PinyinGrep

Yanwen from Group Alpha

HU Ke, Kirby Fung, Wendy Wu

CONTENTS

1. Introduction
2. Functionality
3. Highlight
4. Procedure
5. Future work
6. Acknowledgement

Introduction

/01

Introduction: PinyinGrep

- **Background:**
  - Corpora are large in sizes(in TB).
  - Parabix provides efficient parallelism tools.
- **PinyinGrep:**
  - a grep program using regex-like pinyin syllables to grep Chinese characters from a list of files.
- **Possible Users:**
  - Linguists and professional scholars who need to process a large number of Chinese corpora.
  - Engineers who need to do data cleaning on numerous Chinese corpora.

Introduction: Design Philosophy

- **No news is good news.**
  - Inherit from UNIX philosophy.
  - Common for the grep program in Linux.
- **Being customer-oriented.**
  - Bother users as few as possible.
  - Robustness makes convenience.
  - Trust users and give them more flexibility.
- **Design in an extensible way.**
  - Consider extensibility at the beginning.

Functionality

/02

Functionality: Basic Functionality

```
$ pinyingrep <pinyins> <list of files>
```

Robustly supports pinyin inputs in various format:

- **Individual Pinyin Syllables Support:**
  - Pinyin syllables in alphabetic characters(“`ü`” and “ê” as exception) without tones specified, e.g. zhong.
  - Alphabetic pinyin syllables with tones specified by Arabic numbers(`0-4`), e.g. zhong1
  - Pinyin syllables with tones specified by toned characters in Unicode, more specifically, Latin characters like “ǎ” or “ō”, e.g. zhōng.
  - Arbitrarily embedded upper case characters(including upper case Latin characters with two types of encodings), e.g. ZHŌng

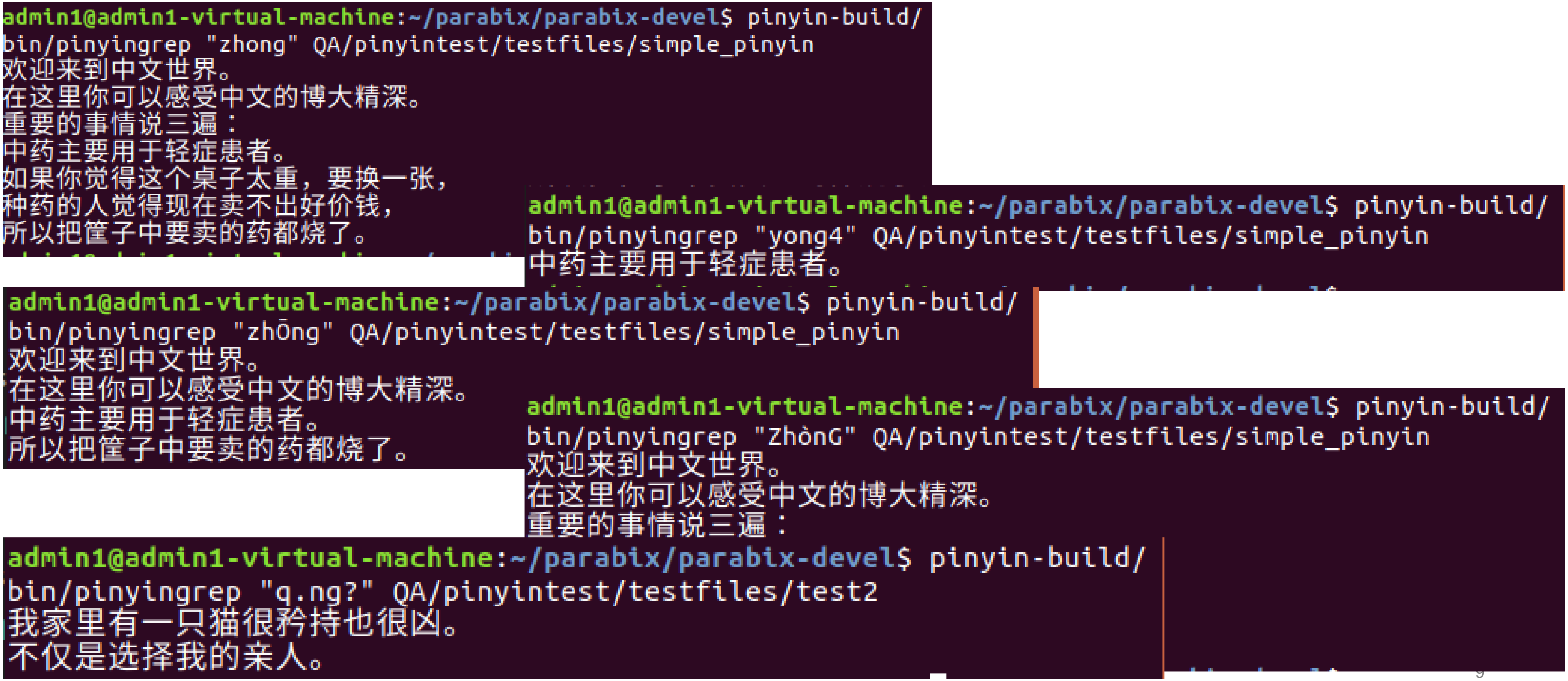
Functionality: Basic Functionality

- **Regex-like syllables support:**
  - Regular expression feature “.”, equivalent to arbitrary alphabetic characters(including “v”) ; ultimately, only legitimate syllables will get Chinese characters matched.
  - E.g. “zh.ng” is equivalent to “zhong” or “zhang” or “zheng” …..

  - Regular expression feature “?”, making the previous character alternative.
  - E.g. “zh?ang” is equivalent to “zang” or “zhang”.

- **Sequences of syllables support:**
  - Sequences of syllables in arbitrary lengths, e.g. “gong xi fa cai”.

Functionality: Basic Functionality



```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "zhong" QA/pinyintest/testfiles/simple_pinyin
欢迎来到中文世界。
在这里你可以感受中文的博大精深。
重要的事情说三遍：
中药主要用于轻症患者。
如果你觉得这个桌子太重，要换一张，
种药的人觉得现在卖不出好价钱，
所以把筐子中要卖的药都烧了。
```

```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "yong4" QA/pinyintest/testfiles/simple_pinyin
中药主要用于轻症患者。
```

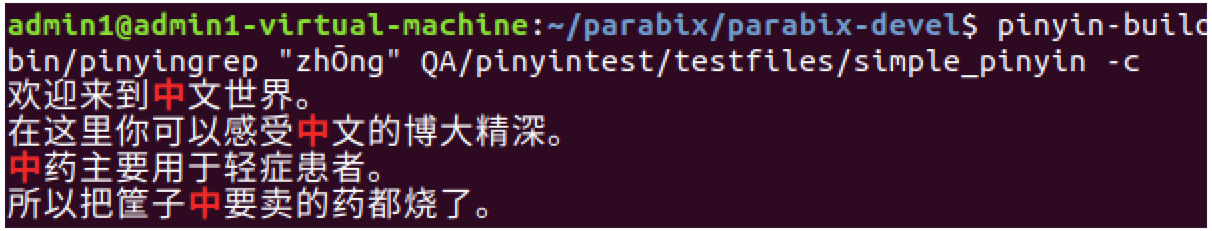
```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "zhŌng" QA/pinyintest/testfiles/simple_pinyin
欢迎来到中文世界。
在这里你可以感受中文的博大精深。
中药主要用于轻症患者。
所以把筐子中要卖的药都烧了。
```

```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "ZhònG" QA/pinyintest/testfiles/simple_pinyin
欢迎来到中文世界。
在这里你可以感受中文的博大精深。
重要的事情说三遍：
```

```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "q.ng?" QA/pinyintest/testfiles/test2
我家里有一只猫很矜持也很凶。
不仅是选择我的亲人。
```

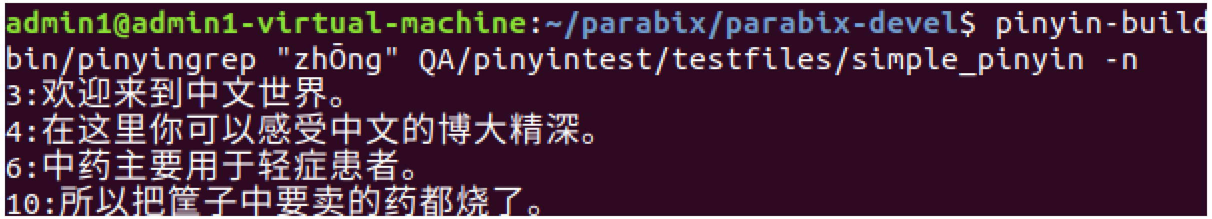
Functionality: Grep Features

- **Coloring:**
  - Command Line Flag “-c”



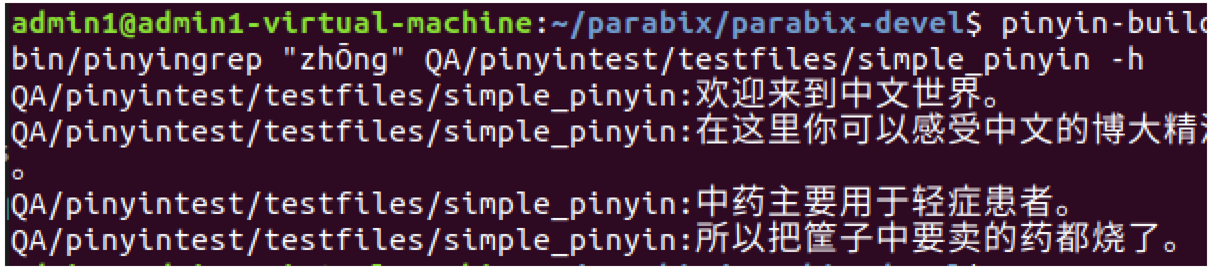
```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-buil
bin/pinyingrep "zhŌng" QA/pinyintest/testfiles/simple_pinyin -c
欢迎来到中文世界。
在这里你可以感受中文的博大精深。
中药主要用于轻症患者。
所以把筐子中要卖的药都烧了。
```

- **Line numbers:**
  - Command Line Flag “-n”



```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build
bin/pinyingrep "zhŌng" QA/pinyintest/testfiles/simple_pinyin -n
3:欢迎来到中文世界。
4:在这里你可以感受中文的博大精深。
6:中药主要用于轻症患者。
10:所以把筐子中要卖的药都烧了。
```

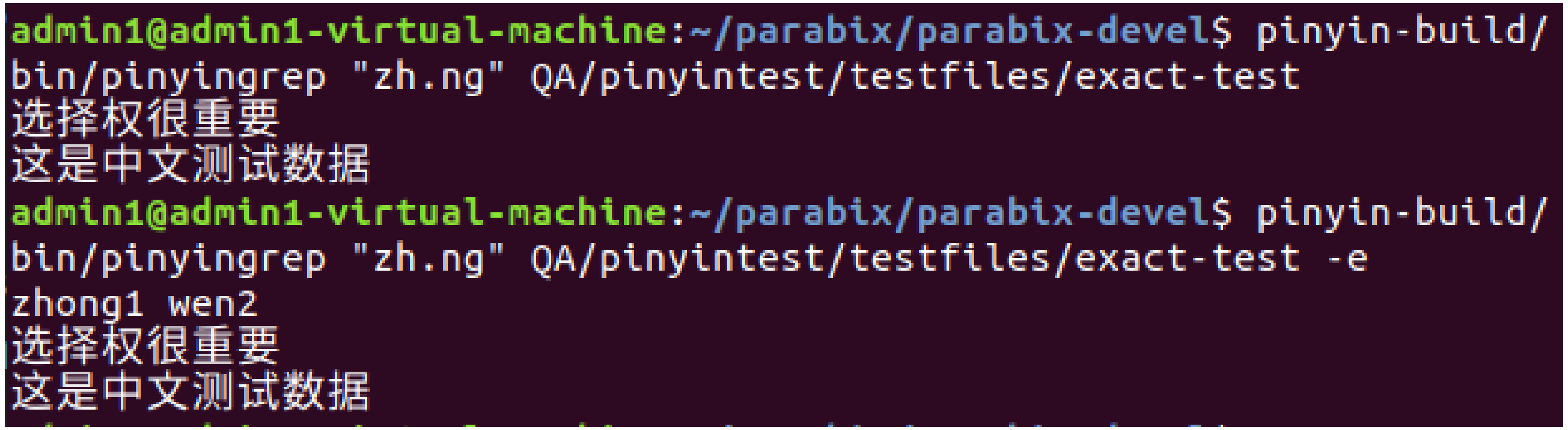
- **File names:**
  - Command Line Flag “-h”



```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-buil
bin/pinyingrep "zhŌng" QA/pinyintest/testfiles/simple_pinyin -h
QA/pinyintest/testfiles/simple_pinyin:欢迎来到中文世界。
QA/pinyintest/testfiles/simple_pinyin:在这里你可以感受中文的博大精
。
QA/pinyintest/testfiles/simple_pinyin:中药主要用于轻症患者。
QA/pinyintest/testfiles/simple_pinyin:所以把筐子中要卖的药都烧了。
```

Functionality: Exact Match Mode

- **Exact Match:**
  - Command Line Flag “-e”
  - Interpret the input directly also as a regular expression to match possible pinyin syllables.



```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "zh.ng" QA/pinyintest/testfiles/exact-test
选择权很重要
这是中文测试数据
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "zh.ng" QA/pinyintest/testfiles/exact-test -e
zhong1 wen2
选择权很重要
这是中文测试数据
```

Functionality: Database Selection

- **Different Database:**
  - XHC1983: From “kXHC1983” fields of Unihan_reading.txt in Unihan database; Collected from 《现代汉语词典》
  - HanyuPinyin: From “kHanyuPinyin” fields and “kMandarin” fields of Unihan_reading .txt in Unihan database; “kHanyuPinyin” is collected from 《漢語大字典》while “kMandarin” is the most customary pinyin reading for this character according to Unihan database description.

- **Selecting Different Database:**
  - -xhc is from XHC1983, which is set as default.
  - -kpy is from HanyuPinyin

Functionality: Database Selection

- **Why Different Database:**
    - “kHanyuPinyin” includes rare readings.
        - E.g. meng4 for 明, which is probably from non-standardized Chinese.
    - “kHanyuPinyin” has few supports for Simplified Chinese.
        - E.g. 药 doesn’t have kHanyuPinyin field, but 葯 has.
    - We have no idea which data sources are actually wanted by our users, so we provide them both.

Functionality: Traditional/Simplified Options

- Data from Unihan_Variants.txt in Unihan database.

- **Rich Command Line Options:**
  - `-all` greps all characters, simplified or traditional. Default.
  - `-trd` greps characters used in traditional Chinese. Including used in both traditional and simplified.
  - `-sim` greps characters used in simplified Chinese. Including used in both traditional and simplified.
  - `-tonly` greps character used *ONLY* in traditional Chinese
  - `-sonly` greps character used *ONLY* in simplified Chinese

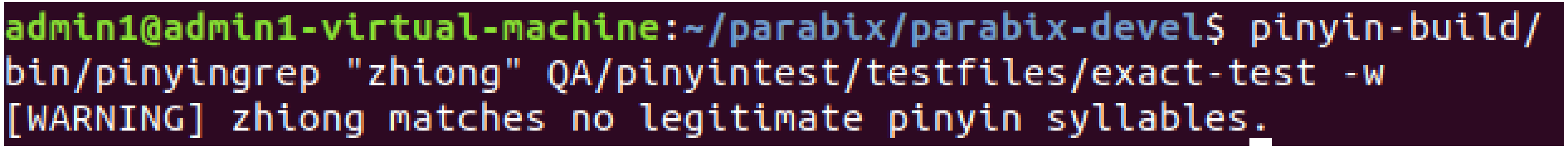
Functionality: Traditional/Simplified Options

- **Why So Many Options?**
  - For example: `井` is used in both traditional and simplified Chinese; `書` is only used in traditional Chinese while `学` is only used in simplified Chinese.
  - For more information: http://www.unicode.org/reports/tr38/#SCTC

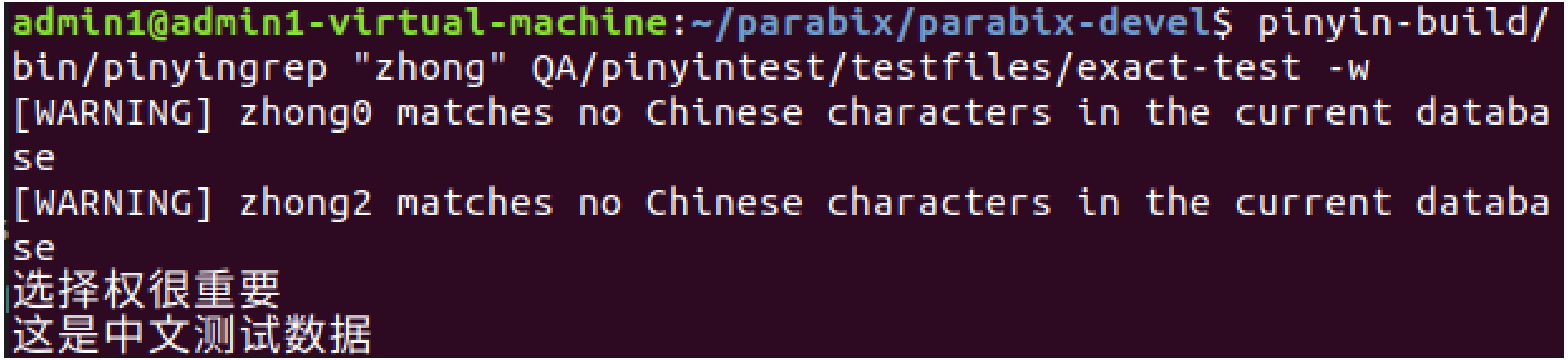
Functionality: Warning Mode

- Command Line flag `-w`
- **Warnings when:**
  - their pinyin syllable input is not legitimate;
  - when the interpreted toned syllable is not in the current database.
- **The basic principle:**
  - to give users more flexibility when using the software without making them bothered by unwanted warning/error messages.

Functionality: Warning Mode



```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "zhiong" QA/pinyintest/testfiles/exact-test -w
[WARNING] zhiong matches no legitimate pinyin syllables.
```



```
admin1@admin1-virtual-machine:~/parabix/parabix-devel$ pinyin-build/
bin/pinyingrep "zhong" QA/pinyintest/testfiles/exact-test -w
[WARNING] zhong0 matches no Chinese characters in the current databa
se
[WARNING] zhong2 matches no Chinese characters in the current databa
se
选择权很重要
这是中文测试数据
```

Highlight

/03

Highlight

- Robust
- Functional
- User-friendly

Procedure

/04

Procedure

- **Tools:**
  - Git – Version Control
  - Wechat – Group Communication
  - Python Scripts – Generating Header Files

- **Work Assignment:**
  - Decoupling
  - e.g. in iteration 2:
    - First, building the framework of “pinyin_interface”
    - Then, Each one implemented certain methods or initialization that were predefined clearly and irrelevant to others’ implementation.

Future work

/05

Future work

- **Many possibilities:**
  - More command line flag like in icgrep.
  - Use more powerful database to do frequency ranking and phrase detection.
  - Input mode without delimiters or with alternative delimiters.
  - More flexible regular expression support.
  - Support for multiple tones syllable like `zhong123`
  - More types of warnings in Warning Mode.

Acknowledgement /06

Acknowledgement

- Many thanks to Professor Cameron for your generous help.
- Appreciation for the source code of “icgrep” and “UCD-scripts” as good references.

Thanks for listening

Yanwen

2020.6.22