Vũ Quang Minh

18110150

MÔN HỌC: PHÁT TRIỂN PHẦN MỀM HƯỚNG ĐỐI TƯỢNG

Đề 3. Thiết kế và cài đặt một ứng dụng cho phép phân tích một tập tin văn bản, xuất ra tần số xuất hiện của của các từ có trong tập tin văn bản. Các ký tự ngăn cách bao gồm khoảng trắng chuẩn và các ký tự ngăn cách câu (dấu ‘,’, ‘.’, ‘;’, ‘:’, ‘?’, ‘!’, ‘…’). Từ là từ có nghĩa trong tự điển từ tiếng Anh, có xét đến số ít, số nhiều, quá khứ (ví dụ: mouse, mice được tính là cùng một từ, study, studies, studied, studying được tính là cùng một từ).

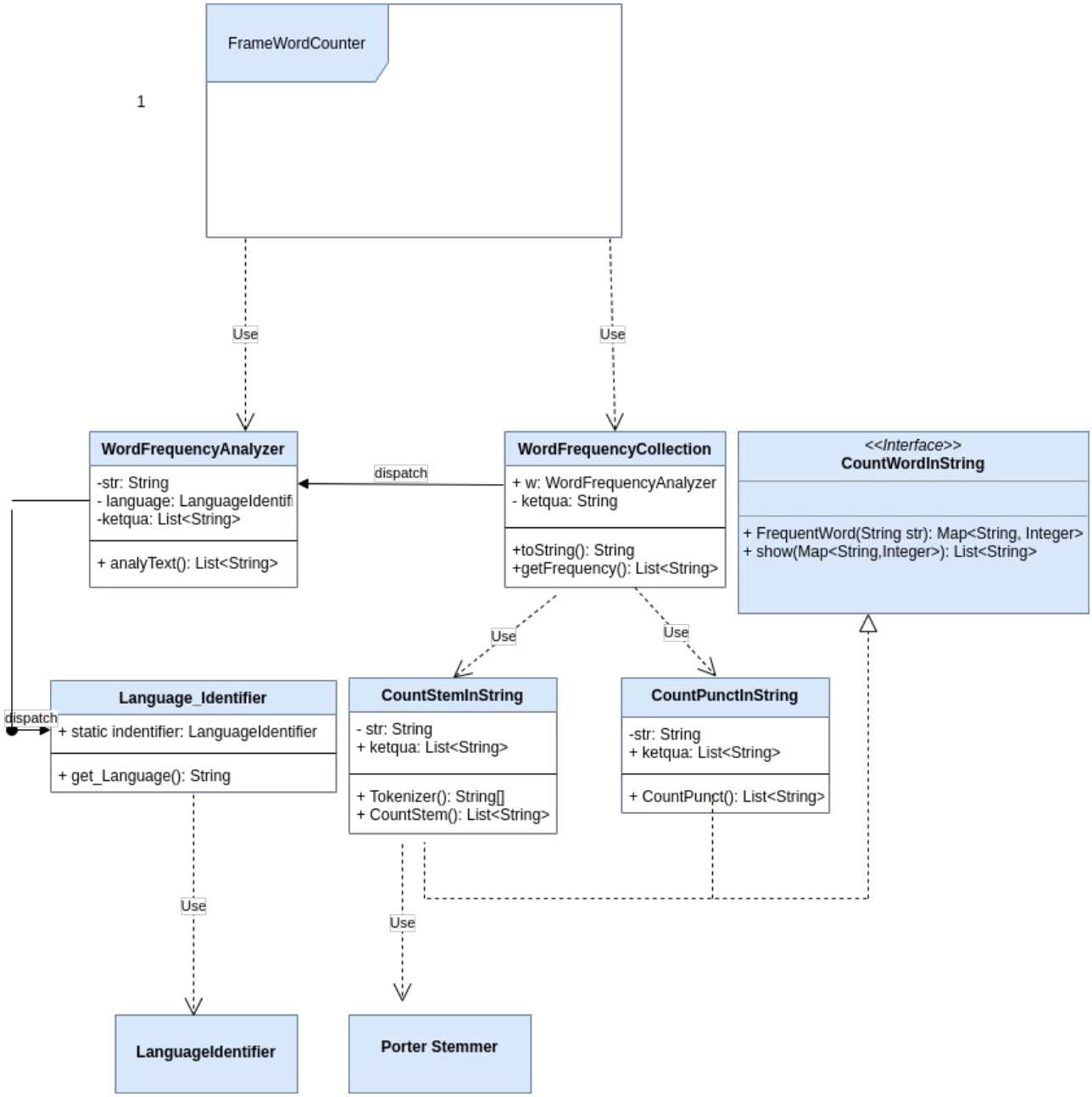
A) Specification

- Người dùng sẽ chọn ***FrameWordCounter*** và nhấn nút Run As Application để chạy chương trình.
- Chương trình sẽ cho mở file bất kì ở vị trí nào trên bô nhớ bằng nút(Browse) nhưng chỉ đọc được File txt nếu gặp file định dạng khác thì không đọc được.
- Chương trình cũng có phần xuất file tcũng được lưu ở vị trí bất kì trên bộ nhớ chỉ lưu được file dưới dạng txt. Nếu lưu dưới dạng khác thì sẽ không chạy được file.
- Sau khi mở file thì đường link sẽ được lưu trên ô Open File, và nội dung sẽ xuất hiện trên ô xuất nội dung.
- Nhấn nút ***Xuất kết quả*** để xử lý chương trình.
- Chương trình định dạng được ngôn ngữ tiếng Anh.
- Chương trình có thể đếm được các kí tự(,.?:;!…)
- Chương trình sẽ liệt kê tất cả các từ và các loại từ của chúng(ở đây là số ít, số nhiều, quá khứ, hiện tại, tương lai).
- Chương trình sẽ đếm các loại từ.

B) Phạm vi đề tài

- Chương trình định dạng được ngôn ngữ tiếng Anh.
- Chương trình có thể đếm được các kí tự(,.?:;!…).
- Chương trình sẽ liệt kê tất cả các từ và các loại từ của chúng(ở đây là số ít, số nhiều, quá khứ, hiện tại, tương lai).
- Chương trình sẽ đếm các loại từ.
- Chương trình có phần giao diện từ được hỗ trợ bởi *java1.8* và *WindowBuilder* được tải từ Marketplace.
- Chương trình có 3 thư viện được thêm vô *org.apache.tika* để định dạng ngôn ngữ, *opennlp-uima-1.9.4.jar và opennlp-tools-1.9.4.jar* để để xử lý các từ cùng loại.

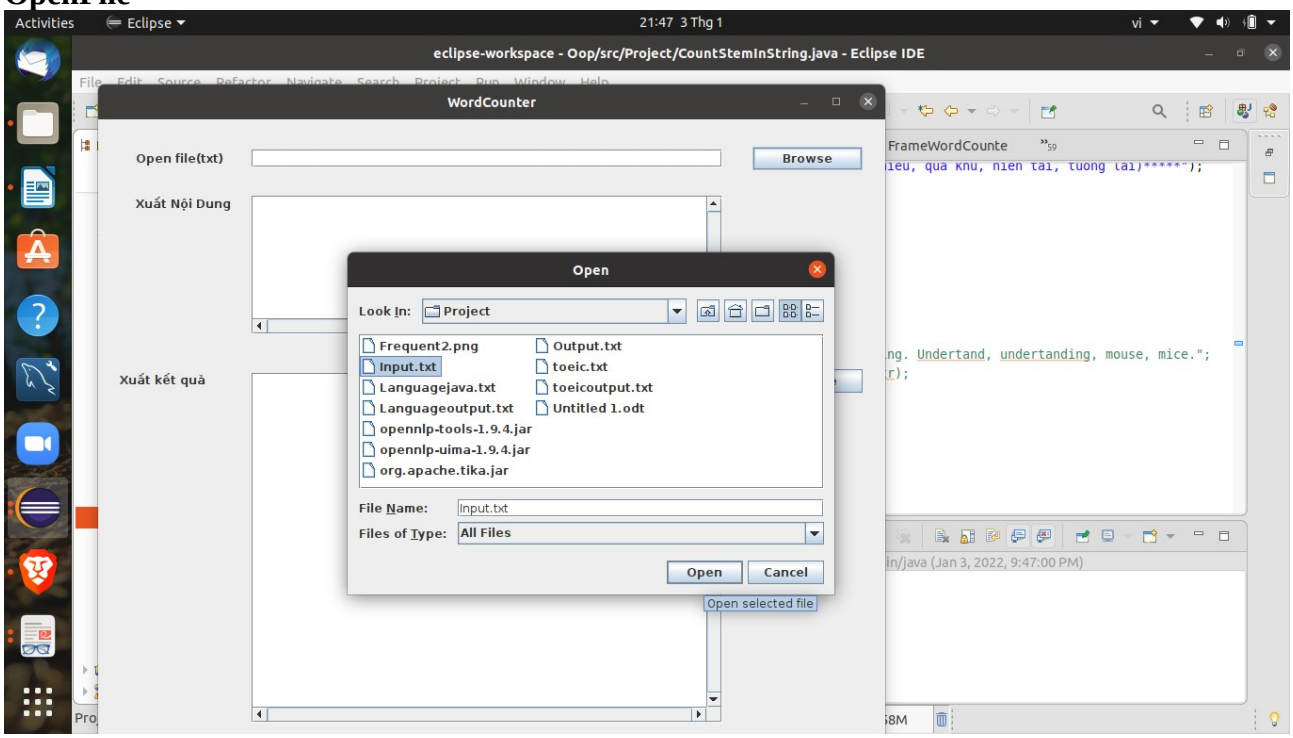
C) Sơ đồ và Code



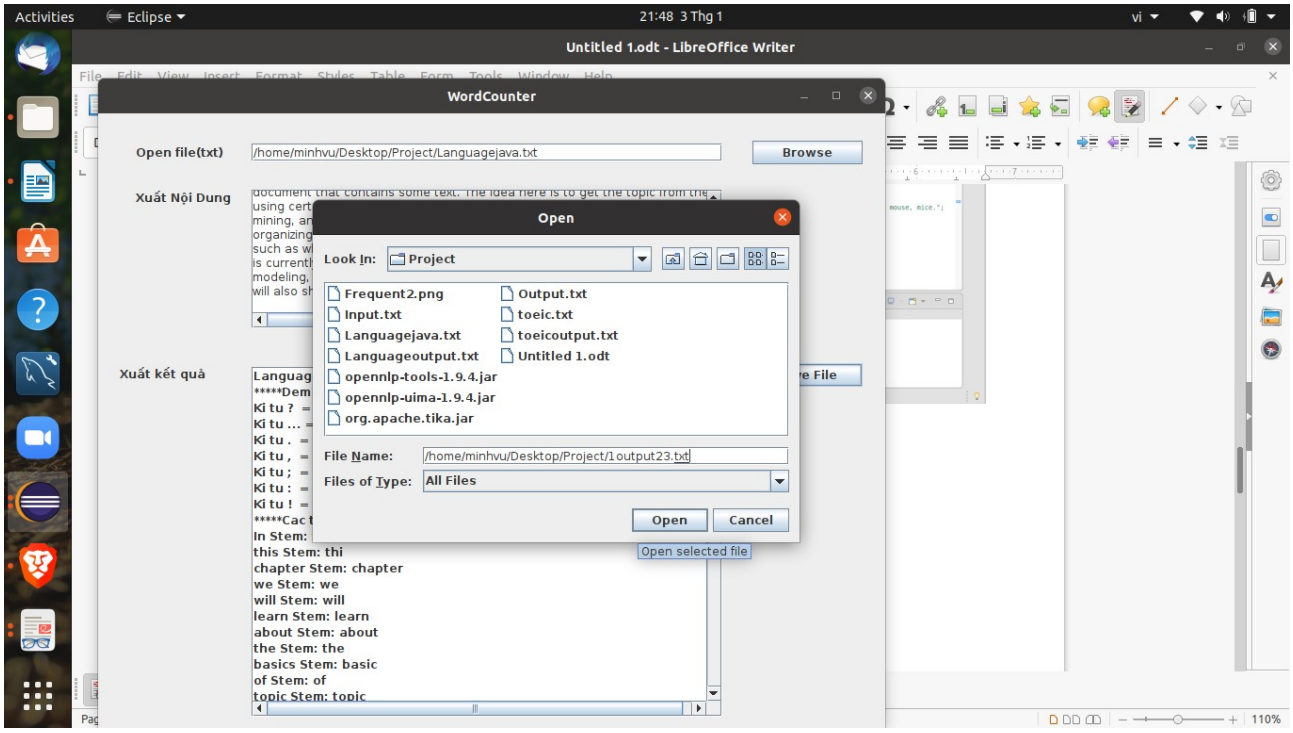
- Người dùng sẽ chọn *FrameWordCounter* để chạy chương trình.
- *FrameWordCounter* sẽ sử dụng *WordFrequencyCollection* để đếm kí tự và các từ cùng loại và *WordFrequencyAnalyzer* để định dạng ngôn ngữ.
- *WordFrequencyCollection* sẽ sử dụng *CountStemInString* để đếm các từ cùng loại và *CountPunctInString* để đếm các kí tự. *CountStemInString* sẽ sử dụng *PorterStemmer* để xử lấy các từ cùng loại và được hỗ trợ bởi thư viện opennlp
- Hai hàm *FrequentWord(String str)* và *show(Map<String, Integer>)* của *CountStemInString* và *CountPunctInString* được implement từ interface *CountWordInString*.
- *WordFrequencyCollection* sẽ được tham chiếu từ *WordFrequecyAnalyze.*
- *WordFrequencyAnalyze* khi khởi tạo sẽ tham chiếu đến *Language_Identifier* để định dạng ngôn ngữ.

D) Demo và hướng dẫn sử dụng

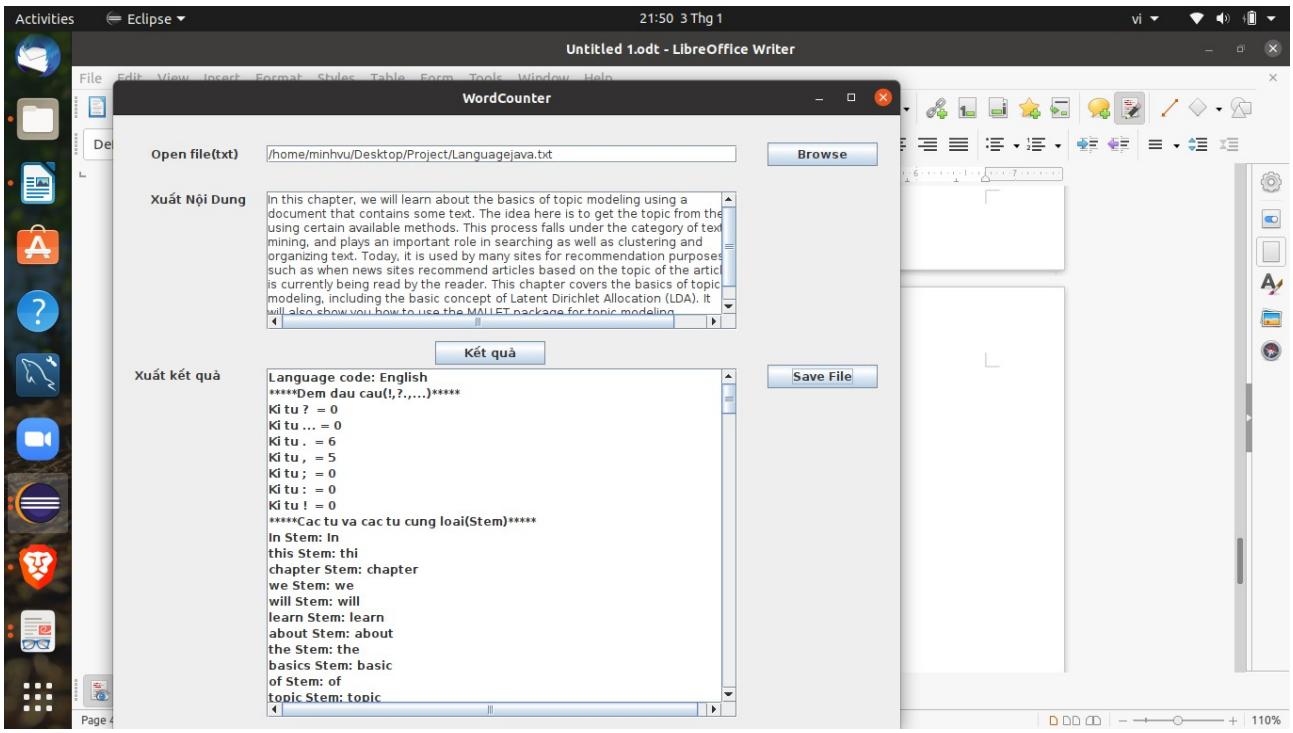
OpenFile

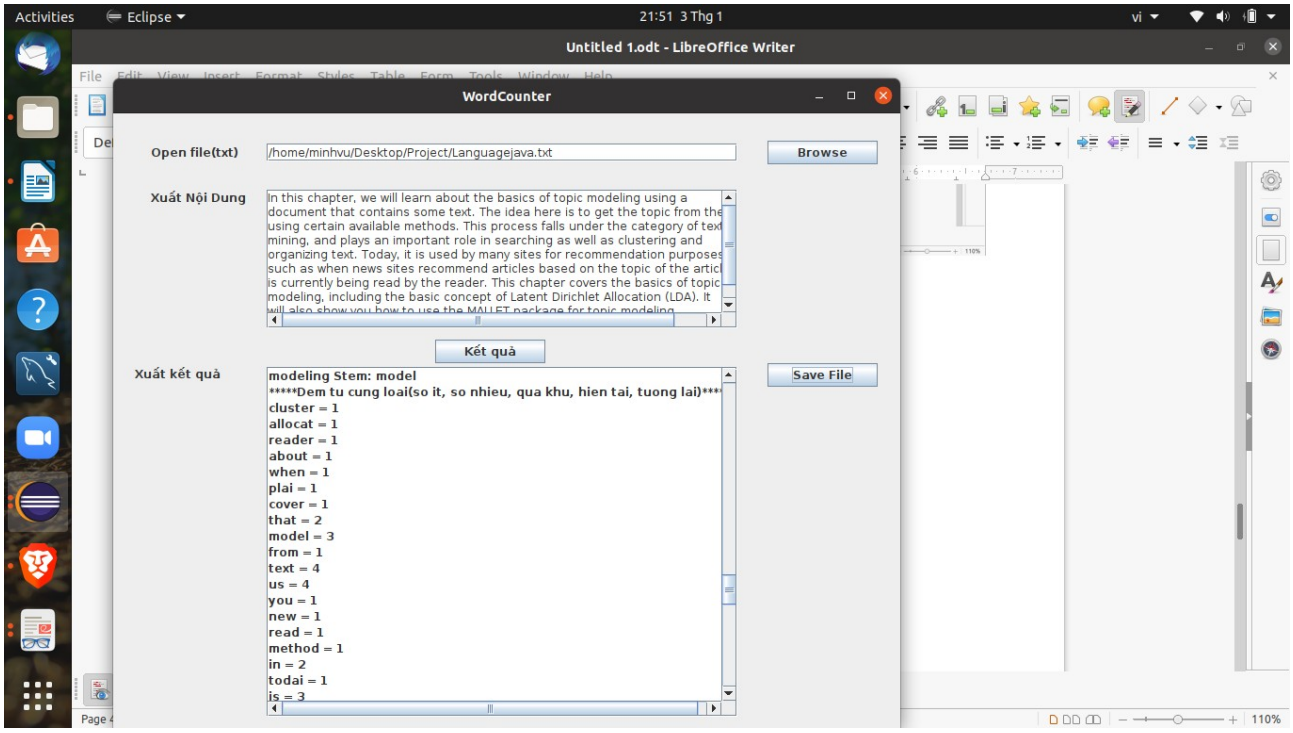


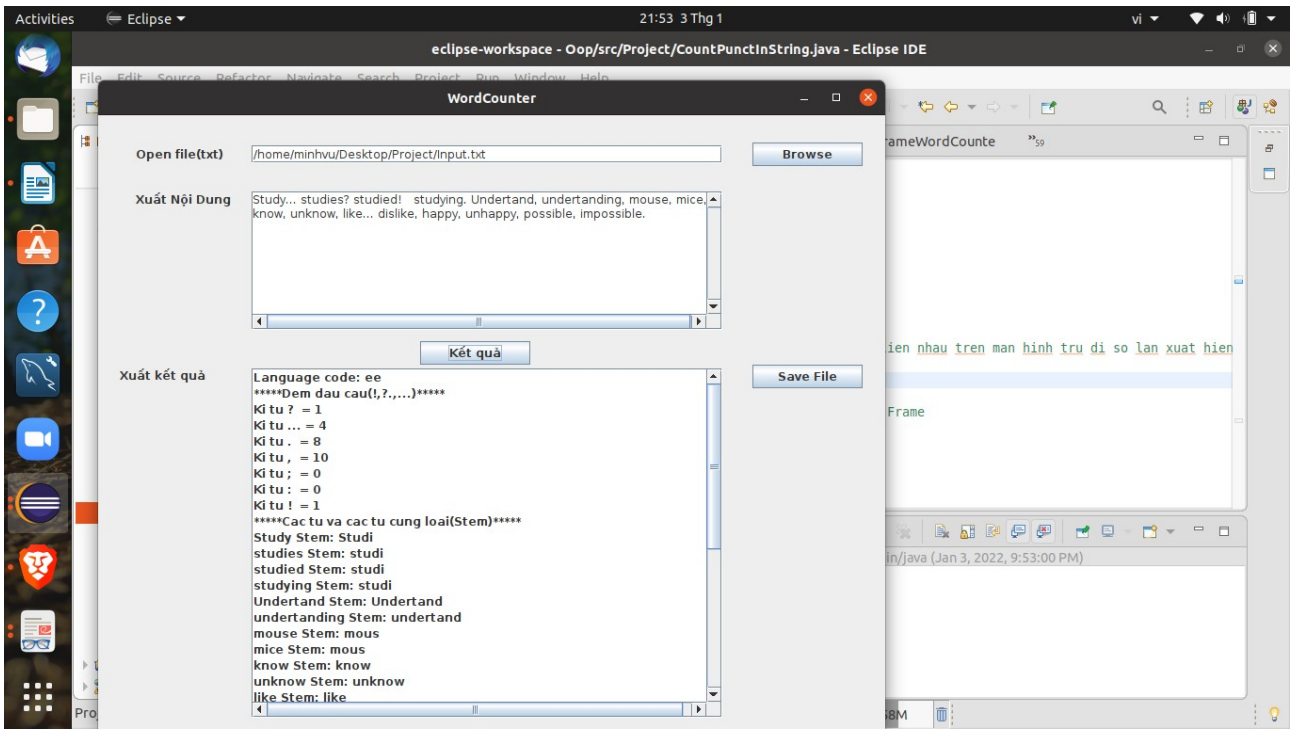
SaveFile

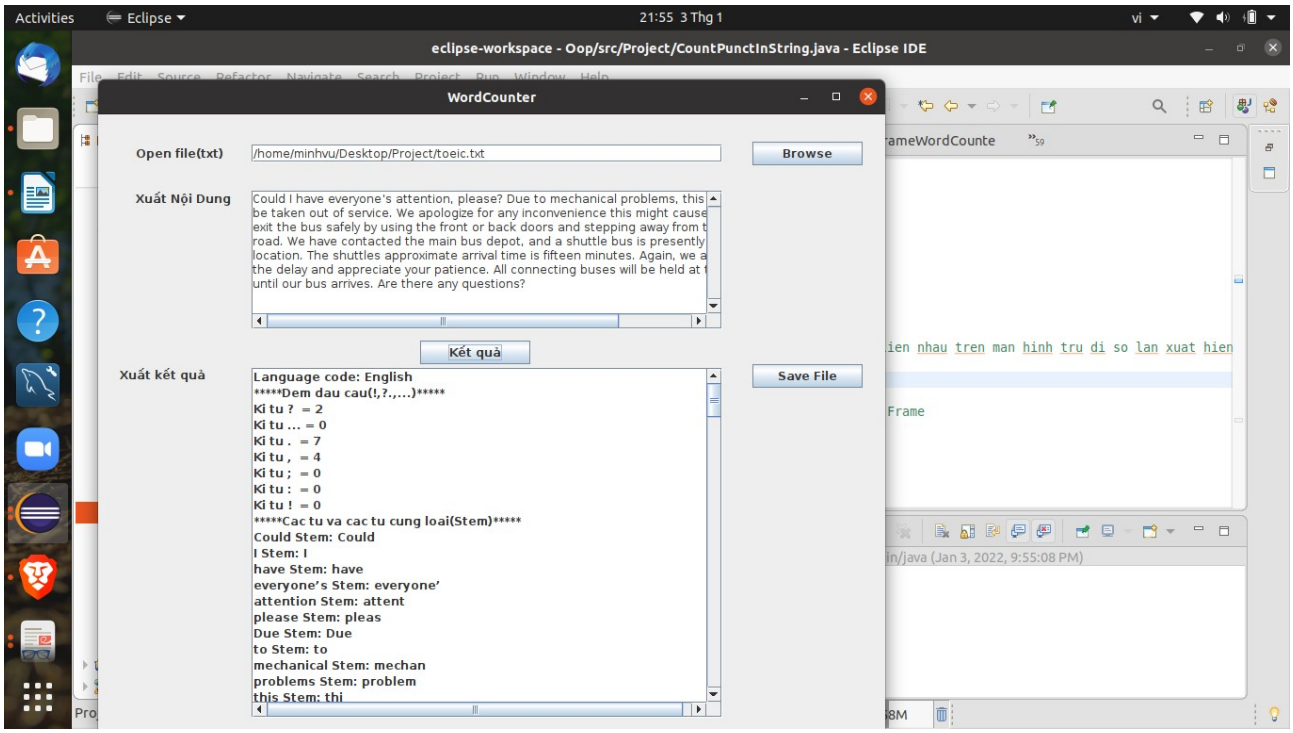


Demo vài đoạn chương trình


Activities
Eclipse
21:50 3 Thg 1
vi
Untitled 1.odt - LibreOffice Writer
WordCounter
Open file(txt)
/home/minhvu/Desktop/Project/Languagejava.txt
Browse
Xuất Nội Dung
In this chapter, we will learn about the basics of topic modeling using a document that contains some text. The idea here is to get the topic from the using certain available methods. This process falls under the category of text mining, and plays an important role in searching as well as clustering and organizing text. Today, it is used by many sites for recommendation purposes such as when news sites recommend articles based on the topic of the articl is currently being read by the reader. This chapter covers the basics of topic modeling, including the basic concept of Latent Dirichlet Allocation (LDA). It will also show you how to use the MALLET package for topic modeling
Kết quả
Xuất kết quả
Save File
Language code: English
*****Dem dau cau(!,?,,...)*****
Ki tu ? = 0
Ki tu ... = 0
Ki tu . = 6
Ki tu , = 5
Ki tu ; = 0
Ki tu : = 0
Ki tu ! = 0
*****Cac tu va cac tu cung loai(Stem)*****
In Stem: In
this Stem: thi
chapter Stem: chapter
we Stem: we
will Stem: will
learn Stem: learn
about Stem: about
the Stem: the
basics Stem: basic
of Stem: of
topic Stem: topic
110%


Activities
Eclipse
21:51 3 Thg 1
vi
Untitled 1.odt - LibreOffice Writer
WordCounter
Open file(txt)
/home/minhvu/Desktop/Project/Languagejava.txt
Browse
Xuất Nội Dung
In this chapter, we will learn about the basics of topic modeling using a document that contains some text. The idea here is to get the topic from the using certain available methods. This process falls under the category of text mining, and plays an important role in searching as well as clustering and organizing text. Today, it is used by many sites for recommendation purposes such as when news sites recommend articles based on the topic of the articl is currently being read by the reader. This chapter covers the basics of topic modeling, including the basic concept of Latent Dirichlet Allocation (LDA). It will also show you how to use the MALLET package for topic modeling
Kết quả
Xuất kết quả
Save File
modeling Stem: model
*****Dem tu cung loai(so it, so nhieu, qua khu, hien tai, tuong lai)*****
cluster = 1
allocat = 1
reader = 1
about = 1
when = 1
plai = 1
cover = 1
that = 2
model = 3
from = 1
text = 4
us = 4
you = 1
new = 1
read = 1
method = 1
in = 2
todai = 1
is = 3
110%


Activities
Eclipse
21:53 3 Thg 1
vi
eclipse-workspace - Oop/src/Project/CountPunctInString.java - Eclipse IDE
WordCounter
Open file(txt)
/home/minhvu/Desktop/Project/Input.txt
Browse
Xuất Nội Dung
Study... studies? studied! studying. Undertand, undertanding, mouse, mice, know, unknow, like... dislike, happy, unhappy, possible, impossible.
Kết quả
Xuất kết quả
Save File
Language code: ee
*****Dem dau cau(!,?,...)*****
Ki tu ? = 1
Ki tu ... = 4
Ki tu . = 8
Ki tu , = 10
Ki tu ; = 0
Ki tu : = 0
Ki tu ! = 1
*****Cac tu va cac tu cung loai(Stem)*****
Study Stem: Studi
studies Stem: studi
studied Stem: studi
studying Stem: studi
Undertand Stem: Undertand
undertanding Stem: undertand
mouse Stem: mous
mice Stem: mous
know Stem: know
unknow Stem: unknow
like Stem: like


Activities
Eclipse
21:55 3 Thg 1
vi
eclipse-workspace - Oop/src/Project/CountPunctInString.java - Eclipse IDE
WordCounter
Open file(txt)
/home/minhvu/Desktop/Project/toeic.txt
Browse
Xuất Nội Dung
Could I have everyone's attention, please? Due to mechanical problems, this be taken out of service. We apologize for any inconvenience this might cause exit the bus safely by using the front or back doors and stepping away from t road. We have contacted the main bus depot, and a shuttle bus is presently location. The shuttles approximate arrival time is fifteen minutes. Again, we a the delay and appreciate your patience. All connecting buses will be held at t until our bus arrives. Are there any questions?
Kết quả
Xuất kết quả
Save File
Language code: English
*****Dem dau cau(!,?,...)*****
Ki tu ? = 2
Ki tu ... = 0
Ki tu . = 7
Ki tu , = 4
Ki tu ; = 0
Ki tu : = 0
Ki tu ! = 0
*****Cac tu va cac tu cung loai(Stem)*****
Could Stem: Could
I Stem: I
have Stem: have
everyone's Stem: everyone'
attention Stem: attent
please Stem: pleas
Due Stem: Due
to Stem: to
mechanical Stem: mechan
problems Stem: problem
this Stem: thi