of Prolific Academic data with Dynamic topic modelling from Bertopic**

We used the dynamic topic modelling feature of BERTopic to idenfiy changes in the specify\_positive field of the Prolific Academic surveys over time. From the plot of the evolution of topics it was seen that in the beginning of the lockdown April 2020, the factor of the first topic about the enjoyment of the lockdown peaked off whereas it gradually decreased as the families became used to of the new routine being at home all the time. Same goes for topic 3 about excess time with family since excessive family time has been a welcome change in the initial days but till November 2020 this element is not that frequent. i.e. its importance dropped with time.