**Слово**: *движение*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Определение** | **Пример** | **Частота** | **Ключевые слова** |
| 1 | Изменение положения предмета или его частей, перемещение; состояние, противоположное неподвижности, покою. | *Машина пришла в движение.* | 51 | Муравьиный, спокойный, непрерывный, скорость, механический, вид, броуновский, круговой, без, хаотический, реактивный, поступательный, постоянный |
| 2 | Перемещение в пространстве в каком-л. направлении; передвижение. | *Она продолжала свое движение в его сторону.* | 58 | Направление, трамвайный, начинать, начать, уличный, темп, прекращение, траектория, график, дорожный, встречный, воздушный, прекратить, план, медленный, ускорять, правосторонний |
| 3 | Изменение положения тела или его частей; телодвижение, жест. | *Его движения были быстры.* | 42 | Пластика, рука, искусство, координация, глаз, голова, резкий, палец, резкий, неверный, слаженность, угловатый, судорожный, делать, решительный, энергичный |
| 4 | Общественная деятельность, преследующая определенные цели. | *Революционное движение.* | 28 | Коммунистический, забастовочный, демократический, олимпийский, тимуровский, женский, рабочий, революционный, партизанский, общественный |
| 5 | Внутреннее побуждение, душевное переживание. | *Движение сердца.* | 11 | Мысль, сердце, совесть, душа, душевный, психологический |
| 6 | Количественное или качественное изменение; рост, развитие. | *Движение вперед по плану.* | 10 | Отражать, ожидать, развиваться, вперед, тормозить |

1. **Какого качества вам удалось достигнуть? Лучшее или худшее качество получилось в вашем случае по сравнению с результатами для английского языка?**

Baseline – 37,5%

Ниже представлены только лучшие модели из групп алгоритмов.

Trees: RandomForest – 48%

Meta: MultiClassClassifier – 47,5%

Lazy: KStar – 44%

Functions: MultilayerPerceptron – 48%

Bayes: NaiveBayes – 43%

Таким образом, лучшая модель (+ ввиду быстроты) – Random Forest. Основная ошибка моделей связана с несбалансированностью выборки.

Непонятно, что имеется в виду под результатами для английского языка. Если речь про слово interest, то безусловно да, мы получали качество выше 80%. Если речь о данных по слову motion, то непонятно, откуда взять данные.

1. **Сбалансирован ли ваш датасет по количеству значений? Как это влияет на итоговый результат?**

Датасет не сбалансирован. На три довольно больших класса приходится один средний (что не так уж плохо) и два маленьких (что плохо).

Влияние на результат: если смотреть в абсолютных частотах, алгоритм охотнее выбирает наиболее частотные классы, а наименее частотные вообще отказывается выявлять. С этой проблемой связано качество построенных моделей: наиболее частотное предсказание – класс 2.

1. **Какие значения лучше всего различаются? Какие хуже?**

На baseline хорошо находятся все первые значения, но к нему приписываются и все остальные значения, поэтому различение по остальным классам сильно хромает.

На лучшей модели лучше всего различается второе значение, более-менее различаются первое, третье и четвертое значение. Малочисленные значения пять и шесть не различаются.

1. **Улучшилось ли качество для ключевых слов?**

Результаты по baseline не улучшились.

Baseline – 37.5%

Лучшая модель на прошлом датасете улучшилась:

RandomForest – 52,5%

Попробовала другие модели:

Tree: HoeffdingTree – 71.5%

Meta: IterativeClassifierOptimizer, MultiClassClassifierUpdateable – 72%

Lazy: LWL – 69.5%

Function: SMO – 73%

Bayes: NaiveBayes – 71.5%

Соответственно, для нового датасета лучшая модель SMO.

Может показаться, что я переусердствовала с ключевыми словами. Однако с движением может использоваться открытый класс слов, поэтому я выделила действительно ключевые слова, но они оказались частотны в моей выборке.