Being lazy with machine learning

For fun and profit

Problem: picking book title

- Book data from our vendors has inconsistent quality
- Very common case (titles for the same book from 3 vendors):
  - Urban Fortunes: The Political Economy of Place
  - Urban fortunes : the political economy of place
  - Urban fortunes

Our current solution: **pick longest**

- The “winner” in this case is: *“Urban fortunes : the political economy of place”*
- Correct title: *“Urban Fortunes: The Political Economy of Place”*

Simple fix: just add couple “if's” and be done with it?

- But there's more, for e.g.:
  - *\Accompaninge the Players\": Essays Celebrating Thomas Middleton*
  - *On Record #*
  - *Essentials of Pharmacology for Nurses.*
  - *AS Level Sociology: The Complete Course for the AQA Specification (The Complete Course for t...*
  - *The Kite Runner (Alex Awards (Awards))*
  - ..and many more

Clean up?

- Fixing title case is language specific
- Everything else is just too random

Combine both

- Write some code to rate “quality” of titles, pick best and then clean
- We will end up with a huge tree of rules, exceptions and cleaning actions.. I'm too lazy for this.

I heard there's this magic thing called “machine learning”

Idea is simple

- Pick bunch of examples
- Classify (manually)
- Train by feeding pairs of (class, example) to appropriate algorithm
- Magic
- … ??..
- We now have function which selects best likely class to given data
- I believe this is called “supervised machine learning”

First attempt

- TextBlob: Simplified Text Processing
  http://textblob.readthedocs.org/en/dev/
- “I have no idea what I'm doing”

Second attempt

- scikit-learn - Machine Learning in Python http://scikit-learn.org/
- “I still have no idea what I'm doing, but...”
- Provides various classifiers for vectorized data
- Book title is not exactly a vector, so here's what I've done...