

---

# **Exercise 1: Text Tokenization**

# Text Tokenization

---

- ❑ URL: <https://news.ifeng.com/c/8EOcnBCpJir>
- ❑ Crawler: [www.topcoder.com/thrive/articles/web-crawler-in-python](http://www.topcoder.com/thrive/articles/web-crawler-in-python)
- ❑ Chinese word tokenization: <https://github.com/fxsjy/jieba>

# Text Tokenization

## □ Submission

- Code file
- Excel file with tokenized Chinese words:

URL	Title	Content
<a href="https://news.ifeng.com/c/89TNOI">https://news.ifeng.com/c/89TNOI</a>	我/爱/中山大学	中山大学/是/一所/985/高校/...

- Zip the file with filename “2022NLP-exercise 1-学号-姓名.zip/rar”
- Send the result to
  - Email: sysucusers@163.com
  - Subject: 2022NLP-exercise 1-学号-姓名, 例如 “2022NLP-exercise 1-xx-xxx” ;
- Deadline: 2022-3-22, 24:00