Further normalization

Normalizing capital letters

- Us, us → us (if both are pronoun)
- us, US (could be pronoun and country)
- We can use heuristics:
  - lowercasing the beginning of the sentence
  - lowercasing words in titles
  - leave mid-sentence words as they are
- Or we can use machine learning to retrieve true casing → hard

Acronyms

- eta, e.t.a., E.T.A. → E.T.A.
- We can write a bunch of regular expressions → hard