Оформление домашнего задания

Домашнее задание должно быть оформлено в виде pdf-файла или ipython-тетрадки с развернутыми ответами на все вопросы и описанием проделанных шагов. Данные к домашнему заданию можно найти на wiki.cs.hse.ru.

Домашнее задание 3 (до 09/03/17). Классификация имен

В этом домашнем задании мы рассмотрим задачу бинарной классификации. Пусть дано два списка имен: мужские и женские имена. Требуется разработать классификатор, который по данному имени будет определять мужское оно или женское.

1. **[1 балл]** Предварительная обработка данных: 1) удалите неоднозначные имена (те имена, которые являются и мужскими, и женскими одновременно), если такие есть; 2) создайте тестовое множество по следующему принципу: 20% от общего количества имен на каждую букву (т.е. 20% от имен на букву A, 20% имен на букву B, и.т.д.).

2. **[4 балла]** Используйте метод наивного Байеса для классификации имен: в качестве признаков используйте символьные n-граммы. Сравните результаты, получаемые при разных $n = 2, 3, 4$ по F-мере и аккуратности. В каких случаях метод ошибается?

   Для генерации n-грамм используйте `from nltk.util import ngrams`.

3. **[4 балла]** Используйте сеть с двумя слоями LSTM для определения пола. Представление имени для классификации в этом случае: 2-мерный бинарный вектор количество букв в алфавите $\times$ максимальная длина имени. Обозначим его через x. Если первая буква имени a, то $x[1][1] = 1$, если вторая – b, то $x[2][1] = 1$. Не забудьте про регуляризацию нейронной сети дропаутами. Если совсем не получается запрограммировать нейронную сеть самостоятельно, обратитесь к туториалу тут: `https://github.com/divamgupta/lstm-gender-predictor/blob/master/train_genders.py`. Сравните результаты, получаемые при разных значениях дропаута, разных числах узлов на слоях нейронной сети по F-мере и аккуратности. В каких случаях нейронная сеть ошибается?

4. **[1 балл]** Сравните результаты классификации разными методами. Какой метод лучше и почему?