School of Computing and Information Systems
The University of Melbourne
COMP90049, Introduction to Machine Learning, Semester 1 2021

Assignment 3: Classifying the Geolocation of Tweets

Released: Wednesday, April 21st 2021.

Due: **Stage I**: Friday, May 14 5pm
Stage II: Wednesday, May 19 5pm

Marks: The Project will be marked out of 30, and will contribute 30% of your total mark.

1 Overview

In this assignment, you will develop and critically analyse **geolocation classifiers for Tweets**. That is, given a tweet, your model(s) will produce a prediction of where the author of the tweet was located. You will be provided with a data set of tweets that have been annotated with the location of the author. The assessment provides you with an opportunity to reflect on concepts in machine learning in the context of an open-ended research problem, and to strengthen your skills in data analysis and problem solving.

The goal of this assignment is to **critically assess** the effectiveness of various Machine Learning classification algorithms on the problem of determining a tweeter's location, and to **express the knowledge that you have gained in a technical report**. The technical side of this project will involve applying appropriate machine learning algorithms to the data to solve the task. There will be a Kaggle in-class competition where you can compare the performance of your algorithms against your classmates.

The focus of the project will be the report, formatted as a short research paper. In the report, you will demonstrate the knowledge that you have gained, in a manner that is accessible to a reasonably informed reader.

2 Deliverables

Stage I: Model development and testing and report writing (by May 14):

1. One or more programs, written in Python, including all the code necessary to reproduce the results in your report (including model implementation, label prediction, and evaluation). You should also include a README file that briefly details your implementation. *Submitted through Canvas.*
2. An anonymous written report, of 2000 words ($\pm 10\%$) **excluding** reference list. Your name and student ID should **not** appear anywhere in the report, including the metadata (filename, etc.). *Submitted through Canvas/Turnitin.*
3. Predictions for the test set of tweets submitted to the Kaggle[1] in-class competition described in Sec 6.

[1] https://www.kaggle.com/

Stage II Peer reviews (by May 19th):

1. Reviews of two reports written by your classmates, of 200-400 words each.

3 Data Sets

You will be provided with a geolocation-labelled *training* set of Tweets, a geolocation-labelled *development* set which you can use for model selection and tuning, and an unlabeled *test* set which will be used for final evaluation in the Kaggle in-class competition. Train, test and development sets are provided for each of the representations explained below. Each row in the data files contains the twitter user ID and the tweet representation as comma-separated values. Each line in the label files contains the corresponding location label.

Class Labels

In the provided data set, each tweet is labelled with one of four possible regions (i.e., class labels):

[MIDWEST, NORTHEAST, SOUTH, WEST]

Features

To aid in your initial experiments, we have applied some **feature engineering** to the raw tweets. You may use any subset of the representations described below in your experiments, and you may also extract your own features from the tweets if you wish. The provided representations are

1. full The raw tweets represented as a single string, e.g.,

User_ID,*"ill explain everything in a bit ill explain you everything"*

2. count which stands for "Bag of Words". We (1) filtered the words in the data set, removing very frequent and very infrequent words; and (2) mapped each word to a unique ID. The resulting mapping from remaining word strings to their ID is provided in `vocab.txt`. And (3), we represent each tweet as a list of (ID, word_count) tuples. E.g., assuming the following mapping (word:ID): {ill:0, explain:1, everything:2, in:3, a:4, ...}

User_ID,[(0, 2), (1, 2), (2, 2), (3, 1), (4, 1), ...]

word ID count

3. tfidf Same as "count" except that instead of word counts, we provide the tfidf value as a measure of feature importance. E.g.,

User_ID,[(0, 0.465), (1, 0.874), (2, 0.002), (3, 0.336), (4, 0.998), ...]

word ID tfidf

You can learn more about tfidf in Schütze et al. (2008).

4. glove300 We mapped each word to a 300-dimensional Glove "embedding vector". These vectors were trained to capture the meaning of each word. We then average the vectors of each word in a tweet to obtain a single 300-dimensional representation of the tweet. E.g.,

User_ID, 2.05549970e-02 8.67250003e-02 8.83460036e-02 -1.26217505e-01 1.31394998e-02 [. . .]

a_300-dimensional_list_of_numbers

You can consult Pennington et al. (2014) for more information about GloVe.

4 Tasks

Stage I

1. Feature Engineering (optional)

The process of engineering, or selecting, features that are useful for discriminating among your target class set is inherently poorly-defined. Most machine learning assumes that the attributes are simply given, with no indication from where they came. The question as to which features are the best ones to use is ultimately an empirical one: just use the set that allows you to correctly classify the data.

In practice, the researcher uses their knowledge about the problem to select and construct "good" features. What aspects of a tweet itself might indicate a user's location? You can find ideas in published papers, e.g., Cheng et al. (2010).

We have discussed three types of attributes in this subject: categorical, ordinal, and numerical. All three types can be constructed for the given data. Some machine learning architectures prefer numerical attributes (e.g. k-NN); some work better with categorical attributes (e.g. multivariate Naive Bayes) – you will probably observe this through your experiments. (1)

It is **optional** for you to engineer some attributes based on the **full** Tweets dataset (and possibly use them instead of – or along with – the feature representations provided by us). Or, you may simply select features from the ones we generated for you (count, tfidf, and glove300).

2. Machine Learning

Various machine learning techniques have been (or will be) discussed in this subject (Naive Bayes, Decision Trees, 0-R, etc.); many more exist. You are strongly encouraged to make use of machine learning software (2) and/or existing libraries in your attempts at this project (such as `sklearn` or `scipy`).

The objective of your learners will be to predict the classes of unseen data. We will use a **holdout strategy**: the data collection has been split into three parts: a training set, a development set, and a test set. This data will be available on the LMS.

1. The **training phase** will involve training your classifier and parameter tuning where required.
2. The **development phase** is where you observe the performance of the classifier. The development data is labelled: you should run the classifier that you built in the training phase on this data to calculate one or more evaluation metrics to discuss and compare in your report, using tables/diagrams. (3)

3. The **testing phase**: The test data is unlabeled; you should use your preferred model to produce a prediction for each test instance, and submit your predictions to Kaggle website; we will use this output to confirm the observations of your approach.

To give you the possibility of evaluating your models on the test set, we will be setting up a **Kaggle In-Class competition**. You can submit results on the test set there, and get immediate feedback on your system's performance. There is a Leaderboard, that will allow you to see how well you are doing as compared to other classmates participating on-line. assumptions to

You should *minimally* implement and analyse in your report one baseline, and at least two different machine learning models. **N.B.** We are more interested in your *critical analysis* of methods and results, than the *raw performance* of your models.

3. Report

You will submit an **anonymised** report of 2000 words in length ($\pm 10\%$), **excluding** reference list. The report should follow the structure of a short research paper, as discussed in the guest lecture on Academic Writing. It should describe your approach and observations, both in engineering (optional) features, and the machine learning algorithms you tried. Its main aim is to provide the reader with **knowledge** about the problem, in particular, **critical analysis of your results and discoveries** (or maybe some that you haven't!). The internal structure of well-known classifiers should only be discussed if it is important for connecting the theory to your practical observations.

- Introduction: a short description of the problem and data set
- Literature review: a short summary of some related literature, including the data set reference and at least two additional relevant research papers of your choice. (One option is Rahimi et al. (2018), as well as papers cited therein or in (Eisenstein et al., 2010; Pavalanathan and Eisenstein, 2015).)
- Method: Identify the newly engineered feature(s), and the rationale behind including them (Optional). Explain the methods and evaluation metric(s) you have used (and why you have used them)
- Results: Present the results, in terms of evaluation metric(s) and, ideally, illustrative examples
- Discussion / Critical Analysis: which should cover two aspects:
  - Contextualise** the system's behavior, based on the understanding from the subject materials
  - Discuss any ethical issues you may find with developing a geolocation classifier given the data and evaluation used in this assignment. Your discussion may touch on data selection, user discrimination, or other biases introduced in the pipeline. (You may use Pavalanathan and Eisenstein (2015) for inspiration (and cite it appropriately if you do), or come up with your own ideas.)
- Conclusion: Clearly demonstrate your identified knowledge about the problem
- A bibliography, which includes Eisenstein et al. (2010), as well as references to any other related work you used in your project. You are encouraged to use the APA 7 citation style, but may use different styles *as long as you are consistent* throughout your report.

** Contextualise implies that we are more interested in seeing evidence of you having thought about the task and determined reasons for the relative performance of different methods, rather than the raw scores of the different methods you select. This is not to say that you should ignore the relative performance of different runs over the data, but rather that you should think beyond simple numbers to the reasons that underlie them.

We will provide LaTeXand RTF style files that we would prefer that you use in writing the report. Reports are to be submitted in the form of a **single PDF file**. If a report is submitted in any format other than PDF, we reserve the right to return the report with a mark of 0.

Your name and student ID should **not** appear anywhere in the report, including any metadata (filename, etc.). If we find any such information, we reserve the right to return the report with a mark of 0.

Stage II

During the reviewing process, you will read two anonymous submissions by your classmates. This is to help you contemplate some other ways of approaching the Project, and to ensure that every student receives some extra feedback. You should aim to write 200-400 words total per review, responding to three *'questions'*:

- Briefly summarise what the author has done in one paragraph (50-100 words)
- Indicate what you think that the author has done well, and why in one paragraph (100-200 words)
- Indicate what you think could have been improved, and why in one paragraph (50-100 words)

5 Assessment Criteria

The Project will be marked out of 30, and is worth 30% of your overall mark for the subject. The mark breakdown will be:

Report Quality: (26/30 marks available)
You will produce a formal report, which is commensurate in style and structure with a (short) research paper. You must express your ideas clearly and concisely, and remain within the word limit (2000 words $\pm 10\%$) **excluding** reference list. You will include a short summary of related research. You can consult the marking rubric on the Canvas/Assignment 3 page which indicates in detailed categories what we will be looking for in the report.

Kaggle: (2/30 marks)
For submitting (at least) one set of model predictions to the Kaggle competition.

Reviews: (2/30 marks available)
You will write a review for each of two reports written by other students; you will follow the guidelines stated above.

6 Using Kaggle

The Kaggle in-class competition URL will be announced on LMS shortly. To participate do the following:

- Each student should create a Kaggle account (unless they have one already) using your Student-ID
- You may make up to 8 submissions per day. An example submission file can be found on the Kaggle site.
- Submissions will be evaluated by Kaggle for accuracy, against just 30% of the test data, forming the public leaderboard.
- Prior to competition close, you may select a final submission out of the ones submitted previously – by default the submission with highest public leaderboard score is selected by Kaggle.

- After competition close, public 30% test scores will be replaced with the private leaderboard 100% test scores.

7 Assignment Policies

7.1 Terms of Data Use

The data set is derived from the resource published in Eisenstein et al. (2010):

> Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277-1287).

This reference **must** be cited in the bibliography. We reserve the right mark of any submission lacking this reference with a 0, due to violation of the Terms of Use.

Please note that the dataset is a sample of actual data posted to the World Wide Web. As such, it may contain information that is in poor taste, or that could be construed as offensive. We would ask you, as much as possible, to look beyond this to the task at hand. If you object to these terms, please contact us (`lea.frermann@unimelb.edu.au`) as soon as possible.

Changes/Updates to the Project Specifications

We will use Canvas announcements for any large-scale changes (hopefully none!) and Piazza for small clarifications. Any addendums made to the Project specifications via the Canvas will supersede information contained in this version of the specifications.

Late Submission Policy

There will be **no extensions** granted, and **no late submissions** allowed to ensure a smooth peer review process. Submission will close at **5pm on May 14th**. For students who are demonstrably unable to submit a full solution in time, we offer to reduce the weighting of the mark of this assignment towards the overall course grade (but you will still have to submit your solutions by the deadline).

Academic Misconduct

For most people, collaboration will form a natural part of the undertaking of this project. However, it is still an individual task, and so reuse of ideas or excessive influence in algorithm choice and development will be considered cheating. We highly recommend to (re)take the academic honesty training module in this subject's Canvas. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy[2] where inappropriate levels of collusion or plagiarism are deemed to have taken place.

[2] http://academichonesty.unimelb.edu.au/policy.html

References

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287.

Pavalanathan, U. and Eisenstein, J. (2015). Confounds and consequences in geotagged twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rahimi, A., Cohn, T., and Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.