Course title: Digital Humanities Prof. J. Savoy

Serie 6. Classification with the Naïve Bayes

For this series, you can work with a group of TWO.

A new folder was created in the folder "Exercises" entitled "Corpus" with different corpora useful for the next practical exercises.

During the preprocessing, the stemming function must be performed quickly. As a second operation, we need to remove the stop words.

Question 1: When removing the stopwords (a list between 150 top 500 items), it is faster to store the stopwords into a list or a dictionary? With python do you see a difference in performance between these two possible implementations (e.g., using the sms_spam.txt dataset)?

Question 2: Apply the naïve Bayes to solve the spam detection problem. You can use the file sms_spam.train.csv to train your classifier and sms_spam.train.cvs (361 entries) to evaluate your system (I obtained an accuracy of 77.28% with 60 features = 30 for each category). As features, you can consider the m most frequent words per category, or according to another way.

Question 3: Apply the naïve Bayes to solve the authorship attribution problem related to the *Federalist Papers* (federalist-papersNew2.csv) with the twelve disputed papers. As features, you can use the following words: {to, upon, would}. Of course, in this case, you must not apply the stopword list. We consider only Hamilton or Madison as the possible author of the disputed 12 papers.

Write your own Naïve Bayes procedure.

Use the log and compute a sum over all features instead of trying to multiply all the features (according to the two categories).