Course title: Digital Humanities Prof. J. Savoy

Serie 8. Delta and authorship attribution

For this series, you can work in a team of two (or alone).

Question 1.

Based on the well-known corpus the *Federalist papers*, apply the Delta model to solve the authorship attribution with the twelve disputed papers. As features, you can use *m* most frequent word-types (with *m* varying from 30 to 500).

What is the importance of the choice of the (hyper) parameter *m*?

Which value of *m* is the most appropriate and why?

Do you see a short number of more problematic papers (in the test set) for this Delta model?

Question 2.

Apply the Delta model to the Twitter case (author profiling) to identify the author gender. The training set is composed to the two files `tweets.female.txt` and `tweets.male.txt`. The test sample is composed by the file `tweet.female.test.txt`, and the file `tweets.male.test.txt`.

Which value of *m* seems the most appropriate (because we face with many different authors in both the female and male group)?

Which accuracy do you obtain?

Question 3.

The Delta model is based on a (weighted) Manhattan distance. But considering small differences between some features is not so important. One can consider those small differences as not really important. On the other hand, we must take (only?) account of large differences or give to them a larger weight. How could you propose to change the distance computation inside the Delta model to take account of those comments?

Can you improve the accuracy with this modification?