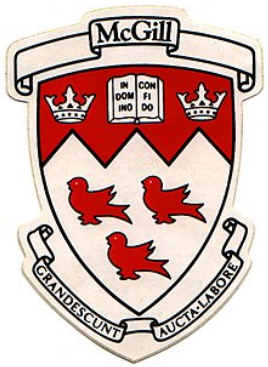
Practical 1
Preliminaries to Recommender Systems

Practical 1.1: Handling large data

Instructor	Etienne Tajeuna	etienne.tajeuna@mcgill.ca
Teaching assistant	Dima Al Saleh	dima.alsaleh@mcgill.ca

School of Continuing Studies

McGill University



Practical 1.1: Handling big data

In this first practical work, we want to manipulate a large volume of data collected from Tweets surrounding the COVID-19[1]. Full data available here: `https://drive.google.com/file/d/1Dn7VfY8XmGJybFT2dJ0VS9Qo1sVWKOEx/view?usp=sharing`

```
> db.Sentiment_Tweets.findOne()
{
        "_id" : ObjectId("62591a947012ae68e09536b3"),
        "user_id" : "1319491585",
        "tweet_timestamp" : ISODate("2020-01-27T16:44:36Z"),
        "keyword" : "wuhan",
        "country/region" : "Malaysia",
        "valence_intensity" : 0.336,
        "fear_intensity" : 0.575,
        "anger_intensity" : 0.505,
        "happiness_intensity" : 0.184,
        "sadness_intensity" : 0.507,
        "sentiment" : "negative",
        "emotion" : "fear"
}
```

Figure: Sentiment record sample.

[1] `https://www.openicpsr.org/openicpsr/project/120321/version/V12/view`

Practical 1.1: Handling big data

1. For the period ranging from January 28^{th}, 2020 to September 1^{st}, 2021 display the average number of tweets made per continent (i.e. North America, South America, Africa, Europe, Middle-East and Asia+Australia). Explain your strategy;
2. Knowing that users may present one of the sentiments: positive, negative, or neutral, we want to know if the provided features (valence, fear, anger, happiness, sadness) can enable recognizing these sentiments. For this, you are asked to perform an unsupervised learning process.
    1. Due to the large volume of the dataset, you are asked to randomly select 20% of the dataset and run the k-means algorithm with $k = 2, 3$.
    2. Using the PCA, project your clusters in a two-dimensional space to visualize the different clusters. Give an interpretation of your result.
    3. From the visual aspect, you want to further explain the quality of the results plotted. For this, you are asked to calculate the silhouette score (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`) of your clusters and give an interpretation.
    4. For each case of $k = 2, 3$, using the cosine similarity function, assign the remaining 80% of the dataset to their respective clusters. Explain your strategy.
    5. For the particular case of $k = 3$, over the 20% of your dataset evaluate the quality of your clustering using the homogeneity score (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html`). Repeat the same process after assigning the 80% to their corresponding clusters (using the cosine similarity).
3. The sampling performed in step 2 may not be accurate. Based on the positive, negative, and neutral sentiments, you are asked to perform a stratified sampling and run back the whole step 2. Compare your results you what you obtained in step 2.