Exercise 2: Word2Vec

Recall: Exercise 1 - Text Tokenization

- URL: *https://news.ifeng.com/c/89TNORdIths*
- Crawler: www.topcoder.com/thrive/articles/web-crawler-in-python
- Chinese word tokenization: https://github.com/fxsjy/jieba

Exercise 2: Word2Vec on Tokenization

- ❑ Apply Word2Vec on tokenized sentences from Exercise 1
- ❑ Word2Vec: https://radimrehurek.com/gensim/models/word2vec.html
- ❑ Submissions:
  1) Code file
  2) Word vector file in '.txt' with format:

```
中国：0.2, 10, 20.1, …, 2.36
中国：0.2, 10, 20.1, …, 2.36
中国：0.2, 10, 20.1, …, 2.36
```

  3) Find 10 most similar words with "中国" and list their probabilities in '.txt' file

```
国家：0.84
印度：0.76
世界：0.67
```

Exercise 2: Word2Vec after Tokenization

- Zip the 3 files with filename “2021NLP-exercise 2-学号-姓名.zip/rar”
- Send the assignment to
  - Email: sysucsers@163.com
  - Subject: 2021NLP-exercise 2-学号-姓名，例如“2021NLP-exercise 2-xx-xxx”；
- Deadline: 2021-10-18, 24:00