ttnt59

# Problem 1

I used BeautifulSoup to scrape each page of search results. I determined relevance using firstly only /news/ articles, but including short article video articles /av/. I then further determined relevance to a given keyword by a search result containing that search result in the title or in the description. This is why there are no results for the more complex keywords such as “targeted threat”, as the articles that appeared would just end up matching to “threat”, and there would be lots of articles about petty (relatively) crimes.

# Problem 2

I used BeautifulSoup again and simply reduced each article to text content only, but including embedded tweet contents, as I found this would often include fairly meaningful data.

# Problem 3

The algorithm looks for instances of other keywords in an article relevant for a given keyword, and from a starting distance of one, divides by two if the word is present. Furthermore, it divides by 1.25 if the keyword is present in the Wikipedia definition for the given keyword, as this is less relevant data.

However, this still leaves gaps for the more complex terms. There is no data for “targeted threat”, “Advanced Persistent Threat”, “DoS attack” and “malicious bot”, as the BBC has no relevant articles for them. The algorithm’s results would be improved with more data, but the keywords provided

# 