MSc Data Science - Interim Report

Peter Moore

2018-09-12

Project Details

Title: Building a holiday recommendation engine based on Twitter timelines

Course: AC53011

Studentid: '170000983'

Project Summary

The aim of this project is to recommend exotic holiday destinations for people. "Exotic" in this case meaning places they may never have thought of going but that would suit them. These would be recommended via an analysis of their own holiday snaps per their Twitter feed and a comparison with their peers.

Project Specifications

The recommendation engine will follow the following steps and considerations.

Data understanding

Online part: reviewing "obvious" hashtags such as #vacation to identify holiday snaps; comparing and contrasting these with, for example, #work. Using this to build a library of terms to create training set. **Data part**: regency, frequency, intensity analysis (RFI) of tweets versus holiday (do Tweets cluster around holidays); how does place name vary. What are the pitfalls, for example someone is away on business?

Data preparation

Technology selection: in the first instances, tweets will be downloaded using the Python[1] Tweepy[2] package, data will be initially stored securely. Image detection will be performed using Tensorflow[3]. Initial process will be on a MacBook using host CPU, with (probable) expansion to Google's Cloud[4] computing platform using NVIDIA GPU[5]. Tweet collection: development of download script. Tweet image pseudonymisation: co-opting of technology to remove personally identifying information (PII) Tweet text pseudonymisation: removal of names and references.

Modelling

The fundamental unit of data collection will be the JSON from a tweet[6]. This includes information such as photographs, Hashtag and emoji. From these, the following models will need be created:

[1] https://www.python.org
[2] http://www.tweepy.org
[3] https://www.tensorflow.org
[4] https://cloud.google.com/gpu/
[5] https://www.nvidia.com/en-us/data-center/tesla-k80/
[6] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html

Twitter downloader

Purpose: to be able to get the dataset for the project. **Considerations**: inline processing, pseudonymisation.

Object detection

Purpose: For pseudonymisation of faces; facial sentiment (e.g. pouting); for detection of holiday effects (e.g. pints of beer) **Preliminary considerations**: sunsets, pints of beer, planes, smiles, oceans

Sentiment analysis

Purpose: The purpose is to recommend holidays that make people happy. Sentiment analysis will be used to investigate what makes people happy **Preliminary considerations**: Hashtag, Emoji, Text,Place, RFI

Recommendation engine

Building

Purpose: To assess the holiday effects, compare them to those of others and create recommendation via, for example, collaborative filtering methods **Preliminary considerations**: Flora, fauna, background, activities, timing

The recommendation itself

Purpose: To allow the user to interact with the product, for example, using R-shiny to create a navigable front-end, that allow, not only the real-time parsing of their Twitter feed but also the parametrisation of their suggestion via dimensions such as adventurousness and relaxation.

Evaluation

Iteration of above process to assess emergent problems. Also, the area where known problems are assessed, for example it is illegal under GDPR article to parse sexuality or religious belief (say) and these are known constraints.

Deployment

Liaison with business to see how the app would be useful for business. Production of project poster. Curation of stand for project demonstration day in Dundee

Dependencies

Technology: CPU processing is available to the project but GPU processing on NVIDIA is not. This makes NVIDIA a dependency, and/or, Google Cloud (who offer a rental version of this). Twitter is vital to the project and as such is a primary dependency.

Literature Review

The successful implementation of this project requires the use of object recognition as described by LeCun, Bengio, and Hinton (2015) and that this be rapid (Krizhevsky, Sutskever, and Hinton (2012)). The photographic analysis will rest on a variety of pre-trained models, in particular, those in the the ImageNet dataset (Deng et al. (2009)).

A stretch goal of the project is to move beyond traditional object detection and into semantic scene classification, for example to detect a sunset. This has been looked at by Shen et al. (2003) but not form a deep learning perspective. A more recent edition Of his body of work comes from Chen et al. (2018) who look at cityscapes with Tensorflow.

Before a holiday is recommended it is necessary to ascertain whether or not a person is on holiday. For example, does a photo containing a pint of beer imply a holiday; ditto a sunset. The interaction between photographs and tourism has been described by song2006tourism.From the metadata, the place of a tweet is sometimes known and this may imply away from home; modal location may be used to describe home but when does away from home constitute holiday? This is an interesting question that has been tangentially looked at by Toyama, Logan, and Roseway (2003).

Finally, in order to recommend holidays to people, it is necessary to know what makes them happy. Whilst this can be achieved by sentiment analysis, some traditional methods for doing this have been found wanting due to sarcasm (see González-Ibánez, Muresan, and Wacholder (2011)). A potential way of avoiding this is via the use of emoji as studied by Taboada et al. (2011) who show that emojis may be used to detect sarcasm.

These documents form the starting point for the problem as understood.

Current progress

An initial twitter downloader has been built using Tweepy with in Python. The JSON was interrogated. This acts as a successful proof of concept.

The project is presently looking into the ImageNet database as an (at least partially) solved object detection problem.

Ethics forms have been submitted and have had preliminary approval from the project supervisor and are awaiting ethics committee approval.

Establishment of a happiness/relaxation index and the recommendation engine itself are yet to be commenced.

Future Project Timeline

September: gathering and pseudonymisation of photographic data from Twitter, labelling of images, training of algorithm. October: data mining of photos, for example identification of holidays, clustering of data to produce recommendations (for example via association rules) November: definition and creation of product, for example, development of API December: testing and showcasing January: write-up and presentation

References

Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *arXiv Preprint arXiv:1802.02611*. https://arxiv.org/pdf/1802.02611.pdf.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–55. Ieee. https://www.researchgate.net/profile/Li_Jia_Li/publication/

221361415_ImageNet_a_Large-Scale_Hierarchical_Image_Database/links/00b495388120dbc339000000/ImageNet-a-Large-Scale-Hierarchical-Image-Database.pdf.

González-Ibánez, Roberto, Smaranda Muresan, and Nina Wacholder. 2011. "Identifying Sarcasm in Twitter: A Closer Look." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, 581–86. Association for Computational Linguistics.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553). Nature Publishing Group: 436. https://www.evl.uic.edu/creativecoding/courses/cs523/slides/week3/DeepLearning_LeCun.pdf.

Shen, Xipeng, Matthew Boutell, Jiebo Luo, and Christopher Brown. 2003. "Multilabel Machine Learning and Its Application to Semantic Scene Classification." In *Storage and Retrieval Methods and Applications for Multimedia 2004*, 5307:188–200. International Society for Optics; Photonics. https://s3.amazonaws.com/academia.edu.documents/46558675/Multilabel_machine_learning_and_its_appl20160616-29151-1ftbw2m.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1536131162&Signature=xh0U/%2BzssJdEF3whyIWmi0qZP70s/%3D&response-content-disposition=inline/%3B/%20filename/%3Dtitle_Multilabel_machine_learning_and_i.pdf.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37 (2). MIT Press: 267–307. https://www.mitpressjournals.org/doi/pdfplus/10.1162/COLI_a_00049.

Toyama, Kentaro, Ron Logan, and Asta Roseway. 2003. "Geographic Location Tags on Digital Images." In *Proceedings of the Eleventh Acm International Conference on Multimedia*, 156–66. ACM.