For my first attempt, I documented how I built a simple web crawler that grabbed a set of documents from the Internet and summarized them using gensim.

Step

Step 1: Determine which Zhihu page to crawl

I use the requests library to get the page HTML. Here, I will be available from https://www.zhihu.com/search?type=content&q=%E6%9C%89%E6%9C%BA%E9%A3%9F%E7%89%A9 search results page of "organic food”.

Step 2: Parse the page

I use the beautifulsoup4 library to parse HTML so that I can easily fetch the text in the page.

Step 3: Grab the text

I used the beautifulsoup4 library to extract all the text from the page and store it in a string.

Step 4: Preprocess the text

I use the gensim.utils.simple_preprocess function to convert the text to lower case and split it into words. (I may need to remove stops and punctuation.)

Step 5: Summarize using gensim

Using gensim. Summarization. Summarize function to summarize the text.

Perception and summary

I designed a sample code using Python to crawl the relevant content of the keyword "organic food" in Zhihu search results and perform text summary.

This code uses the requests library to send HTTP requests to Zhihu to retrieve the HTML content of the search results page. The HTML is then parsed into a BeautifulSoup object using the BeautifulSoup library, from which the text in all <p> tags is extracted and stored in a string variable. Next, use the simple_preprocess function in the Gensim library to convert the text to lower case and perform simple processing (such as removing punctuation and stopping words). Finally, use the summarize function in the Gensim library to generate a text summary from the processed text and store it in a string variable.

Finally, the code prints a text summary to the console. Among them, the summary length is 10% of the length of the original text, which is achieved by passing the parameter ratio=0.1 to the summarize function.