Báo cáo bài tập 2 phần 2:

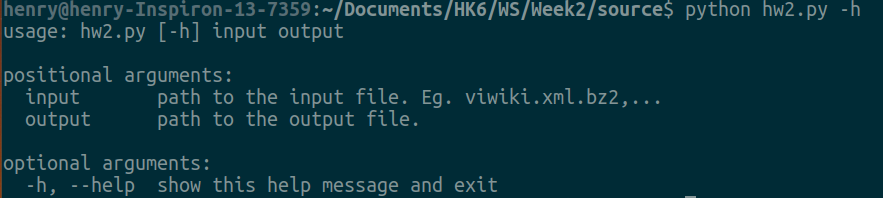
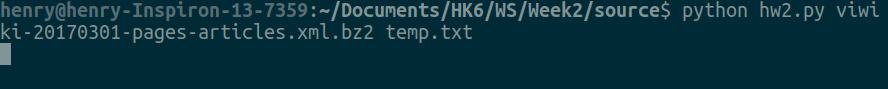
Thành viên:

|  |  |  |
| --- | --- | --- |
| MSSV | Họ và tên | Email |
| 1412161 | Lưu Đức Hiếu | [henryhieu1996@gmail.com](mailto:henryhieu1996@gmail.com) |
| 1412498 | Nguyễn Minh Thành |  |

Ngôn ngữ lập trình:

* + Python

Cách sử dụng chương trình:

* Với câu lệnh: “python hw2.py -h” để hiện thị hướng dẫn sử dụng
* 
* Với câu lệnh: “python hw2.py viwiki-20170301-pages-articles.xml.bz2 temp.txt” để chạy chương trình:
* 

**Cách thực hiện:**

* + Nhóm dùng thư viện BZ2File để giải nén file bz2
  + Sau đó dùng thư viện lxml để lấy từng dòng tag xây dựng thành cây etree
  + Với mỗi <page>, nhóm kiểm tra nếu ns = 0 thì sau đó mới xử lý tiếp, còn không bỏ qua page đó.
  + Với title và text lấy được từ 1 page, ta thực hiện quá trình tiền xử lý:
    - Trước tiên: xóa các heading và text trong đó
    - Xóa các thông tin trong ngoặc {...}, gồm các infobox và wikitable.
    - Xóa các tập tin và hình ảnh
    - Xóa các tham khảo
    - Xóa các đường link, nhưng giữ lại các thông tin hiển thị
    - Xóa các câu đặc biệt, #Redirect
    - Xóa các kí tự đặc biệt
  + Sau đó, nhóm dùng word\_counter để đếm tần suất của các từ.