COMP30027 Machine Learning
Blog Author Age Identification

Semester 1, 2018

Jeremy Nicholson & Tim Baldwin & Karin Verspoor

THE UNIVERSITY OF MELBOURNE

Lecture Outline

1. Project 2 Intro

What is Author Identification? I

Romeo & Juliet	Shakespeare
Hamlet	Shakespeare
Richard III	Shakespeare
Letter of Advice to the Queen	Bacon
The Sussex Speech	Bacon
The Plantation of Ireland	Bacon
Love's Labour's Lost	?

What is Author Identification? II

Various theories on how to determine authorship:

- Word choice
- Grammatical structures
- Spelling variation
- Handwriting
- Publication dates/places
- ...

What is Short Text Author Identification? I

Hillary Clinton @HillaryClinton · 9 Nov 2016

"To all the little girls watching...never doubt that you are valuable and powerful & deserving of every chance & opportunity in the world."

79K 679K 1.3M

Donald J. Trump @realDonaldTrump · 6h

Very thankful for President Xi of China's kind words on tariffs and automobile barriers...also, his enlightenment on intellectual property and technology transfers. We will make great progress together!

7.2K 13K 60K

A TOTAL WITCH HUNT!!!

60K 26K 101K

What is Short Text Author Identification? II

- More difficult, because there is less information within a given document
- Otherwise, same strategies apply

Building a Baseline Authorship Classifier

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

One obvious problem:

- A typical document collection has *many* different words
- Most words are uninformative to the author

A less obvious, but bigger problem:

- There are many different authors

ML becomes difficult, and (often) slow

Improving on the Baseline

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

Authorship Age

- Instead of considering the author problem, for Project 2, we have transformed the data so that the class is a category describing the age of the author:
  - 14-16
  - 24-26
  - 34-36
  - 44-46
- (Obviously, not all authors are in these categories!)
- Hypotheses: different age cohorts write text differently, and these differences can be inferred automatically
- In a conceptual sense, it isn't clear whether this problem is easier or more difficult that the authorship problem
- In an ML sense, this is definitely easier!

Project 2

Your job:

- Develop a classifier for blog author age, 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
- My office hours
- We might talk more in the lectures...