turn off setRes X
need 10 mins of setup for display issues.
leave desks as is: not discussion-oriented today.

make copies

9/5/13

Lecture #3 → 1 handout.
~~At 1st milestone~~ deadline is ...
(a) Correlates of review helpfulness
(b) The making of a research ~~project~~ paper. (WWW 2009)

(a) is continuing on the themes we ~~do~~ were discussing last time
(b) is a project on this topic that I was involved in.
I want to discuss it not just b/c it's on-topic,
but also to give some insight into the research process, ~~of~~

• starting w/ a little question, vs. starting w/ a big question
Our particular case was this

Sometimes people are resistant to starting w/ little questions,
And then are people who protest how incremental ~~other people's work~~ work seems to be 'these days'.

• On the other hand, I think that for younger grad students, too early a focus ~~on the~~ only on "big" questions can be a bit paralyzing.

• Also I don't know of any conference where everything is a breakthrough; the pace of science just usually isn't that fast.
• The joint progress of the research community tends to go at a rate of a little bit @ time. (those years where there were lots of breakthroughs were preceded by lots of groundwork)
– anyway, the version of the story I'll tell here is quite diff. from how I give talks on the work.

O.K. last time we did @ that exercise where ~~we~~ you ~~guys~~ guys wrote down some factors you ~~both of~~ personally look for in deciding whether a review is helpful.
Now let's look @ some features that have been considered in the literature.

→ First example: Jahna Otterbacher '09 'Helpfulness'... ~~a measure of online~~, Table 3 (pg 958)
(use reference tab, so people can see pub info (screen width/resolution permitting)

handout better??

I wanted to show this table b/c it divides the description nicely into ~~intuitive~~ the "~~to~~ general concept" of each feature vs. how it was implemented. ~~Some of these~~ You might want to choose a few for your pilot studies
Also, the thing is super-convenient.
(so, in terms of presentation, this is nice)
• These are what she tried, not all of which turned out to be significant
And, ~~not @ all a~~ she was not the first to introduce many of these ~~~~ in the context of review mining.
Link to survey of previous work on the course homepg.

4th column ("Explanation/Justification" is perhaps easiest to go through?) Observations?

intrinsic quality: both the review and, in (5)-(9), the reviewer

(1-3). "trick" "commonly" used for coverage/objectivity of a review

Note: citation to [6] is probably wrong.

just talk

(4, 10, 11)

extreme

different

(12) ~~similar~~ ~~Same~~ (this is like what Gilbert; Karahalios found some of their interviewees

commonalities saying: "A completely unique review wouldn't serve any purpose"

perhaps best to consider extreme wrt other review ratings

contextual row: " ~~as we saw w/ our example last time, sometimes helpfulness~~ ~~depends on other reviews, if you~~

seems to be a nod to the idea that relationship to other reviews can matter.

But, the actual features in this category don't seem that linked to that concept.

Looking @ the 2011 version of Ghose; Ipeirotis '11, table 1.

as another example of features that have been tried.

* note: some of these features were used to predict sales rank, so things like in the individual (so, helpfulness rating not used in helpfulness prediction)

more reviewer characteristics

more ~~focus~~ readability features (5th row)

subjectivity measures, using same idea that the product description can be treated as definitely objective material (as it happens, to train the subj. detector)

comment about devprob: could be to get confidence for subj. prob, or <sthg smart I can't recall>

And other studies look @ similar features, as well.

So, ~~these were the kinds of things that were in the air~~

Any comments/questions/, esp. re: people's potential pilot studies for A1? recommendations

One q: finding a very specific feature that few people have talked about, like the fault in the review example we saw last time?

O.K. now I want to talk about ~~so~~ the project that a group of collaborators and I did on this topic, while such ideas were "in the air".

→ Show dreamweaver entry for today, to see names.

- at the time ~~this~~ we started this, which was early 2008, ~~there were~~ there were a few papers looking @ ~~com~~ predicting review helpfulness. ~~So~~

And, Jon Kleinberg, who's a ~~theoretician by training~~ ~~professor of computer science~~ theoretical computer-scientist by training, was working w/ Georgi Kossinets, his postdoc, whose PhD was in sociology ~~w/~~ ~~but w~~ w/ Duncan Watts.

Together, they were fooling around with a ~~colle~~ crawl of Amazon reviews.

Meanwhile, Cristian was a first-year grad student who'd taken my grad course that fall, and was interested in working with me.

Jon & Georgi asked us if we'd like to join in wrt looking for interesting things to do w/ the review text, esp. ~~given past work @ Cornell on sentiment analysis~~.

We thought we'd start w/ ~~the~~ helpfulness prediction, as a potentially interesting variant of sentiment categorization [where you'd try to predict the # of stars a review gives, based on the text.]

"helpfulness prediction". 'real data': "x out of y found this helpful".

= regression? or (binary) classification?

we ended up going w/ this, choosing 60% as the gold-label decision threshold ~~thresd~~ [based on dataset stats, altho' it will turn out this didn't matter b/c we turned away from classification.

(@) problem: text features not adding to accuracy (vs. meta-features)

we tried a bunch of things, as you can imagine, altho' we ~~weren't~~ ~~as sophisticated as~~ didn't try anything as sophisticated as what's on yr handout.

so do you stop there and go, oh well?

~~Seems like a w~~

Sometimes the answer is yes!! And maybe not such an interesting prob anyway...

But, we thought @ least, let's (*) try to analyze errors → switch topics? maybe ~~text~~ NLP not suited to stuff.

reaction: { (@ maybe text doesn't matter! (*)) look @ the data, try to figure out what's wrong. are there interesting correlates/diffs b/twn text & metafeatures? & what meta-features doing?

(b) many duplicate reviews ~~(removed via plagiarism detection techniques)~~

writing this down b/c may be useful for your projects.

(c) due to cross-posting by Amazon among 'editions', we recrawled all the data

This is a difficult decision to make — to throw out all your data and start over.

But sometimes you gotta do it.

All the same, sometimes it's a bad idea.

- q's about: ~~were~~ were there ~~plagiarism~~ ~~or~~ duplicates you were going to care about?

(d) went through lots of correlation plots...

one point: someone noticed a correlation btwn agreement w/ avg & helpfulness. (slide 37 of gatech)

...> but maybe everyone simply has the same opinion? and their helpfulness is their str?

idea: look @ places where there's polarization, (when people don't all have same opinion. assumes helpfulness raters are the same as the writers [LM])

variance plot - q's that might

...> but maybe this just looks like it's not textual, and really there's a correlation in the text.

for σ^2 = max, there's sparse data in the middle, but for the lower σ's there's enough data

show plagiarized reviews.

- natural experiment: highly similar text pairs.

---

how do you 'sell' an 'obvious' result?