Admistrivia

- Projects!

- Posted the last two assignment. One is serious—on “Alice in wonderland,” and the other is an R-exercise to make sure you can actually do the commands we have been talking about.

Slashdot quote

Linguists Out Men Impersonating Women On Twitter Posted by timothy on Thursday July 28, @06:58PM Hugh Pickens writes

“Remember when the Gay Girl in Damascus revealed himself as a middle-aged man from Georgia? On a platform like Twitter, which doesn’t ask for much biographical information, it’s easy (and fun!) to
take on a fake persona but now linguistic researchers have developed an algorithm that can predict the gender of a tweeter based solely on the 140 characters they choose to tweet. The research is based on the idea that women use language differently than men. ‘The mere fact of a tweet containing an exclamation mark or a smiley face meant that odds were a woman was tweeting, for instance,’ reports David Zax. Other research corroborates these findings, finding that women tend to use emoticons, abbreviations, repeated letters and expressions of affection more than men and linguists have also developed a list of gender-skewed words used more often by women including love, ha-ha, cute, omg, yay, hahaha, happy, girl, hair, lol, hubby, and chocolate. Remarkably, even when only provided with one tweet, the program could correctly identify gender 65.9% of the time. (PDF). Depending on how successful the program is proven to be, it could be used for ad-targeting, or for socio-linguistic research.”

Words

Aside: I titled the last lecture “you can Mathematize anything.” Mathematize is a word you will only hear about 30 times in your life. Still you probably didn’t think is exotic! (Method: the web is about 1000 times larger than what you will hear in your life.)
Example: author identification

(Frederick Mosteller and David L. Wallace. Inference and Disputed Authorship: The Federalist. Addison-Wesley, Reading, Mass., 1964.)

- Who wrote the federalist papers? (85 in all)

- Official writer: Publius

- Most likely though:
  - Hamilton:
    
    AFTER an unequivocal experience of the inefficacy of the subsisting federal government, you are called upon to deliberate on a new Constitution for the United States of America. The subject speaks its own importance; comprehending in its consequences nothing less than the existence of the UNION, the safety and welfare of the parts of which it is composed, the fate of an empire in many respects the most interesting in the world. It has been frequently remarked that it seems to have been reserved to the people of this country, by their conduct and example, to decide the important question, whether societies of men are really capable or not of establishing good government from reflection and choice, or whether they are forever destined to depend for their political constitutions on accident
and force. If there be any truth in the remark, the crisis at which we are arrived may with propriety be regarded as the era in which that decision is to be made; and a wrong election of the part we shall act may, in this view, deserve to be considered as the general misfortune of mankind.

– Madison

AMONG the numerous advantages promised by a well-constructed Union, none deserves to be more accurately developed than its tendency to break and control the violence of faction. The friend of popular governments never finds himself so much alarmed for their character and fate, as when he contemplates their propensity to this dangerous vice. He will not fail, therefore, to set a due value on any plan which, without violating the principles to which he is attached, provides a proper cure for it. The instability, injustice, and confusion introduced into the public councils, have, in truth, been the mortal diseases under which popular governments have everywhere perished; as they continue to be the favorite and fruitful topics from which the adversaries to liberty derive their most specious declamations. The valuable improvements made by the American constitutions on the popular models, both ancient and modern, cannot certainly be too
much admired; but it would be an unwarrantable partiality, to contend that they have as effectually obviated the danger on this side, as was wished and expected. Complaints are everywhere heard from our most considerate and virtuous citizens, equally the friends of public and private faith, and of public and personal liberty, that our governments are too unstable, that the public good is disregarded in the conflicts of rival parties, and that measures are too often decided, not according to the rules of justice and the rights of the minor party, but by the superior force of an interested and overbearing majority. However anxiously we may wish that these complaints had no foundation, the evidence, of known facts will not permit us to deny that they are in some degree true. It will be found, indeed, on a candid review of our situation, that some of the distresses under which we labor have been erroneously charged on the operation of our governments; but it will be found, at the same time, that other causes will not alone account for many of our heaviest misfortunes; and, particularly, for that prevailing and increasing distrust of public engagements, and alarm for private rights, which are echoed from one end of the continent to the other. These must be chiefly, if not wholly, effects of the unsteadiness and
injustice with which a factious spirit has tainted our public administrations.

– Jay

WHEN the people of America reflect that they are now called upon to decide a question, which, in its consequences, must prove one of the most important that ever engaged their attention, the propriety of their taking a very comprehensive, as well as a very serious, view of it, will be evident.

• Can’t go by the “ideas” since many were new and so couldn’t be related to the individual authors

• Use of grammer and syntax! (Or more accurately word counts)

• Look at the distribution of words:
  – Can’t use “freedom” and “government” since the different articles will use them differently based on topic
  – Can use “the”, “this”, “that”!

**Related problems**

• hand coded data

• spam
Expensive data

- We only have about 40 labelled documents
- But there are 1 million words in English
- We can’t use all of them
- Ideas:
  - Use most frequent
  - Use stepwise regression
- Still, we can’t end up with very many in our model
- But each will only contribute a small amount of accuracy.
- What to do? Naive bayes to the rescue.

Calibration of scales in science

Example:

- Suppose I had 100 scales
- I only have a two items of know weight
- I want to build a predictive model based on the 100 measurements.
- Easy: just average the weights!
• Better: weighted average (use variances for weights)

Naive bayes

See the wiki page on Naive bayes.