Assignment 6

L245
Document Classification

Due Thursday, April 9

1. Select two articles in the Indiana Daily Student (http://www.idsnews.com/) written by different authors.

   (a) Give the title, author’s name, and date of each article.

   (b) What are some prominent stylistic differences between the two authors? Describe at least three differences.

   (c) Could these differences be detected automatically? How?

2. I want you to read certain parts of a paper by Koppel, Schler, & Argamon (2009) (Computational Methods in Authorship Attribution, http://dl.acm.org/citation.cfm?id=1461959 & available on oncourse under Resources) and answer the following questions:

   (a) **Introduction** (p. 9): Describe in your own words the difference between the profiling, needle-in-a-haystack, and verification problems.

   (b) **Profiling** (p. 15–17): Of the four types of profiling tasks mentioned: i) which seems most difficult? ii) how do the most important features differ for the tasks?

   (c) Pick one of the four profiling tasks (Gender, Age, Native Language, Personality) and some set of users on a social media or user-generated content site (Facebook, Twitter, etc.). Do the features mentioned in the paper seem to work for the users you look at? If so, provide examples and explain. If not (and even if so), describe other features you see that might distinguish the types of users (e.g., males from females).

   (d) **Bonus:** Read the section on **Authorship Verification** and describe in your own words: a) why this is such a hard problem; and b) the intuition behind the unmasking method, in particular the use of degradation curves (i.e., pretend you have to describe this to someone who isn’t technically-inclined).

3. Do question #3 from chapter 5 of the textbook (p. 151–152).

4. Do question #4 from chapter 5 of the textbook (p. 152).

5. Do question #5 from chapter 5 of the textbook (p. 152). (moved to next assignment!)

   • Section 5.5.1 has a miscalculation (http://purl.org/lang-and-comp/errata/), whereas the calculations on our slides 29–30 are correct.