Nanos Machine Learning Task Approach

Web Scraping

I used two approaches on Web Scraping. One is by downloading the website's html code directly using python and the next approach is by using a webdriver for Javascript intensive websites. The latter approach also works better for websites that have anti-bot mechanisms and certificate checkers.

Text preprocessing

I used regex to clean the text by removing digits and punctuation and lemmatized the text to avoid word duplicates and provide a uniform starting point.

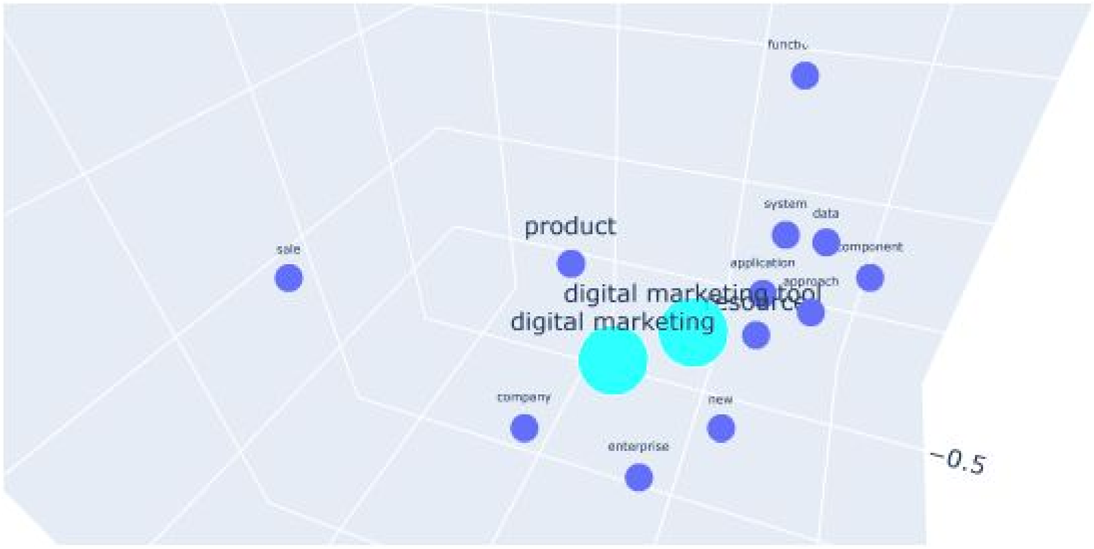
Text matching

I employed pre-trained word vectors which i.e 1 million word vectors trained on Wikipedia 2017, UMBC web based corpus and statmt.org news dataset (16B tokens). I loaded the model and calculated the word vectors for the search words. I did this by splitting the word vectors e.g **digital marketing tool** to digital, marketing and too. I then calculated the vectors for the individual words and finally the average to the vector representation of the whole.

I collected the vector representation for each word from the website and calculated the cosine distances from that word to the search phrase vector. The shorter the distance the more similar the word is to the phrase.

Plotting

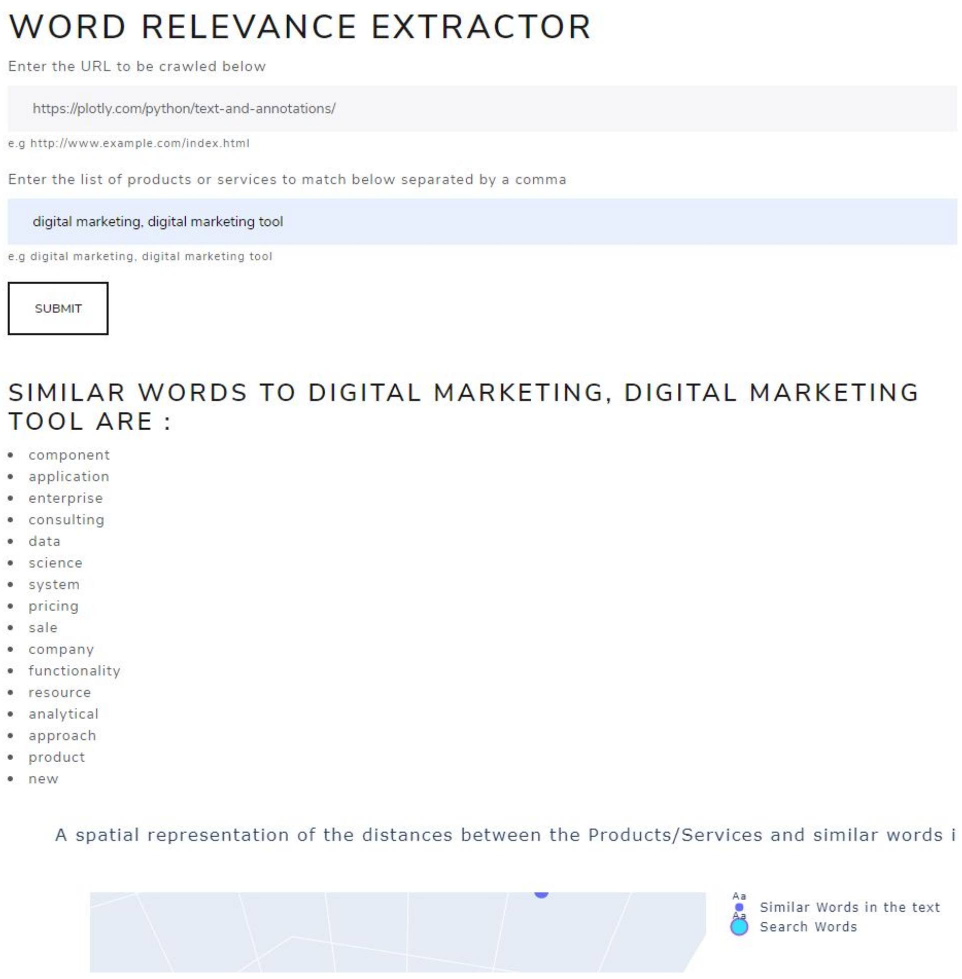
I reduced the vector representations from 300 dimensions to 3 dimensions using principal component analysis and plotted the words in a 3d plot for a better understanding of how the words relate to each other and the search phrases in a 3d space.



Deployment

The app is deployed in a flask-like web-server which reduces the load time of the model since it's only loaded once when the app is run

Here is what you should see



WORD RELEVANCE EXTRACTOR

Enter the URL to be crawled below

https://plotly.com/python/text-and-annotations/

e.g http://www.example.com/index.html

Enter the list of products or services to match below separated by a comma

digital marketing, digital marketing tool

e.g digital marketing, digital marketing tool

SUBMIT

SIMILAR WORDS TO DIGITAL MARKETING, DIGITAL MARKETING TOOL ARE :

- component
- application
- enterprise
- consulting
- data
- science
- system
- pricing
- sale
- company
- functionality
- resource
- analytical
- approach
- product
- new

A spatial representation of the distances between the Products/Services and similar words i