Определение классов

Классы:

1. Автомобили.
2. Технологии.
3. Политика.
4. Трейлер.
5. Игры.
6. Экономика.
7. Общество
8. Аниме.
9. Манга.
10. Интернет.
11. Наука.
12. Происшествия.
13. Спорт.
14. Культура.
15. Прочее.

Вектор признаков

Пусть: A – множество уникальных слов на русском языке (кроме предлогов, частиц, союзов, слов на латинице и неизвестных слов) в рамках обучаемой выборки. Тогда:

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

где a_i – i-ое уникальное слово в нормальной форме;

n – размер множества A.

Пусть: V – вектор признаков. Тогда:

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

где v_i – частота встречаемости слова a_i для входного текста.

Способ векторизации текстов

Шаги:

1. Вектор V для заданного текста автоматически заполняется нулями.
2. Из текста извлекаются нормальные формы слов $B = \{b_1, b_2, \dots, b_m\}$, где m – количество слов, которые не принадлежат служебным частям речи, написаны не на латинице и найдены в словаре морфологического анализатора.
3. Проверяется принадлежность каждого b_j множеству A.
4. Если $b_j \in A$ и $b_j = a_i, i \in 1..n$, то $v_i = v_i + 1$.

Обучение модели различными методами

Для обучения и сравнения были выбраны следующие 2 метода:

- Bag of Words
- Word2Vec

Оценки их результатов изображены на таблице 1 и таблице 2 соответственно.

Таблица 1 – Оценка Bag of Words

Accuracy	0.83
Precision	0.84
Recall	0.76
F1 score	0.79

Таблица 2 – Оценка Word2Vec

Accuracy	0.74
Precision	0.71
Recall	0.67
F1 score	0.68

На основании полученных данных был выбран метод Bag of Words.