Representations

CMSC 723 / LING 723 / INST 725

Hal Daumé III [he/him]

10 Sep 2019

Logistics

- We…
  - Posted Office Hours
  - Responded to some issues on HW1
- For today:
  - Watch the Fairness in ML webinar
- For next time:
  - Homework 1 – due Thursday Sep 12 by 3:30pm
  - Read NLP 3.1—3.3

Bag of words

- Represent document by indicator (or count) of words therein
- Loses context
- Dimensionality = size of vocabulary

Bag of word ngrams

- Allows capturing some local context
  - Most useful for languages with more fixed word orders
- Dimensionality grows quickly with *n* (but not exponentially)
- **This should be your default baseline for all text classification** (with a linear model)

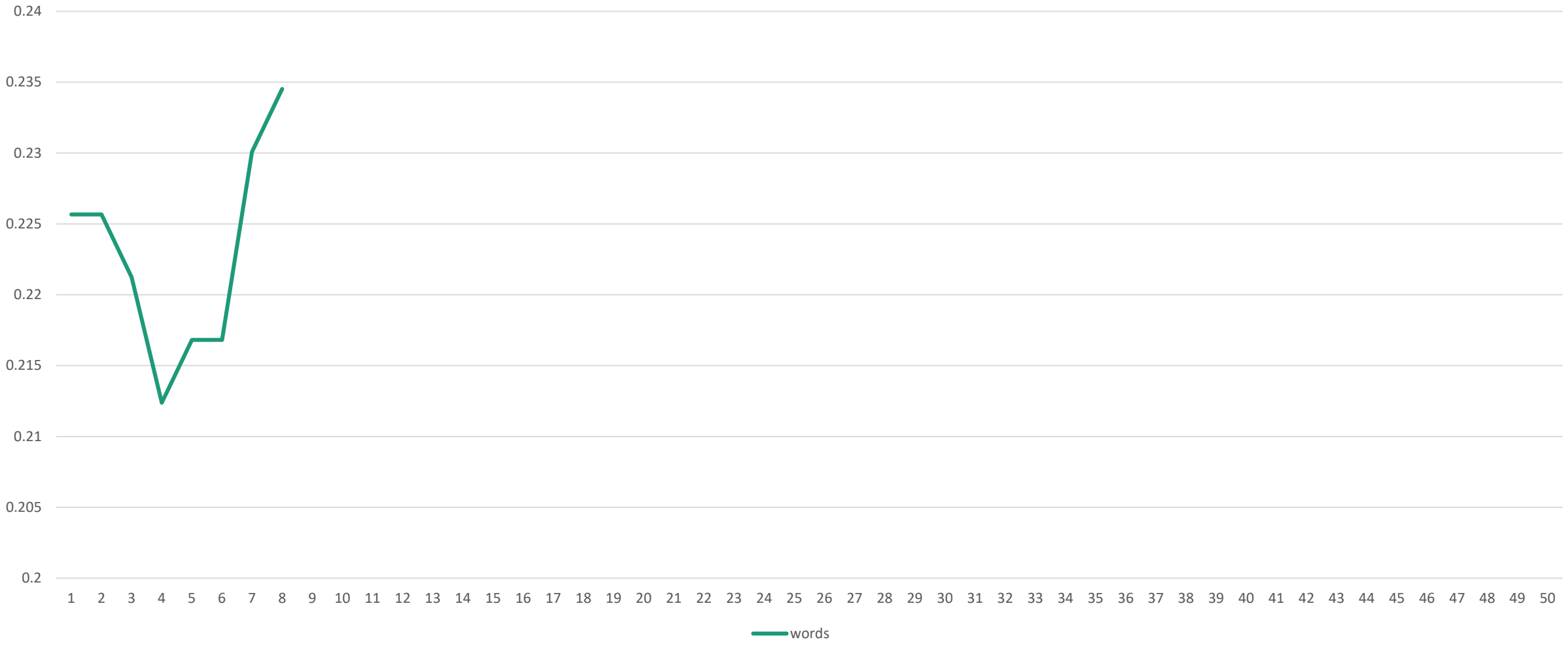
Predicting factuality of questions

- Data from Tsvetomila Mihaylova (github.com/tsvm/factcheck-cqa)

Q1: My son passed th entry exam in Doha Acadmey and they asked for the school fees; Is it recommended school?

Q2: i want someone to share my life with... but i want labrador retriever.. is there any chance i can find here in qatar?

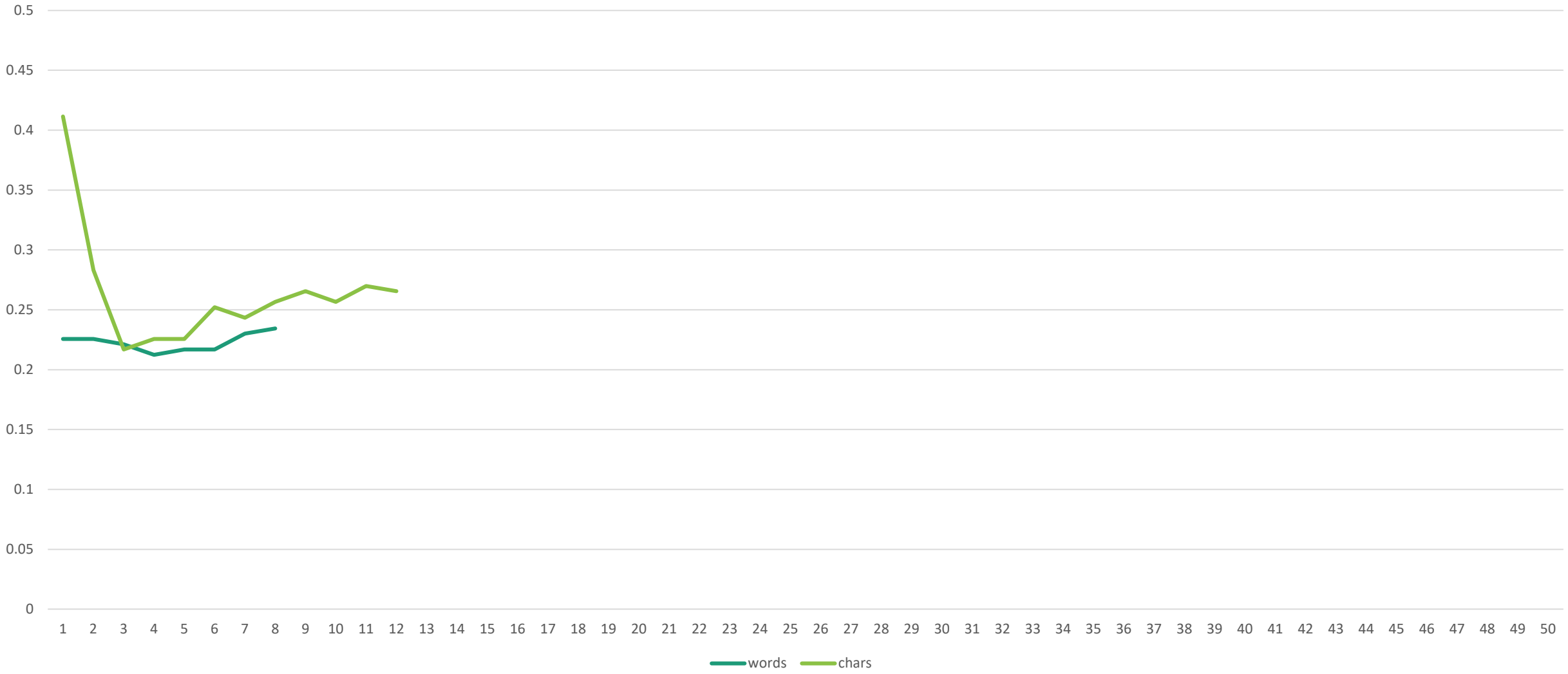
Predicting factuality of questions (error rate)



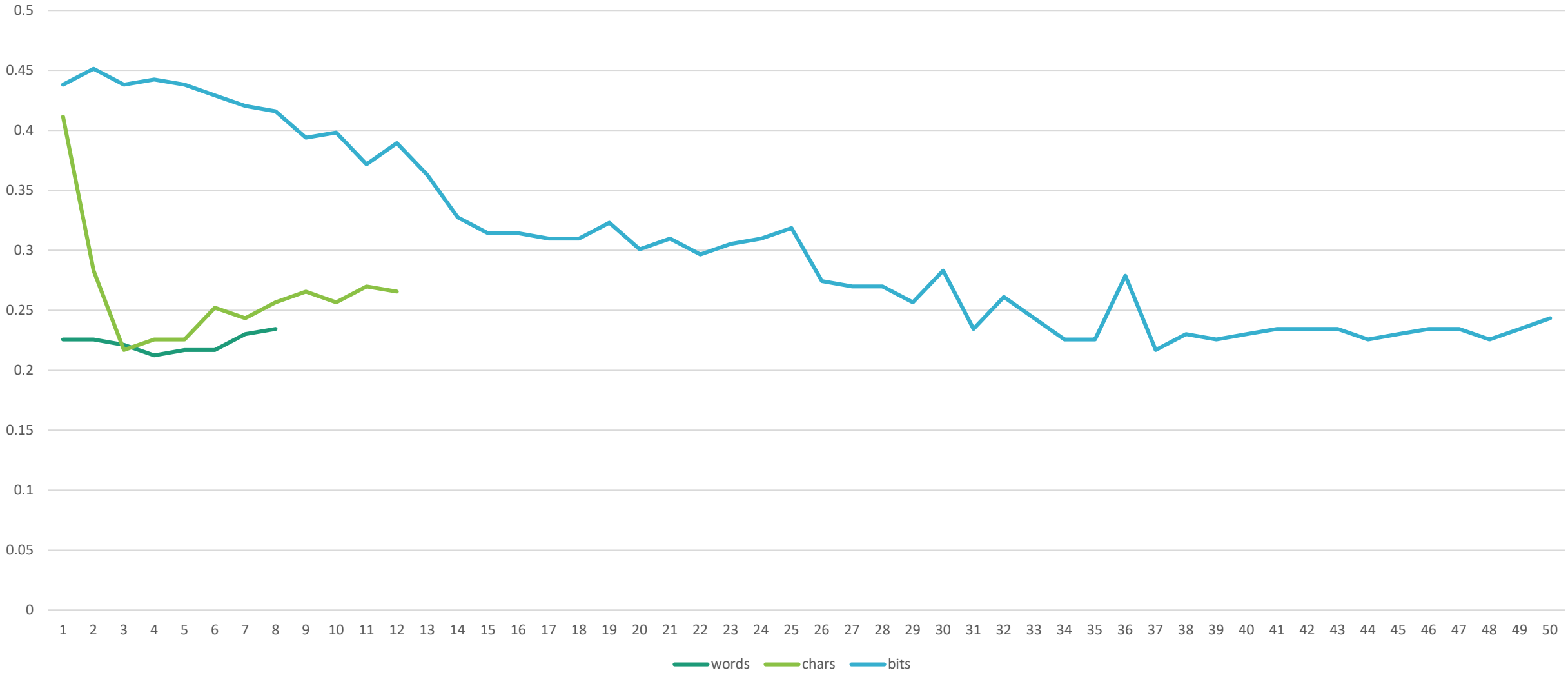
Bag of character ngrams

- Better handling of misspellings/creative language use
- In some ways easier than bag of words (no need to tokenize, etc.)
- More robust cross-linguistically
- Lots of redundant ngrams
- Very popular for stylometrics

Predicting factuality of questions (error rate)



Predicting factuality of questions (error rate)



A few more comments about stylometry

- Lots of (purported) applications…
  - Authorship identification
  - Identity identification (age, “gender”, “race/ethnicity”, etc.)
  - Psychological prediction (personality type, depression, suicidality, etc.)
- Lots of dual use of most of these applications
- Need to be *very careful* to avoid conflating stylometry with topic

Other off-the-shelf representations

- Spelling features:

  HelloWorld! -> `AaAa~`
  R2D2 -> `A0A0`
  socio-technical -> `a-a`

  (in general: [A-Z] -> A, [a-z] -> a, [0-9] -> 0, some punct stays the same and rest to ~, remove duplicates)
  very useful for things like named entity recognition

- Affix features: first/last *k* characters of each word

  unhappiness -> `un-, unh-, -ess, -es`
  defeated -> `de-, def-, -ted, -ed`
  went -> `we-, wen-, -nt, -ent`

  very useful for things like syntactic analysis (in languages with some morphology)

Before next class

- Finish HW1