# 1. Выберите себе многозначное слово в [этой таблице](https://docs.google.com/spreadsheets/d/19VRKiVOvL4R_40tnag6LCbL5SOjvl59puH_ggMb0Wz4/edit?usp=sharing), запишите его за собой.

Свет

# 2. Составьте список значений слова, взяв в качестве основы толкования из словаря МАС

|  |  |  |  |
| --- | --- | --- | --- |
| Значение | Примеры | Количество |  |
| 1. **Электромагнитное излучение, воспринимаемое глазом и делающее видимым окружающий мир.** | *Солнечный свет.*  *Свет свечи.* | 66 |  |
| **2. Освещение, характерное для какой-л. части суток.** Рассвет, восход солнца. | *Дневной свет.*  *На улицах был еще неопределенный, рассеянный свет белых ночей.*  *Еще семь часов до света!*  *Как и в прошлую ночь, едва забрезжил свет, она уже проснулась.* | 8 |  |
| **3. Источник освещения и приспособление для освещения в домах и на улицах.** | *Зажечь в комнате свет.* | 60 |  |
| **4. Место, откуда исходит освещение, освещенное место, пространство, где светло.** | *Встать спиной к свету.*  *Доктор взял ребенка на руки, быстро повернул к свету и заглянул в глаза.* | 8 |  |
| **5. Светлое место, пятно на картине, передающее наибольшую освещенность какого-л. участка изображаемого пространства.** | *Контрасты света и тени.* | 0 |  |
| **6. Блеск глаз под влиянием какого-л. чувства, радостное, ясное выражение лица.** | *Все лицо его светилось тогда каким-то внутренним светом.* | 1 |  |
| **7. То, что делает ясным, понятным мир, то, что делает радостной, счастливой жизнь. Употребляется как символ истины, разума, просвещения или радости, счастья.** | *Свет истины.*  *[Андрей Иванович] стал говорить о свете знания, о святости труда, о широком и дружном товариществе.* | 15 |  |
| **8. (*обычно со словом* «мой»*)* Ласковое, приветливое обращение к кому-л.** | *Свет ты мой! послушай меня, старика: напиши этому разбойнику, что ты пошутил.* | 2 |  |
| 1. **Земля со всем существующим на ней, мир1, вселенная.** | *Части света.*  *Путешествие вокруг света.*  *И знаю, что лучше этого народа никакого народа на свете нет!*  *Мигающий огонек горел на мысе Крильон, самом южном советском мысе в этой стороне света.* | 30 |  |
| 1. **Окружающие люди, общество. – не обязательное высшее** | *И сердце бросил в море жизни шумной, И свет не пощадил — и бог не спас!*  *[Ане] казалось, что весь свет видит ее дешевую шляпку и дырочки на ботинках.* Чехов, Анна на шее. | 14 |  |
| 1. **Ограниченный круг людей, составляющий высший слой привилегированных классов буржуазно-дворянского общества.** | *Большой свет.*  *Бывать в свете.*  *Положение в свете.* | 5 |  |

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.