Bias on Wikipedia?

The project I conducted searched Wikipedia sites on seven different topics related to genetics, which I deem a controversial topic, and evaluates the wikipedia articles' implicit bias. Essentially, I used a "histogram" type system of two for loops to count the number of words in the Wikipedia articles matching with negative or positive words in my lists. This more wordy form of sentiment analysis allowed me meet my goal of better understanding how for loops work, and observing interesting biases (which shouldn't technically exist in Wikipedia articles) between disciplines relating to genetic fields.

The first part of my code translates wikipedia articles into text files. This part of the code searches for articles on wikipedia and then prints them as text files. Pickle and pattern.web need to be imported for this to work. A for loop was used to search wikipedia for them using w.search of that term. Included in the for loop is something that takes prints article sections and puts them into a text file with the name of the element and encoded. The text file output is then ready to be processed by the function.

The function I wrote to find certain terms in the wikipedias uses the same search term list as well as lists containing positive and negative words to find the number of positive or negative words in each article so they are available for comparison. The function does this by splitting up the documents into words that can be processed individually through for loops in such a way that the can be individually processed by one for loop that sees if they are a negative word and counts if they are and one for loop that sees if they're a positive word can counts if they are. I chose to use this system of comparing sentiment rather than the simpler one because I wanted to learn from writing effective for loops, which is something I sometimes struggle with.

Interesting things I found from the data is that Wikipedia seems to be implicitly positive about subjects regarding genes, but that it is more obviously so on topics like conservation genetics versus topics like cloning, which I expected. One thing I didn't expect is for so many negative words to be associated with gene therapy, but this might come from fear of "designer babies" or from the risks associated with the process. Genetic engineering in general surprised me as well, as it seemed far more positive than negative, but this is a very nuanced subject, so it could be that Wikipedia simply represented more of the subjects people feel favorably about within this category.

Here is a table representing the results:

Name of Subject	# of Positive Words	# of Negative Words	Ratio
Cloning	61	30	~2:1
Genetic Engineering	100	31	~3:1
Gene Therapy	66	43	~3:2
CRISPR	75	26	~3:1
Genetically Modified Organisms	41	33	~3:2

Genetics	101	39	~2:1
Conservation Genetics	32	4	~8:1

The results in this table would imply that CRISPR is viewed more positively than cloning, and that genetic engineering is viewed more favorably than genetics. This is very interesting, but my suspicion is that it has something to do with CRISPR editing its own page and a coincidence for the latter. Either way, I find the results of running this program to be interesting.

I think my execution of the code went well: I liked that I could find an interesting way to evaluate sentiment that seemed to work to some degree. I could improve by choosing websites that I know present a bias so that I can be sure that my code is working to evaluate what it should rather than pointing out coincidences. I think my unit tests were alright, but I wished they could have been more fleshed out. What I've learned for going forward is that at this point, I should have a program whose results I can entirely predict or expect in order to further my learning. I wish I had known this before I began this process.