COMP30027 Machine Learning
Short Text Location Identification

Semester 1, 2019

Jeremy Nicholson & Afshin Rahimi & Tim Baldwin

THE UNIVERSITY OF MELBOURNE

Lecture Outline

1. Project 2 Intro

What is Short Text Location Identification? I

Source:

`https://twitter.com/andrew_lund/status/1122999744475815936`

What is Short Text Location Identification? II

Various useful sources of information:

- The message can be "geotagged"
- The user has set their location (n.b. possibly unreliable)
- Other users in the social network may have location information

What is Short Text Location Identification? III

... And if we don't have access to the user's metadata? The text itself can vary from location to location!

- Language/script
- Preference for certain words/phrases (lexical choice)
- Dialectical differences:
  - grammatical structures
  - spelling variation
- ...

词汇的

What is Short Text Location Identification? IV

- A difficult problem — the patterns are difficult to describe there is very little information within a single document
- The research community has some ideas about how to approach the problem, but we don't know how to solve it
- But perhaps Machine Learning can bridge the gap: perhaps we can learn how to solve this problem!

Building a Baseline Location Classifier

- Instance representation in form of "bag" (= multiset) of words
- Features:
  - Word frequencies
  - Metadata (if we have it)

One obvious problem:

- A typical document collection has *many* different words
- Most words are uninformative to the location ("the")

A less obvious, but bigger problem:

- There are many different locations

ML becomes difficult, and (often) slow

Improving on the Baseline

- Choosing the features more carefully
  - document pre-processing
  - feature selection
- Use grammatical structure
  - through Natural Language Processing techniques
  - ... or, just hack it with word sequences (but this increases the feature space *a lot*)
- Model the instances as authors instead of documents
- ... Collect more data...

Feature Selection for Multi-class problems I

Multiclass (e.g. Melbourne, Sydney, Brisbane, Perth) classification tasks are usually much more difficult.

- PMI, MI, χ^2 are all calculated *per-class*
- (Some other feature selection metrics, e.g. Information Gain, work for all classes at once)
- Need to make a point of selecting (hopefully uncorrelated) features for *each* class to give our classifier the best chance of predicting everything correctly.

Feature Selection for Multi-class problems II

Actual example (MI):

Sydney	Melbourne	Brisbane	Perth
sydney	melbourne	brisbane	perth
ovoawl	barometer	queensland	annemarie
perth	hpa	perth	brisbane
brisbane	temperature	uuu	wa
nsw	perth	qld	voodoo
waterpolosa	brisbane	uu	freo
melbourne	kmh	sydney	sydney

location

(good predictor for "not Sydney")

Feature Selection for Multi-class problems III

Actual example (χ^2):

Sydney	Melbourne	Brisbane	Perth
sydney	melbourne	brisbane	perth
ovoawl	barometer	queensland	annemarie
nsw	hpa	qld	wa
q	temperature	perth	freo
waterpolosa	kmh	uuu	voodoo
dwpcdevils	mm	uu	freodockers
waterpoloa	humidity	cynwel	gemmatognini

Project 2

Your job:

- Develop classifier(s) for short text location, given training/development data
- Produce a classifier that makes good predictions on the test data (hopefully! (-:)
- **Write a report** that explains what works and what doesn't, and explain why
- (Later) Write reviews for some reports written by students in this subject

Other places to get help

What if I have more questions?

- Chat with other students (principles, not details!)
- Post to the Discussion Forum
- Our office hours