1. Introduction

To connect the frontend and backend, I use Django to develop my website. Typing “python manage.py runserver”, the server will be started. The program begins from scraping articles from TNW and Bloomberg. Thereafter, the website is built and we can type “127.0.0.1:8000/index/” in the address bar of Chrome to open it. The website allows users to key in a keyword and return 10 most related articles scraped from TNW and Bloomberg respectively. For more details of the website, please read the README on Github.

1. Technical skills
   1. Web scraping

I use the packages “requests” to get the web page source code and “Beautiful soup” to analyze it. I scrape for articles from 20 pages in the “tech section” from TNW (<https://thenextweb.com/section/tech/>) and the “artificial intelligence”, “big data”, “cloud computing”, “mobile” sections from Bloomberg. Approximately 300 articles are crawled and prepared as a dataset.

* 1. Analysis

I apply the “word2vec” model from the package “genism” to analyze the title of each articles. For data preprocessing, I remove the stopwords such as “we”, “you”, ”are”. Then I use “word2vec” model to embed the remain words into vectors with dimension 150. The vector for each title can be represented by averaging the vectors of the words which are contained in the title. When receiving a keyword from the user, the cosine similarities is calculated between the vectors of the keyword and the title. The articles with title that have the top 10 highest similarities will be returned to the user.

* 1. Web development

I use CSS and HTML to design the appearance of the website and javascript to implement the slide show at the homepage.

1. Difficulties

Bloomberg will block the crawler if just using “request.get(url)”. The solution is to add cookie in “request.get(url)”.

1. Improvements:

There are two points that require improvement. First, the word2vec model is not good enough so that the searching results are similar respect to many different keywords. Second, I should have set up a button that can switch over the results from another website. For example, when users are reviewing the results from TNW, there should be a button that can switch to the results from Bloomberg.