Предсказание ридабилити научно-популярных текстов

или история одной неудачи

Что такое ридабилити?

- Это “удобочитаемость” текстов: их понятность для читателя с содержательной точки зрения, стилистической, и терминологической

Как его считать?

- Обычно используются различные статистические характеристики длина текста, средняя длина предложений, количество “сложных” слов, и так далее.
- На основе этих характеристик рассчитываются специальные метрики, результат которых определяет уровень сложности текста

```
def flesch_RE(text):
    ASL = avg_sentence_length(text)
    ASW = avg_syllab_per_word(text)
    FRE = 206.835 - float(1.3 * ASL) - float(60.6 * ASW)
    return round(FRE, 2)
```

Причем тут науч-поп?

- Оценка ридабилити -- одна из задач НИСа по исследованию научно-популярных текстов. Можно ли определить ее при помощи машинного обучения?

Данные и инструменты

- Тексты из научно-популярных ресурсов (n+1, ПостНаука, Чердак, Индикатор, Полит.ру, Geektimes)
- Экспертная оценка текстов: выделение трех уровней сложностей (очень просто, средне, сложно). Корпус из 320 текстов (да, мало, но будем наращивать)
- Питоновский модуль для подсчета метрик со специально подобранными коэффициентами для русского языка

Метрики

- *Flesch reading ease (FRE)* -- индекс удобочитаемости Флеша. Считает соотношение общего количество слов, предложений и количества слогов. Чем меньше значение, тем сложнее текст (отсчет идет от 100 и иногда может уходить в минус)
- *Flesh-Kincaid Grade (FKG)* -- считает индекс Флеша и переводит его в шкалу от 0 до 20 (с возможным превышением), ставя в соответствие необходимый уровень образования (5 -- 5 класс, 12 -- выпускник школы, от 14 и выше -- студенты и специалисты, и т.д.). Такой перевод используется и в остальных метриках.
- *SMOG index* -- считает количество слов, в которых более 3 слогов и их количество в предложениях
- *Coleman-Liau index (CLI)* -- считает не количество слогов, а количество букв в словах и слов в предложениях
- *Dale chall readability score (DCH)* -- количество сложных слов и средняя длина предложений
- *Gunning fog* -- считает сложные слова (более 3 слогов) и их количество в тексте

Обучение

- Коллекция текстов с метками классов (1, 2, 3)
- С помощью модуля считаются статистические метрики для каждого текста. Их пришлось нормализовывать.

```
In [11]: x_train[:10]

Out[11]: [[39.43, 12.6, 13.0, 19.25, 7.65, 15.9],
          [46.01, 11.5, 12.9, 17.97, 7.63, 15.8],
          [12.02, 17.7, 15.4, 19.26, 8.57, 19.3],
          [23.53, 17.5, 16.9, 21.12, 8.93, 21.2],
          [24.75, 15.5, 14.8, 18.27, 8.34, 18.4],
          [40.24, 13.9, 14.8, 18.16, 8.15, 18.3],
          [13.09, 19.9, 18.3, 23.04, 9.46, 23.3],
          [47.52, 12.4, 12.8, 16.59, 7.24, 15.6],
          [12.39, 21.6, 18.9, 20.84, 9.38, 24.4],
          [44.66, 12.6, 13.3, 16.7, 7.55, 16.3]]
```

Обучение

- Берем классификатор
- Обучаем. Что же получается?

Метод опорных векторов

	precision	recall	f1-score	support
1	0.62	0.60	0.61	40
2	0.42	0.63	0.51	35
3	0.20	0.05	0.08	21
avg / total	0.45	0.49	0.45	96

```
[[24 15  1]
 [10 22  3]
 [ 5 15  1]]
```

Метод К-ближайших соседей

	precision	recall	f1-score	support
1	0.49	0.68	0.57	40
2	0.47	0.49	0.48	35
3	0.20	0.05	0.08	21
avg / total	0.42	0.47	0.43	96

```
[[27  9  4]
 [18 17  0]
 [10 10  1]]
```

Ну такое... Что тут можно сделать?

- Увеличить корпус для обучения? Увеличить количество метрик?
- А может, связи между статистическими метриками и пониманием научно-популярных текстов просто нет? (в конце концов, они все довольно сложные)
- Добавить более сложные признаки (модная тема): количество терминов, оценка тональности текста, количество абстрактных слов, и т.д...