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.

Агоритм сдачи домашнего задания появится в ближайшее время на странице курса.

Домашнее задание 1 (до 9/2/17). Сравнение стилей

В лингвистике выделяют 5 стилей речи (или текста): научный стиль, официально-деловой стиль, публицистический стиль, разговорный стиль, художественный стиль. Эта система стилей представляется вполне очевидной: в научном стиле написаны научные статьи и энциклопедии, в официально-деловом – официальные документы, в публицистическом – новости и любые сообщения в СМИ, в разговорном – блоги, неформальная переписка, чаты, в художественном – художественные тексты.

В этом домашнем задании вам нужно проверить следующую гипотезу: в текстах разных стилей частоты частей речи имеют разные характеры распределений.

1. **[2 балла]** Составьте самостоятельно как минимум две коллекции текстов разных стилей (например, коллекция текстов в публицистическом стиле и коллекция текстов в научном стиле). Коллекции текстов должны быть достаточно большие (порядка 5000 токенов). Посчитайте количество токенов и типов в каждой коллекции.

2. **[5 баллов]** Используя любой морфологический процессор, который вам нравится (pymorphy2, mystem), определите к какой части речи относятся слова из каждой коллекции текстов. При помощи nltk.FreqDist() составьте частотные словари: часть речи – количество слов, к ней относящихся.

3. **[3 балла]** Посчитайте коэффициент корреляции Спирмена для полученных на предыдущем шаге частот частей речи. На основании полученного значения, сделайте вывод: подтверждается ли гипотеза, сформулированная в задании? Если вы рассматривали больше двух стилей, можно ли утверждать, что один стиль больше похож на второй, чем на третий?