Определение классов

Классы:

1. Elite Dangerous
2. Гарри Поттер
3. Метро
4. Звёздные войны
5. The Elder Scrolls
6. Другое

Метод векторизации

Для модели был выбран метод векторизации Bag of Words. Пусть: A – множество уникальных N-грамм в рамках обучаемой выборки. Здесь N-грамма – это группа слов, разделённых пробелами, каждое из которых является последовательностью букв или цифр, ограниченной пробельными символами или знаками препинания. Слова на русском языке приводятся к нормальной форме.

$$A = \{a_1, a_2, \dots, a_n\}$$

где a_i – уникальная N-грамма; n – размер множества A.

Тогда V – вектор признаков.

$$V = \{v_1, v_2, \dots, v_n\}$$

где v_i – частота встречаемости соответствующего слова a_i в конкретном тексте.

Векторизация в модели производится с помощью фильтра StringToWordVector библиотеки Weka, который преобразует строковые атрибуты в набор числовых атрибутов, представляющих информацию о вхождении слов из текста, содержащегося в строках. Итоговое множество A состоит из 3000 N-грамм, каждая из которых может включать от 1 до 3 слов.

Выбранные модели для обучения

Для тестирования были выбраны следующие модели:

- NaiveBayesMultinomialUpdateable – Обновляемый многочленный классификатор Наивного Байеса. Основное уравнение для этого классификатора: $P(C_i|D) = \big(P(D|C_i) \times P(C_i)\big)/P(D)$ (правило Байеса), где C_i – класс i, а D – документ.
- PART – Список решений. Используется принцип «разделяй и властвуй». На каждой итерации строится частичное дерево решений C4.5 и «лучший» лист становится правилом.
- J48 – Дерево решений C4.5.
- IBk – Классификатор K-ближайших соседей. Можно выбрать подходящее значение K на основе перекрестной проверки.
- REPTree – Алгоритм обучения пропозициональным правилам. Повторяющееся постепенное сокращение для уменьшения ошибок (RIPPER).
- JRip – Алгоритм обучения пропозициональным правилам. Повторяющееся постепенное сокращение для уменьшения ошибок (RIPPER). Оптимизированная версия IREP.
- DescisionTable – Простой классификатор большинства по таблице решений.

Обучение и тестирование моделей

Из множества статей было отобрано 900 статей из каждого класса. Всего в выборке присутствовало 5400 статей. Порядка 3600 случайных статей из выборки были использованы для обучения модели. Остальные 1800 – для тестирования обученной модели.

Матрицы неточностей представлены в таблицах 1-7 для моделей, упомянутых выше, соответственно, где:

- a – Elite Dangerous
- b – Гарри Поттер
- c – Метро
- d – Звёздные воины
- e – The Elder Scrolls
- f – Другое

Результаты тестирования полученных моделей отражены в таблице 8 с указанием для каждой модели показателей:

- $Accuracy = \frac{P}{N}$, где P – количество правильно определённых статей, N – общее количество статей;
- $Precision_c = \frac{P_c}{N_c}$, где P_c – количество правильно определённых статей в рамках конкретного класса c, N_c – общее количество статей в рамках конкретного класса c (в таблице приведено среднее значение по всем классам в столбце Precision);
- $Recall_c = \frac{TP}{TP+FN}$, где TP – количество статей, истинно-положительно определённых к конкретному классу c, FN - количество статей, ложно-отрицательно определённых к конкретному классу c (в таблице приведено среднее значение по всем классам в столбце Recall);

- $F_{Measure_c} = 2 * \frac{Precision_c * Recall_c}{Precision_c + Recall_c}$ (в таблице приведено среднее значение по всем классам в столбце F-Measure).

Таблица 1 – Матрица неточностей для NaiveBayesMultinomialUpdateable

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	294	2		1		1
	b		324		1	2	3
	c		2	279		3	6
	d	1	12	2	279		2
	e	3			1	291	8
	f		1		1		317

Таблица 2 – Матрица неточностей для PART

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	256		10	12	18	2
	b	3	294	4	3	22	4
	c	4		258	2	12	14
	d	15	9	6	252	9	5
	e	6	4	10	5	262	16
	f	6	6	1		13	293

Таблица 3 – Матрица неточностей для J48

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	259	4	6	11	14	4
	b	5	286	5	8	18	8
	c	4	8	245	5	24	4
	d	11	9	9	247	18	2
	e	10	16	11	2	247	17
	f	8	6	7	3	11	284

Таблица 4 – Матрица неточностей для IBk

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	268	2	1	2	24	1
	b	138	136	3		53	
	c	167	5	91		27	
	d	119	16	2	90	69	
	e	98	13	1		191	
	f	155	45	8	5	89	17

Таблица 5 – Матрица неточностей для REPTree

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	240	5	8	14	20	11
	b	8	241	19	5	22	35
	c	10	12	216	5	24	23
	d	21	20	21	200	19	15
	e	20	12	11	8	235	17
	f	18	6	9	6	20	260

Таблица 6 – Матрица неточностей для JRip

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	255	1	22	10	8	2
	b	5	263	50	3	8	1
	c	9	2	253	3	17	6
	d	23	5	30	225	8	5
	e	1	4	30	5	261	2
	f	1	7	14		8	289

Таблица 7 – Матрица неточностей для DescisionTable

		Классификатор					
		a	**b**	**c**	**d**	**e**	**f**
Факт	**a**	225	3	17	18	35	
	b	4	264	21		40	1
	c	43	9	222		16	
	d	26	31	12	192	34	1
	e	10	17	8		267	1
	f	22	42	24	3	16	212

Таблица 8 – Общие результаты тестирования моделей.

	Accuracy	Precision	Recall	F-Measure
NaiveBayesMultinomialUpdateable	97.2 %	97.2 %	97.2 %	97.2 %
PART	88 %	88.2 %	88 %	88 %
J48	85.4 %	85.6 %	85.4 %	85.5%
IBk	43.2 %	67.8 %	43.2 %	40.7 %
REPTree	75.8 %	76.4 %	75.8 %	75.8 %
JRip	84.2 %	86 %	84.2 %	84.6 %
DescisionTable	75.3 %	78 %	75.3 %	75.4 %

Вывод

В основном модели демонстрировали значение параметра Accuracy определения в рамках от 75 % до 90 %. Хуже всего себя показал классификатор на основе алгоритма IBk с точностью порядка 45%. Наилучшие показатели продемонстрировала модель на основе Наивного Байесовского распределения с показателем точности в 97%. Среднее значение Precision у почти всех моделей, за исключением NaiveBayesMultinomialUpdateable, выше, чем Accuracy у тех же моделей, в особенности у модели IBk, со значением в 69 %.

Также при анализе матриц неопределённости видно, что модель Naive Bayes Multinomial Updateable совершала значительно меньше ошибок в определении классов текстов по сравнению с другими моделями.

На основании полученных данных для классификатора была выбрана модель Naive Bayes Multinomial Updateable.