Классификатор веб-страниц на базе SVM-Multiclass

© Р.Ф. Кузнецов
Балтийский Государственный
Технический Университет
ruslkuznetsov@gmail.com

Участие в дорожках

- Классификация веб-страниц

Прогоны

- svmmc (1-й прогон)
- svmmcei (2-й прогон с использованием метода извлечения значимой информации из веб-страниц)

svmmc (1-й прогон)

- В обучении используются только страницы сайта 1-го и 2-го уровня
- Размерность пространства признаков - 44430
- Используемая реализация Метода Опорных Векторов - SVM-Multiclass

svmmcei (2-й прогон)

- В обучении используются только страницы сайта 1-го и 2-го уровня
- Размерность пространства признаков - 39735
- Используемая реализация Метода Опорных Векторов - SVM-Multiclass
- Использование метода извлечения значимой информации из веб-страниц

Метод извлечения значимой информации из веб-страниц

- представление HTML-документа как совокупность структурных блоков
- поиск предложений в каждом из блоков
- поиск заголовков для блоков, отнесенных к содержательным

Результаты классификации (and)

	Recall	Precision	F1
1-й прогон	0.2772	0.1342	0.1343
2-й прогон	0.1700	0.0836	0.0839

Результаты классификации (or)

	Recall	Precision	F1
1-й прогон	0.0893	0.1017	0.0758
2-й прогон	6.77E-4	3.38E-4	4.51E-4

Выводы

- Используемый метод извлечения значимой информации из веб-страниц негативно сказывается на результате

Спасибо за внимание