CENG463 Assignment #1

In this assignment, we will work on the problem of author profiling of given texts. The profiling dimension we'll focus on is the gender of author. To be more precise, we will predict the gender of an author using his/her text's vocabulary and syntactic constructions.

Here the key question is in order to characterize the author, what text features must be selected to be fed into the classifier. Literature categorize the types of features that can be used for authorship profiling as content-based features and style-based features. Evidence proved that the most effective style-based features for gender discrimination determiners and prepositions (markers of male writing) and pronouns (markers of female writing). As for content-based features, words related to technology (male) and words related to personal life or relationships (female) are proved to be most useful [1].

After the preprocessing and feature selection steps, you're expected to apply logistic regression algorithm to make the classification. Please run 10-fold cross validation to test your performance. Please follow the experiment design and performance assesment principles and guidelines we discussed in our last lecture. Code must be implemented in Python programming language. The dataset is provided in the assignment folder.

[1] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Commun. ACM, 52*(2), 119-123. doi: 10.1145/1461928.1461959

[2] http://pan.webis.de/clef17/pan17-web/author-profiling.html