**Data Science Research Project – Interim Report**

**Peter Moore**

**Student ID: 170000983**

**Course: AC53011**

**31st August 2018**

## Project Title

Building a holiday recommendation engine based on Twitter timelines

## 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]](#endnote-1) Tweepy[[2]](#endnote-2) package, data will be initially stored securely on Apple iCloud[[3]](#endnote-3). Image detection will be performed using Tensorflow[[4]](#endnote-4). Initial process will be on a MacBook using host CPU, with (probable) expansion to Google’s Cloud[[5]](#endnote-5) computing platform using NVIDIA GPU[[6]](#endnote-6).

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 following models will need be created:

|  |  |  |
| --- | --- | --- |
| Object detection | For pseudonymisation of faces; for detection of holiday effects sunsets, pints of beer, planes, smiles, oceans |  |
| Sentiment analysis | The purpose is to recommend holidays that make people happy. Sentiment analysis will be used to investigate what makes people happy | Hashtag, Emoji, Text,Place, RFI |
| Building the recommendation engine | To assess the holiday effects, compare them to those of others and create a data-informed recommendation | Flora, fauna, background, activities, timing |
| Making the recommendation | Considerations of user interface | Repeat visits, holiday length, time of year |

### 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

Social: twitter holiday studies; emoji sarcasm; holiday snap studies

Technical: object detection for animals, beverages, backdrops; image pixelation and/or removal;

The Interim Report is usually accompanied by a full literature review, which will form the basis of a dissertation chapter. Like all literature reviews, it should provide an overview of the theories and practices that are most relevant to your own work. It is often organised thematically and will demonstrate the student’s ability to contextualise their project within the recent advances in their field. This will probably be the longest section of the Interim Report.

## Current progress

This section of the Dissertation Interim Report will outline what you have achieved so far. It may include mentions of the background research you’ve undertaken, skills training you’ve received, and any practical work you’ve done toward completing your project.

## Obstacles to Progress

In this part of the Interim Report students should note down any difficulties they’ve encountered so far. It should also contain some details of how the student plans to confront these challenges, whether through changes to the research plan or minor adjustments to the overall project. The University is particularly interested in this section of the Interim Report, as they want to ensure the student’s successful completion of the degree.

## Planned Methodology

For most degree programmes students will also need to outline their research strategy for the remainder of the dissertation work. This may focus primarily on text-based research or it may require practical lab work. Regardless, the student should justify their chosen methodology and explain how it will adequately address the research question.

## Future Project Timeline

This section details the student’s plan of action for the remainder of the degree, and is usually broken down into a month-by-month timeline. You should include all the work that is relevant to completing the dissertation, including research, project development, chapter writing and lab activities. You might also include plans for additional skill training, funding bids, and conference presentations.

## Bibliography

@Manual{R-base,

title = {R: A Language and Environment for Statistical Computing},

author = {{R Core Team}},

organization = {R Foundation for Statistical Computing},

address = {Vienna, Austria},

year = {2018},

url = {https://www.R-project.org/},

}

@Manual{R-rmarkdown,

title = {rmarkdown: Dynamic Documents for R},

author = {JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang},

year = {2018},

note = {R package version 1.9},

url = {https://CRAN.R-project.org/package=rmarkdown},

}

@inproceedings{krizhevsky2012imagenet,

title = {Imagenet classification with deep convolutional neural networks},

author = {Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E},

booktitle = {Advances in neural information processing systems},

pages = {1097--1105},

year = {2012},

note = {Hinton et al image identification classification - how they did it},

url = {http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf},

Read = True,

}

@article{lecun2015deep,

title = {Deep learning},

author = {LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey},

journal = {nature},

volume = {521},

number = {7553},

pages = {436},

year = {2015},

publisher = {Nature Publishing Group},

note = {Hinton et al paper in Nature},

url = {https://www.evl.uic.edu/creativecoding/courses/cs523/slides/week3/DeepLearning\_LeCun.pdf},

Read = True,

}

@inproceedings{deng2009imagenet,

title = {Imagenet: A large-scale hierarchical image database},

author = {Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li},

booktitle = {Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on},

pages = {248--255},

year = {2009},

organization = {Ieee},

note = {imagenet database "definition"},

url = {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},

Read = False,

}

@inproceedings{shen2003multilabel,

title = {Multilabel machine learning and its application to semantic scene classification},

author = {Shen, Xipeng and Boutell, Matthew and Luo, Jiebo and Brown, Christopher},

booktitle = {Storage and Retrieval Methods and Applications for Multimedia 2004},

volume = {5307},

pages = {188--200},

year = {2003},

organization = {International Society for Optics and Photonics},

note = {a way of finding backgrounds to photos},

url = {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},

Read = False,

}

@article{chen2018encoder,

title = {Encoder-decoder with atrous separable convolution for semantic image segmentation},

author = {Chen, Liang-Chieh and Zhu, Yukun and Papandreou, George and Schroff, Florian and Adam, Hartwig},

journal = {arXiv preprint arXiv:1802.02611},

year = {2018},

note = {a way of looking at cityscapes with tensorflow},

url = {https://arxiv.org/pdf/1802.02611.pdf},

Read = False,

}

@article{song2006tourism,

title = {Tourism demand forecasting},

author = {Song, Haiyan and Turner, Lindsay and others},

journal = {International handbook on the economics of tourism},

pages = {89--114},

year = {2006},

publisher = {Edward Elgar Cheltenham},

note = {Holiday photos vis-a-vis tourism demand},

url = {https://www.researchgate.net/file.PostFileLoader.html?id=56a58f2e7c192070038b4576&assetKey=AS\%3A321616137916417\%401453690670212},

Read = False,

}

@article{taboada2011lexicon,

title = {Lexicon-based methods for sentiment analysis},

author = {Taboada, Maite and Brooke, Julian and Tofiloski, Milan and Voll, Kimberly and Stede, Manfred},

journal = {Computational linguistics},

volume = {37},

number = {2},

pages = {267--307},

year = {2011},

publisher = {MIT Press},

note = {sentiment analysis of tweets covering emoji use},

url = {https://www.mitpressjournals.org/doi/pdfplus/10.1162/COLI\_a\_00049},

Read = False,

}

@inproceedings{he2016identity,

title = {Identity mappings in deep residual networks},

author = {He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian},

booktitle = {European conference on computer vision},

pages = {630--645},

year = {2016},

organization = {Springer},

note = {First of references in the Tensorflow ResNet, read me, referred to therein as v1},

url = {https://arxiv.org/pdf/1512.03385.pdf},

Read = False,

}

@inproceedings{he2016deep,

title = {Deep residual learning for image recognition},

author = {He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian},

booktitle = {Proceedings of the IEEE conference on computer vision and pattern recognition},

pages = {770--778},

year = {2016},

note = {second of references in the Tensorflow ResNet, read me, referred to therein as v2},

url = {https://arxiv.org/pdf/1512.03385.pdf},

Read = False,

}

@Manual{R-base,

title = {R: A Language and Environment for Statistical Computing},

author = {{R Core Team}},

organization = {R Foundation for Statistical Computing},

address = {Vienna, Austria},

year = {2018},

url = {https://www.R-project.org/},

}

@Manual{R-rmarkdown,

title = {rmarkdown: Dynamic Documents for R},

author = {JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang},

year = {2018},

note = {R package version 1.9},

url = {https://CRAN.R-project.org/package=rmarkdown},

}

@inproceedings{krizhevsky2012imagenet,

title = {Imagenet classification with deep convolutional neural networks},

author = {Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E},

booktitle = {Advances in neural information processing systems},

pages = {1097--1105},

year = {2012},

note = {Hinton et al image identification classification - how they did it},

url = {http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf},

}

@article{lecun2015deep,

title = {Deep learning},

author = {LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey},

journal = {nature},

volume = {521},

number = {7553},

pages = {436},

year = {2015},

publisher = {Nature Publishing Group},

note = {Hinton et al paper in Nature},

url = {https://www.evl.uic.edu/creativecoding/courses/cs523/slides/week3/DeepLearning\_LeCun.pdf},

}

@inproceedings{deng2009imagenet,

title = {Imagenet: A large-scale hierarchical image database},

author = {Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li},

booktitle = {Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on},

pages = {248--255},

year = {2009},

organization = {Ieee},

note = {imagenet database "definition"},

url = {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},

}

@inproceedings{shen2003multilabel,

title = {Multilabel machine learning and its application to semantic scene classification},

author = {Shen, Xipeng and Boutell, Matthew and Luo, Jiebo and Brown, Christopher},

booktitle = {Storage and Retrieval Methods and Applications for Multimedia 2004},

volume = {5307},

pages = {188--200},

year = {2003},

organization = {International Society for Optics and Photonics},

note = {a way of finding backgrounds to photos},

url = {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},

}

@article{chen2018encoder,

title = {Encoder-decoder with atrous separable convolution for semantic image segmentation},

author = {Chen, Liang-Chieh and Zhu, Yukun and Papandreou, George and Schroff, Florian and Adam, Hartwig},

journal = {arXiv preprint arXiv:1802.02611},

year = {2018},

note = {a way of looking at cityscapes with tensorflow},

url = {https://arxiv.org/pdf/1802.02611.pdf},

}

@article{song2006tourism,

title = {Tourism demand forecasting},

author = {Song, Haiyan and Turner, Lindsay and others},

journal = {International handbook on the economics of tourism},

pages = {89--114},

year = {2006},

publisher = {Edward Elgar Cheltenham},

note = {Holiday photos vis-a-vis tourism demand},

url = {https://www.researchgate.net/file.PostFileLoader.html?id=56a58f2e7c192070038b4576&assetKey=AS\%3A321616137916417\%401453690670212},

}

@article{taboada2011lexicon,

title = {Lexicon-based methods for sentiment analysis},

author = {Taboada, Maite and Brooke, Julian and Tofiloski, Milan and Voll, Kimberly and Stede, Manfred},

journal = {Computational linguistics},

volume = {37},

number = {2},

pages = {267--307},

year = {2011},

publisher = {MIT Press},

note = {sentiment analysis of tweets covering emoji use},

url = {https://www.mitpressjournals.org/doi/pdfplus/10.1162/COLI\_a\_00049},

}

@inproceedings{he2016identity,

title = {Identity mappings in deep residual networks},

author = {He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian},

booktitle = {European conference on computer vision},

pages = {630--645},

year = {2016},

organization = {Springer},

note = {First of references in the Tensorflow ResNet, read me, referred to therein as v1},

url = {https://arxiv.org/pdf/1512.03385.pdf},

}

@inproceedings{he2016deep,

title = {Deep residual learning for image recognition},

author = {He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian},

booktitle = {Proceedings of the IEEE conference on computer vision and pattern recognition},

pages = {770--778},

year = {2016},

note = {second of references in the Tensorflow ResNet, read me, referred to therein as v2},

url = {https://arxiv.org/pdf/1512.03385.pdf},

}

@inproceedings{toyama2003geographic,

title={Geographic location tags on digital images},

author={Toyama, Kentaro and Logan, Ron and Roseway, Asta},

booktitle={Proceedings of the eleventh ACM international conference on Multimedia},

pages={156--166},

year={2003},

organization={ACM}

}

1. <https://www.python.org> [↑](#endnote-ref-1)
2. <http://www.tweepy.org> [↑](#endnote-ref-2)
3. <https://www.apple.com/uk/icloud/> [↑](#endnote-ref-3)
4. <https://www.tensorflow.org> [↑](#endnote-ref-4)
5. <https://cloud.google.com/gpu/> [↑](#endnote-ref-5)
6. <https://www.nvidia.com/en-us/data-center/tesla-k80/> [↑](#endnote-ref-6)