Capstone Proposal

Vandana Iyer

Machine Learning Engineer Nanodegree

April 3rd, 2018

Proposal

Automated travel blogging with deep learning

Domain Background

The primary area of focus is Natural Language Processing[NLP] and Deep learning using Recurrent Neural Networks[RNNs]. During the course of nanodegree, I have worked on assignments related to supervised, unsupervised, reinforcement learning and Convolutional Neural Networks. I want to take this learning further by developing a deep learning model that uses RNNs and NLP to achieve automated travel blogging.

This project is based on the paper [Character-level Recurrent Neural Networks in Practice: Comparing Training and Sampling Schemes]

Problem Statement

Many travellers find very little time to blog on-the-go. For some travellers, blogging is one of the means to earn money to fund their future travel. Is it possible to automate travel blogging with artificial intelligence? Wouldn't it be great if your AI assistant helped you keep memories of your travels by automatically writing down stories about the places you visited? It turns out that this task could be tackled using deep learning because there is plenty of data available online.

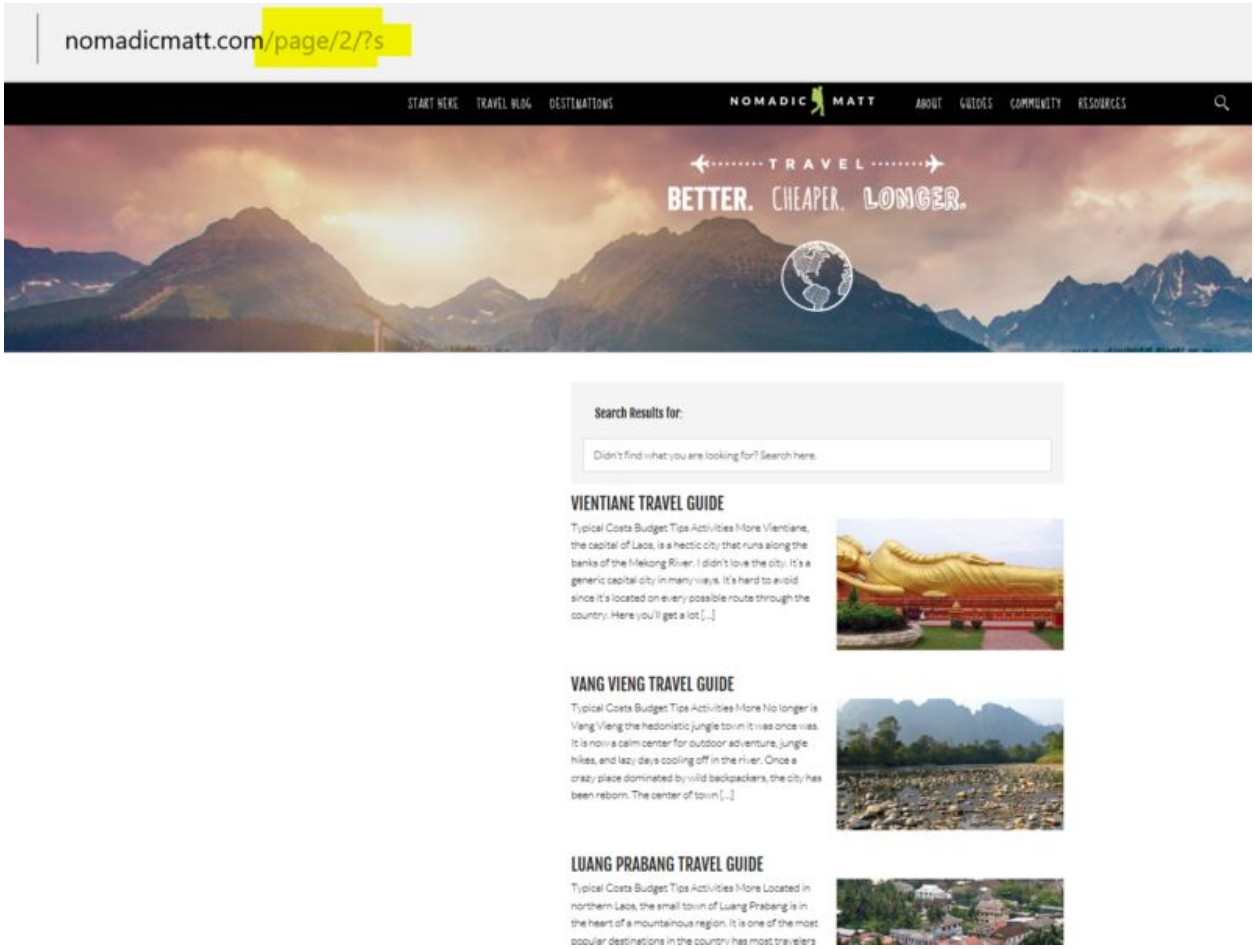
I travel a lot and I find blogging a time consuming process. A tool like this will definitely simplify and speed up my blogging process.

Datasets and Inputs

Most of these travel blogs have been written with Wordpress and they share the same HTML structure. For example, take a look at Nomadic Matt's blog: if you perform an empty search you will get the list of all the blog articles available:

It will be enough to go through all the pages and all the links in the pages with a crawler to get all the articles written. Moreover, once you click on an article all the content can be found in the CSS container labelled with "entry-content". Note that this is true for any blog written with Wordpress:

[

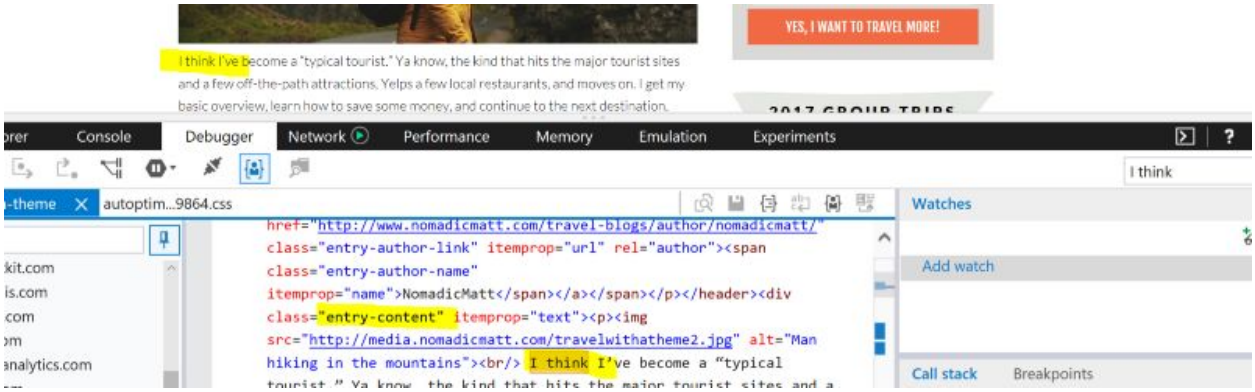


]

It will be enough to go through all the pages and all the links in the pages with the crawler to get all the articles written. Moreover, once you click on an article all the

content can be found in the CSS container labelled with “entry-content”. Note that this is true for any blog written with Wordpress:

[



]

I have found a crawler - Scrapy on GitHub, which crawls the blogs of top 50 bloggers for a period of 4 months and also takes care of removing strange symbols from the retrieved text.

Solution Statement

Once enough data is collected, it is possible to train a character-level Recurrent Neural Network (RNN). I will be using a keras-based implementation of Character-Aware Neural Language Models to train and predict the next character after a sequence of characters. How to Develop a Character-Based Neural Language Model in Keras. A stateful LSTM will be used to achieve this. I shall at a later stage use Word-Level Neural Language Model if the Character-based RNN does not give meaningful sentences.

Benchmark Model

A benchmark model in this case would be a model which auto completes with meaningful sentences for a blog.

Evaluation Metrics

Asking my network to write a travel blog for me from scratch might be a bit too much. On the other hand, it might be possible to get some help from the network to write my blog. In this scenario, I could imagine myself writing a sentence and having the network

to complete it adding further details. For example, I can imagine myself writing: “I am travelling to ”. And the possible outputs could be:

```
“I am travelling to “
```

```
I am travelling to Europe (finding the companions while I was there) on me. You buy reading a night’s
I am travelling to Cornwall: put I follow you to do get the Olymbookality Starbucks pretty packed alo
I am travelling to move on the god dorms time you wan me get this mysterio. Everyone over sizes that
```

Or I could write something like “The hostel was ” and the network would say:

```
“The hostel was “
```

```
The hostel was cool. So day in the world, move cards, and come back to why time while I look at the m
The hostel was on the Anded tourist attractions through Europe. 3. The street to the Musen Gudding is
The hostel was probably not even darked to Travel the Bath. I was celebrated so often, a high day
```

We can see that results are not perfect but not too bad either. It is cool to see that these sentences are not in the original data set and thus the network is really generating new ones.

Project Design

First step will be data collection. I will use Scrapy to crawl the blogs of top travellers. Once I have sufficient data, I will feed this data to a character-based RNN and check if it is suggesting meaningful sentences.

If the character-based RNN's suggestions are not upto the mark, I will then use a word-based RNN to achieve the same. If all goes well, I will take it to the next level and use this data to come up with an automated itinerary plan for those who intend to travel.

References

- Character-level Recurrent Neural Networks in Practice: Comparing Training and Sampling Schemes
- Travel Blogs with Deep Learning
- Crawling Blogs
- Character level deep learning
- The Unreasonable Effectiveness of Recurrent Neural Networks
- How to Develop a Character-Based Neural Language Model in Keras
- Develop a Word-Level Neural Language Model and Use it to Generate Text
- Character-Aware Neural Language Models. A Keras-based implementation