Project #1

Introduction

This project is about how to build the corpus. I get the information about books using spider, save them and build the corpus.

Description

In first part of this project, I build two spiders. one of them is used to crawl the book's title and its URL of
the book, another is used to crawl the description of the book. The second part is about striping, tokenizing and stemming the book's description we get in first step to build a corpus.

Spider

Getting the title and URL of books,

I need to extract the URL of each book from the website pages 1-10. The URL of page `n` is `https://books.toscrape.com/catalogue/page-n.html`. For example, if we want access the 10th page, just
enter its URL `https://books.toscrape.com/catalogue/page-10.html`.

In each book list page, I noticed that all information about a book we need is in the element named `article` and its class is `product_pod`. By finding all element like this, we can get a list of
all books in this page. In each `article product_pod` element, there is `a` element like `<a href="URL" title="TITLE">TITLE</a>` containing all information we need.

> Note: We should extract the title from `title` attribute, not its content. Because the content will hide the last several letters if the title is too long. For example,
> title="**Foolproof Preserving: A Guide to Small Batch Jams, Jellies, Pickles, Condiments, and More: A Foolproof Guide to Making Small Batch Jams, Jellies, Pickles, Condiments, and More**"
> content in element is **Foolproof Preserving: A Guide ...**

Getting the description of each book

In book detail page, the forth `p` element of a element called `article` with `product_page` class contain the book description, and only one `h1` element is the title of the book. We can get them
and store them.
Note: On my test environment, Windows, according to the rule of naming a file on Windows system, / \ " ' * ; - ? [ ] ( ) ~ ! $ { } &lt > # @ & | space tab newline are not allowed. These might be in the title. We must replace these with other character. In my program, blank space is used to replace all of them.

```
book['title'].replace('<','').replace('>','').replace('\\','').
    replace('/','').replace(':','').replace('*','').replace('?','').replace('<','')
```

Build Corpus

Strip

I process the term `...more`, `i/I'm`, `you/You're`, `he/He/she/She/it/It's`, `tab` and `newline`:

```
    text=text.lower()
    text=re.sub(r'\.\.\.more$','more',text)
    text = re.sub(r'n\'t', ' not', text)
    text = re.sub(r'\'am', ' am', text)
    text = re.sub(r'\'re', ' are', text)
    text = re.sub(r'\'s', ' is', text)
```

Tokenize

Add a space before the punctuation, then split the string into a list by spaces.

```
text = re.sub(r'\.', ' .', text)
text = re.sub(r',', ' ,', text)
text = re.sub(r'\?', ' ?', text)
text = re.sub(r'\'', ' \'', text)
text = re.sub(r'"', ' \"', text)
text = re.sub(r';', ' ;', text)
text = re.sub(r':', ' :', text)
text=text.split(' ')
```

Stem

I use `SnowballStemmer` API in `NLTK` package directly

```
stemmer=nltk.stem.snowball.SnowballStemmer('english',ignore_stopwords=True)
word=stemmer.stem(word)
```

Conclusions

I get the information about books using spider, save them and build the corpus.