Problem set guidelines

1. Should you have any questions regarding this homework, please post them on Slack.
2. You can discuss the problems with your classmates but do not share your code or the answers and do not use someone else’s solutions.
3. You can submit writing assignments in any form convenient for you. It could be LaTeX, MS Word, PDF or image. Please make sure it is of good enough quality.
4. For the coding assignment, please submit a Jupyter notebook with your solution. Make sure it can be reproduced without errors and with the same results when running it from scratch.
5. Python 3.6 is recommended for the coding assignment. You can use either the official distribution or Anaconda, which contains many of the required packages pre-installed for you.

Authorship Detection for Ukrainian Songs

In this homework, you are going to use your Natural Language Processing skills to build a classifier that, given the lyrics of a song, can predict whether it comes from Okean Elzy or Tartak.

For this task, you are provided with two datasets (train and test). Each dataset contains a set of text files with song lyrics (one .txt file per song), as well as the correct class labels for each file (labels.json, 0 = Tartak, 1 = Okean Elzy). You need to use the training set to build a model that will generalize well to the unseen test set

You are free to choose any approach for feature engineering or modeling you like. The only requirement is that your model shows at least 90% accuracy on the test set. Other than traditional bag-of-words approaches, you're free to explore the modeling resources available at lang.org.ua.

Tasks

1. **[10 points]** Read and preprocess the datasets.
2. **[25 points]** Extract the modeling features from the datasets. Warning: do not leak any information from the test dataset into your training pipeline.
3. **[15 points]** Build a classifier and make predictions on the training and the test sets.
4. **[10 points]** Compute the accuracy and ROC AUC on the training and the test set. Tune your feature engineering and modeling to achieve at least 90% accuracy and 0.95 AUC score.
5. **[10 points]** Maintain good code style. Organize your notebook into meaningful isolated sections, such as reading data, preprocessing, modeling, evaluation, etc. You can use Markdown formatting in the notebook to enrich the code with descriptions of your approach.
6. **[30 points]** Writeup. What options for feature engineering and modeling did you consider (we expect you to try more than one)? What worked well and what did not? What idea would you try implementing if you had more time? Please, give the reasoning behind your steps.