DSA II Final Project

Introduction & Guidelines

OVERVIEW

- KWIC (Key Word in Context) searcher
- General idea:
  Search in text from a document or a website to show the word in context with additional linguistic information

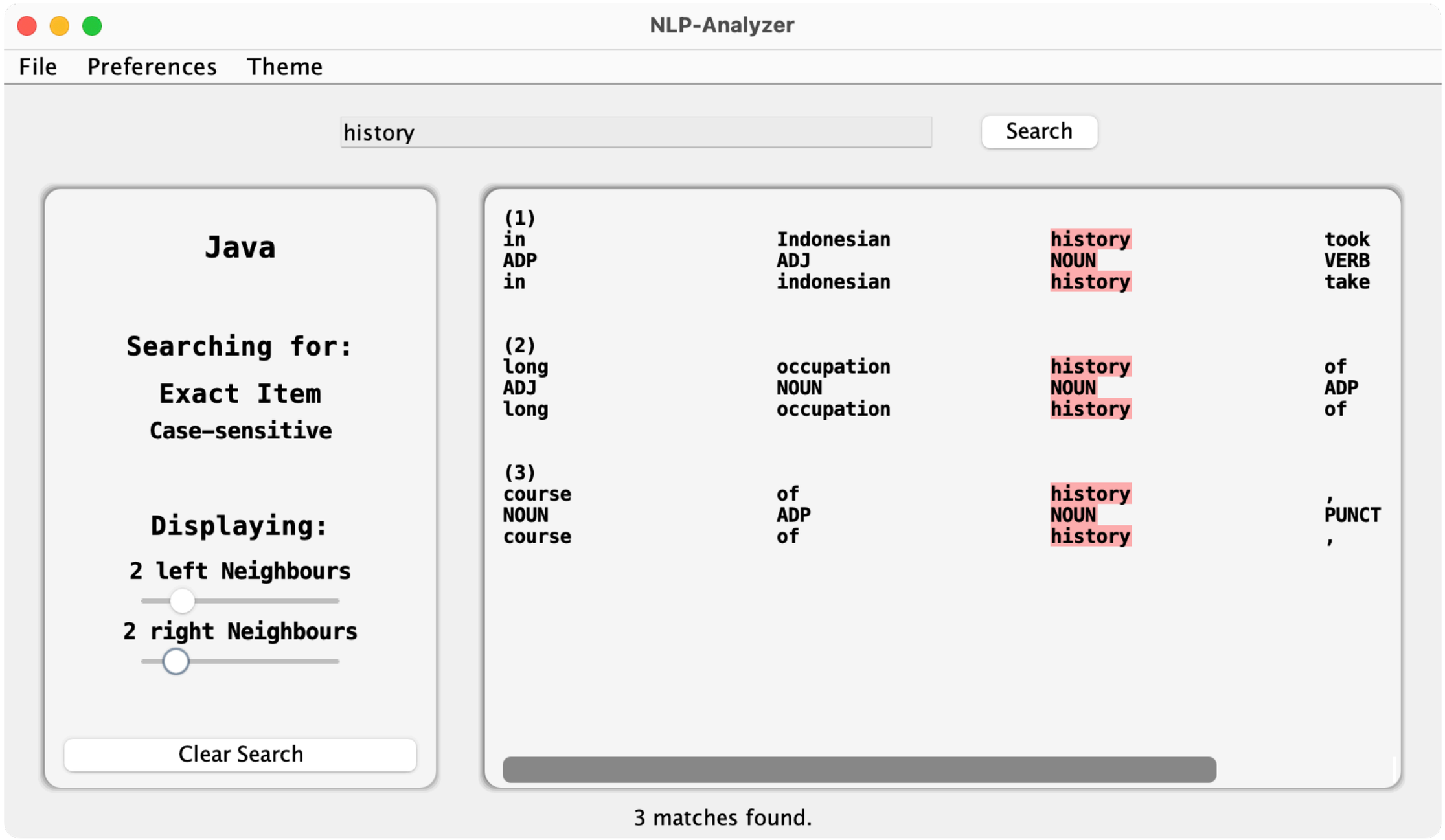
OVERVIEW

Project parts

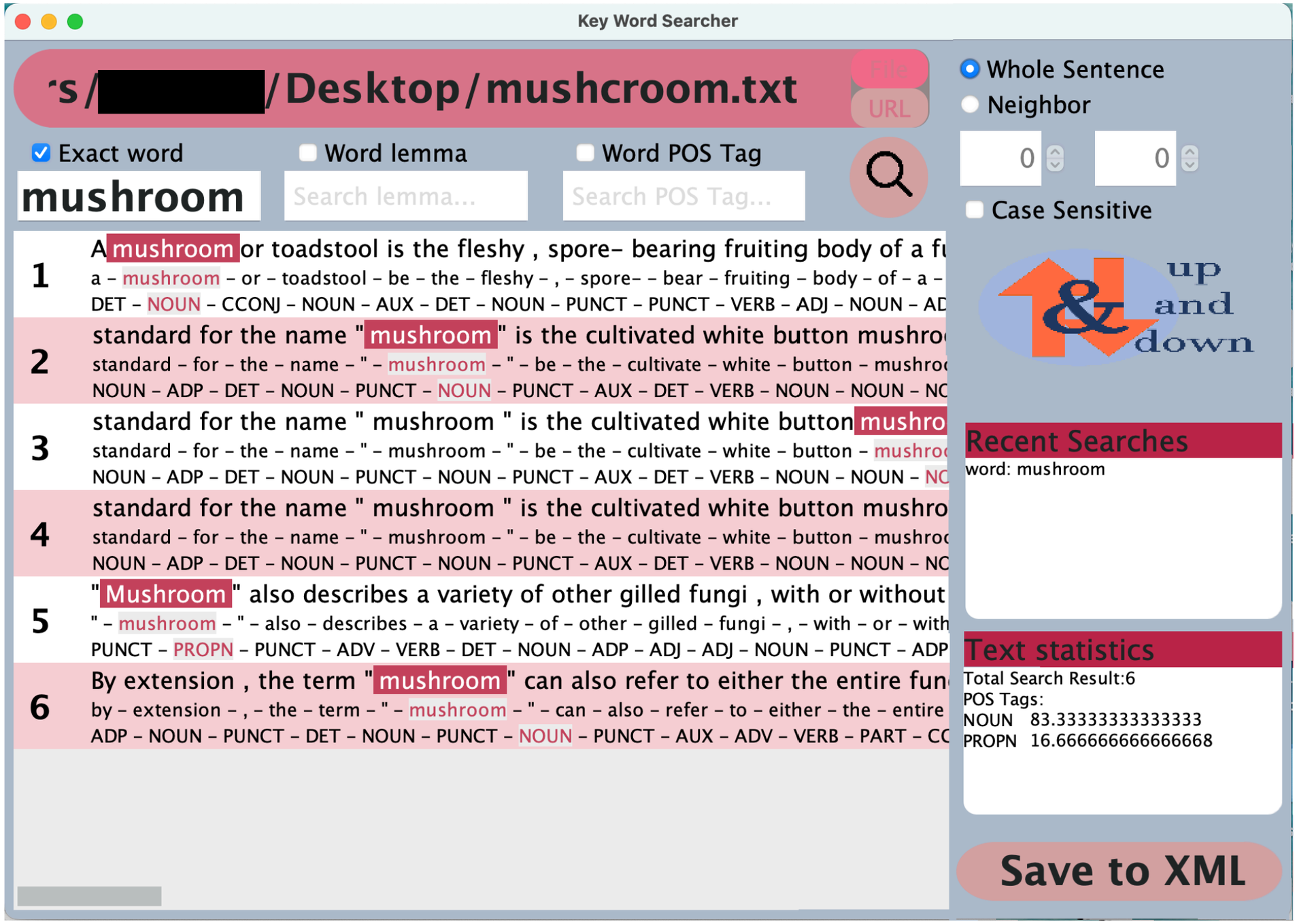
- GUI (with SWING) 15p
- Linguistic Processing (with OpenNLP) 10p
- Web scraping (with JSoup) 10p
- Save to file (XML) 10p
- Create executable JAR file 5p

You can work on multiple parts at the same time.
They do not need to be done in this order.

Example


NLP-Analyzer
File
Preferences
Theme
history
Search
Java
Searching for:
Exact Item
Case-sensitive
Displaying:
2 left Neighbours
2 right Neighbours
Clear Search
(1)
in Indonesian history took
ADP ADJ NOUN VERB
in indonesian history take
(2)
long occupation history of
ADJ NOUN NOUN ADP
long occupation history of
(3)
course of history ,
NOUN ADP NOUN PUNCT
course of history ,
3 matches found.

Example


Key Word Searcher
's/[redacted]/Desktop/mushcroom.txt
File
URL
Whole Sentence
Neighbor
0
0
Exact word
Word lemma
Word POS Tag
mushroom
Search lemma...
Search POS Tag...
Case Sensitive
1
A mushroom or toadstool is the fleshy , spore- bearing fruiting body of a fu
a – mushroom – or – toadstool – be – the – fleshy – , – spore- – bear – fruiting – body – of – a –
DET – NOUN – CCONJ – NOUN – AUX – DET – NOUN – PUNCT – PUNCT – VERB – ADJ – NOUN – AD
2
standard for the name " mushroom " is the cultivated white button mushroo
standard – for – the – name – " – mushroom – " – be – the – cultivate – white – button – mushroo
NOUN – ADP – DET – NOUN – PUNCT – NOUN – PUNCT – AUX – DET – VERB – NOUN – NOUN – NO
3
standard for the name " mushroom " is the cultivated white button mushro
standard – for – the – name – " – mushroom – " – be – the – cultivate – white – button – mushroo
NOUN – ADP – DET – NOUN – PUNCT – NOUN – PUNCT – AUX – DET – VERB – NOUN – NOUN – NO
4
standard for the name " mushroom " is the cultivated white button mushro
standard – for – the – name – " – mushroom – " – be – the – cultivate – white – button – mushroo
NOUN – ADP – DET – NOUN – PUNCT – NOUN – PUNCT – AUX – DET – VERB – NOUN – NOUN – NO
5
" Mushroom " also describes a variety of other gilled fungi , with or without
" – mushroom – " – also – describes – a – variety – of – other – gilled – fungi – , – with – or – with
PUNCT – PROPN – PUNCT – ADV – VERB – DET – NOUN – ADP – ADJ – ADJ – NOUN – PUNCT – ADP
6
By extension , the term " mushroom " can also refer to either the entire fun
by – extension – , – the – term – " – mushroom – " – can – also – refer – to – either – the – entire
ADP – NOUN – PUNCT – DET – NOUN – PUNCT – NOUN – PUNCT – AUX – ADV – VERB – PART – CC
up and down
Recent Searches
word: mushroom
Text statistics
Total Search Result:6
POS Tags:
NOUN 83.33333333333333
PROPN 16.666666666666668
Save to XML

GUI

Graphical User Interface

- Swing
- Well designed and user-friendly
- Exception handling
  upon facing unexpected/invalid input, provide appropriate feedback to the user (don't just say "Error!", explain what's wrong and give the user the opportunity to resolve the issue).

OPEN NLP

Linguistic Processing

- Identifying sentences, word tokens, lemmas and POS tags.
- Search in the text by either word, lemma or POS tag.
- Basically OpenNLP SelfTest.

JSOUP

Web scraping

- Extract the text from a wikipedia article based on a link provided by the user
- Clean up the text - remove ‘junk’ like footnotes and metadata
- Result can be further processed by OpenNLP

XML

Save result to a file

- Save output to a new or existing XML file
- Possible from the UI - dialogue window to specify the file name
- Results from **all** searches in the session

JAR

Make your app into an executable file

- JAR with dependencies
  - Your app as an independant executable
  - Runs also outside of your project folder
- Include it in your submission in addition to your code

JAR

Useful tips in case of issues

- check that your pom.xml (maven project file) is correct and includes all the libraries.
- if you are using external media files in your GUI, make sure to read them as input streams:

```
Image programIcon;
try {
    programIcon = ImageIO.read(Objects.requireNonNull(
            getClass().getClassLoader().getResource( name: "icons/programIcon.png")
    ));
} catch (IOException e) {
    throw new RuntimeException("IO Exception: failed to load program icon.");
}
mainWindow.setIconImage(programIcon);
```

USING GITHUB AS A TEAM

- More people push modifications to the same file => “merge conflicts” (pain to resolve)
- So get used to do *git pull* and *git push* before and right after you do any changes to the code.
  - *git pull*: downloads the changes done by other group members.
  - *git push*: uploads your own changes.

CHECKPOINTS

- For giving advice, checking the progress and also helping you keep up with the flow of the project.

  projects are meant to be independent work,
  so there won't be much debugging from our side.

- During lab hours on Fridays

- **All group members** present what they have done so far

REPORTS

- Before each checkpoint:
  - Write a short report in PDF describing what each group member has done.
  - Add it to your project repositories
  - Push all the changes and new code described in the report
- Deadline: **Thursdays at 6:00PM**
  before the checkpoint sessions for your group

CHECKPOINT DATES

- 07.06: groups 1-6
- 14.06: groups 7-12
- 21.06: groups 1-6
- 28.06: groups 7-12
- 5.07: all & Q&A before the exam
- 12.07: DSA II Exam
- 19.07: all

GRADING

- The score may be different for different team members!
  - communicate and help each other,
  - divide the responsibilities equally and outline them in your final report.
- Bonus points for implementing additional features

Not submitting a report every two weeks will result in -2p each time

YOUR OWN PROJECT IDEA

- You can propose us to do something else as a project, as long as the following is guaranteed:
  - All the team members agree.
  - The workload and task types are similar (GUI, external libraries etc.).
  - NLP-related.
  - You describe us your idea and we approve it.

Questions?