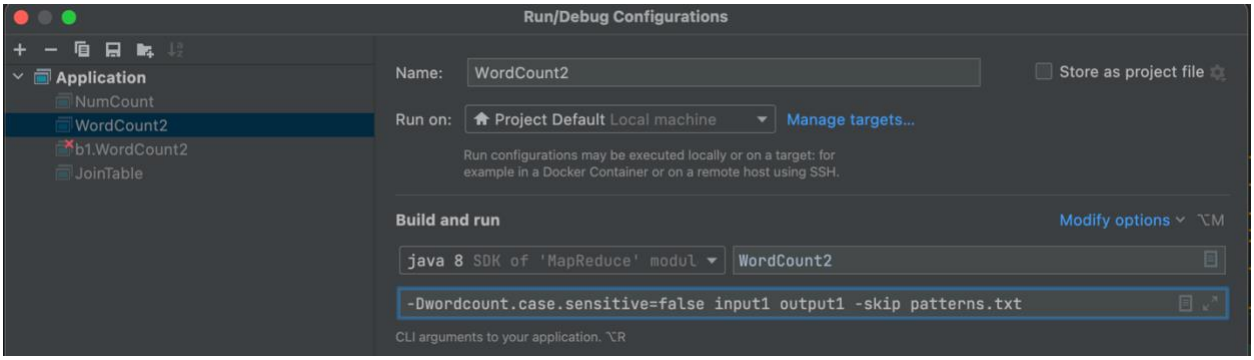
Bài 1: WordCount

- Em sử dụng ví dụ WordCount2 trên trang chủ apache
  Apache Hadoop 3.3.3 – MapReduce Tutorial

- **Lưu ý: Để chạy được trên Intellij cũng như trên HDFS các folder ouput cần phải chưa tồn tại nên khi chạy code phải xoá các folder output đã có hoặc sửa lại tên folder output**

- **Trên server trong folder /home/hadoop/anhlq36/b1**
  - Hai câu lệnh chạy như sau:
    - yarn jar WordCount.jar WordCount2 /anhlq36/b1/input/in.txt /anhlq36/b1/output -skip /anhlq36/b1/patterns.txt
    - -skip để bỏ qua các kí tự trong file patterns.txt
    - yarn jar WordCount.jar WordCount2 -Dwordcount.case.sensitive=false /anhlq36/b1/input/alice.txt /anhlq36/b1/output1 -skip /anhlq36/b1/patterns.txt
    - -Dwordcount.case.sensitive=false không phân biệt chữ hoa và chữ thường
  - Output nằm ở phần em gạch chân /part-r-0000 trên HDFS.
  - Em đã -get về thư mục **/home/hadoop/anhlq36/b1/** 2 file output của 2 câu lệnh trên

- Cách chạy trên Intellij cấu hình các argument như ảnh dưới:
  - -Dwordcount.case.sensitive=false input1 output1 -skip patterns.txt

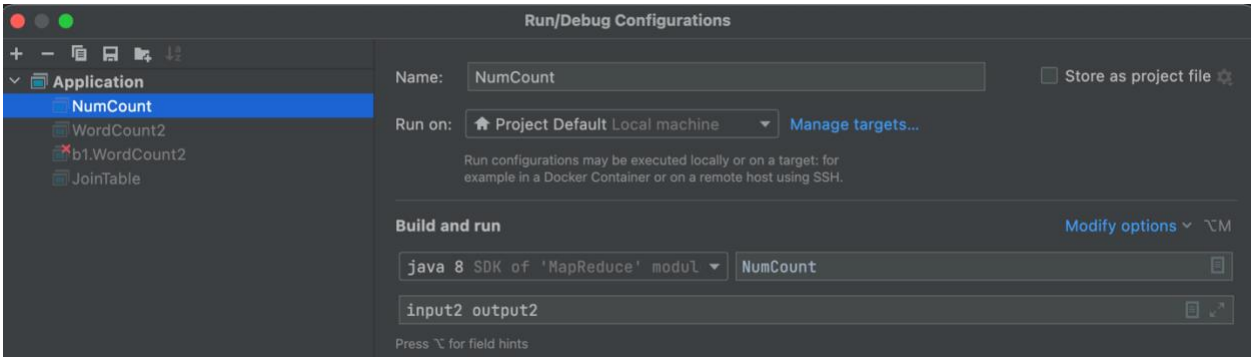


Bài 2: NumCount

- Em sử dụng ví dụ WordCount ở link trên và sửa code phần reduce. Em sử dụng biến int count và tăng biến này lên trong hàm reduce rồi ghi ra ở hàm cleanup.

- **Trên Server trong folder /home/hadoop/anhlq36/b2**
    - Câu lệnh chạy như sau:
        - yarn jar NumCount.jar NumCount /anhlq36/b2/input/count_distinct.csv /anhlq36/b2/output
    - Output nằm ở phần em gạch chân /part-r-0000 trên HDFS
    - Em đã -get về thư mục **/home/hadoop/anhlq36/b2/** file output của câu lệnh trên

- Cách chạy trên Intellij cấu hình các argument như ảnh dưới:
    - input2 output2



Bài 3: JoinTable

- Sử dụng 2 mapper cho 2 file csv để output ra <JobKey, JoinGenericWritable>
- Tạo 2 biến final SALARY_RECORD = 0 và PEOPLE_RECORD = 1 để lấy ra giá trị salary trước sau đó append vào các bản ghi people

- **Trên Server trong folder /home/hadoop/anhlq36/b3**

    - Câu lệnh chạy như sau:
        - yarn jar Join.jar JoinTable /anhlq36/b3/input/salary.csv /anhlq36/b3/input/people.csv /anhlq36/b3/output

    - Output nằm ở phần em gạch chân /part-r-0000 trên HDFS
    - Em đã -get về thư mục **/home/hadoop/anhlq36/b3/** file output của câu lệnh trên

- Cách chạy trên Intellij cấu hình các argument như ảnh dưới:
    - input3/salary.csv input3/people.csv output3