Sentiment analysis

In this project, you will experience how to **build a sentence-level sentiment analysis model** from scratch. You first need to organize the data according to some topic, train the model with the collected data, and finally evaluate the model.

1. Study a state-of-the-art (SOTA) sentiment analysis approach and present your findings.
   - The SOTA algorithm should be published within five years by reliable organizers.
   - Your findings should cite the following from the original material: an overview of the algorithm (pseudo-code, demonstrative figures of the proposed model, et.), experimental data, and discussions on its pros and cons.
2. Collect the data from a variety of sources
   - Choose a target product of interest and collect corresponding reviews from the seller's webpages and any related webpages
   - You may consider Selenium, BeautifulSoup, or Scrapy, to crawl the web pages and parse the HTML documents.
   - Whether the text is in English, or Vietnamese, depends on your choice. Vietnamese is more challenging, yet it may bring you a better score.
3. Preprocess the collected data and label the sentences
   - Break each of the above reviews into multiple sentences such that: 1) each sentence is a simple sentence, and 2) the content of the sentence subjects to the target product.
   - Normalize the sentence heuristically, e.g., remove spelling mistakes and typos, fix the informal abbreviations, etc.
   - Manually label each sentence as Positive or Negative.
   - Your dataset after processing should include at least 1000 positive and 1000 negative samples. However, the more data, the better model.
4. Create the embeddings for each of the above sentences using the following text representations
   - Approaches: TF-IDF, fastText, and BERT (or PhoBert for Vietnamese)
5. Organize your data in a document-store NoSQL database
   - You must define an appropriate schema to store each sentence as a document.
   - A document includes the preprocessed sentence, label, the corresponding three embeddings, and the reference to the original review.

6. Build your sentiment analysis model using any conventional machine learning approach and report the accuracy, precision, recall, and F1-score overall as well as for each class.

You need to prepare the following materials for your group submission.

- The link(s) to the Google Colab implementation. You must guarantee that no edit is made after the deadline.
- A report addresses Question 1 and presents how you accomplish the tasks in Questions 2 to 6.

Important notes:

- This project gives you a 20% course grade.
- Strictly avoid plagiarism in any circumstance.