Bộ dữ liệu gồm 6 file:

1. **BagOfSentences.txt**
    - Biểu diễn các từ trong các document dưới dạng id của từ. Id của từ là index của từ đó trong file Wordlist. Các document được biểu diễn liên tiếp nhau, trong đó mỗi document được biểu diễn dưới dạng:
        - Số câu trong document - n. Ví dụ: **24**
        - n dòng tiếp theo, mỗi dòng là biểu diễn một câu, bao gồm các từ đứng lần lượt. Ví dụ

          152 158 120 159 18

          Câu trên gồm 5 từ, có id của từ là 152, 158, 120, 159, 18
2. **POS.txt**
    - Cấu trúc tương tự file **BagOfSentences.txt**, thay vì là id của word thì tương ứng là **Postag** của word đó.
3. **sentidoc.txt**
    - Nhãn sentiment của lần lượt từng document, theo thứ tự document trong file BagOfSentences.txt và POS.txt
        - label = 1: document đó được đánh giá là **positive**
        - label = 0: document đó được đánh giá là **neutral**
        - label = -1: document đó được đánh giá là **negative**

      Chú ý: ở phạm vi test của bài toán, **không sử dụng tới nhãn neutral**, tức là loại bỏ các document có nhãn là neutral ở bước tính accuracy. (Vẫn sử dụng ở bước Training của mô hình)
4. **SentiWords-0.txt**
    - Danh sách các từ **Positive** - tri thức Prior
5. **SentiWords-1.txt**
    - Danh sách các từ **Negative** - tri thức Prior
6. **WordList.txt**
    - Danh sách từ điển. Mỗi dòng là một từ, được **đánh id từ 0.**

      Ví dụ, trong bộ data **Amazon**, từ **Good** có id là 0, từ **nice** có id là 1. id này được sử dụng trong file **BagOfSentences.txt**

Giải thích một số biến quan trọng

- **C_ds:**
  - số từ trong document d được gán sentiment S.
  - IntegerMatrix(numDoc, NUM_SENTIMENTS + 1)
- **C_st:**
  - số từ được gán sentiment S và aspect T (trong toàn dataset)
  - IntegerMatrix(NUM_SENTIMENTS + 1, NUM_ASPECTS)
- **C_dst**
  - C_dst[i] : Số từ trong document i được gán sentiment s và aspect t
  - IntegerMatrix(NUM_SENTIMENTS + 1, NUM_ASPECTS)[numDoc]
- **C_stw**
  - C_stw[i] : Số lần từ w được gán sentiment S và topic T (trong toàn dataset)
  - IntegerMatrix(NUM_ASPECTS, V)[NUM_SENTIMENTS + 1]
- **S**
  - Sentiment được gán cho từng từ trong từng document
  - ArrayList<ArrayList<Integer>>()
- **Z**
  - Topic được gán cho từng từ trong từng document
  - ArrayList<ArrayList<Integer>>()

- **C_ds_mean:**
  - số từ trong document d được gán sentiment S trong tất cả các lần sampling
  - IntegerMatrix(numDoc, NUM_SENTIMENTS + 1)
- **C_st_mean:**
  - số từ được gán sentiment S và aspect T trong tất cả các lần sampling(trong toàn dataset)
  - IntegerMatrix(NUM_SENTIMENTS + 1, NUM_ASPECTS)
- **C_dst_mean**
  - C_dst_mean[i] : Số từ trong document i được gán sentiment s và aspect t trong tất cả các lần sampling
  - IntegerMatrix(NUM_SENTIMENTS + 1, NUM_ASPECTS)[numDoc]
- **C_stw_mean**

    - C_stw_mean[i] : Số lần từ w được gán sentiment S và topic T trong tất cả các lần sampling (trong toàn dataset)
    - IntegerMatrix(NUM_ASPECTS, V)[NUM_SENTIMENTS + 1]

Các tham số trong file settings.txt có ý nghĩa:

- **NEUTRAL**: tham số tiên nghiệm của từ không có trong danh sách từ cảm xúc biết trước
- **SAME_TYPE**: tham số tiên nghiệm của từ cảm xúc tốt/xấu với cảm xúc tốt/xấu tương ứng
- **DIFFERENT_TYPE**: tham số tiên nghiệm của từ cảm xúc tốt/xấu với cảm xúc ngược lại xấu/tốt
- **NUM_SENTIMENTS**: số lượng cảm xúc
- **NUM_ASPECTS**: số lượng chủ đề
- **MINIBATCH_SIZE**: kích thước một mini-batch
- **DOC_FILE** = BagOfSentences.txt: file dữ liệu theo cấu trúc bag of sentence
- **POS_FILE** = POS.txt: file từ loại theo cấu trúc bag of sentence
- **WORD_LIST_FILE** = WordList.txt: file từ điển
- **POSITIVE_FILE** = SentiWords-0.txt: danh sách từ tích cực
- **NEGATIVE_FILE** = SentiWords-1.txt: danh sách từ tiêu cực
- **LABEL_FILE** = sentidoc_train.txt: nhãn tích cực/tiêu cực của các văn bản
- **burn_in**: số lần burn-in trong Gibbs Sampling
- **sampling**: số lần lấy mẫu sampling trong Gibbs Sampling
- **k**, **tau0**, **kappa**, **gamma**: các tham số của mô hình

Cách thức sử dụng chương trình:

- Sau khi đặt dữ liệu và settings vào đúng thư mục Data và Settings, chỉnh số lượng file settings và bộ dữ liệu trong file main/RunStreamingLearning.java
- Chạy file main/RunStreamingLearning.java
- Kết quả sẽ được ghi vào các folder Settings tương ứng trong folder output.