Света Павлова

группа 131

Домашнее задание 5

Значения слова *ряд* по МАС:

1. предметы в ряду

(в ряду́). Совокупность предметов, лиц, расположенных один к одному, друг за другом, в одну линию. || Строй в одну линию; шеренга. || Места для сидения в театре, кино и т. п., расположенные в одну линию.

2. события в ряду

(в ряду́). Совокупность явлений, событий, отрезков времени, следующих одно за другим.

3. количество чего-либо

только ед. ч. (в ря́де) (*часто в сочетании с прил.* „целый“) *чего.* Некоторое, обычно немалое количество чего-л. *Ряд вопросов. Целый ряд причин.* □ *Во второй части своих сочинений Лессинг напечатал ряд писем, содержанием которых были исследования о старинной литературе.* Чернышевский, Лессинг, его время, его жизнь и деятельность. *В полку сохранился целый ряд воспоминаний о подвигах прежних Лугининых.*Мамин-Сибиряк, Человек с прошлым.

4. группа людей

мн. ч. (ряды́, -о́в). Совокупность лиц, принадлежащих к какой-л. организации, группе, среде и т. д.

5. магазины

(в ряду́). Ларьки или прилавки для торговли однородными товарами, расположенные на рынке в одну линию. || *Устар.* Несколько лавок, магазинов, торгующих однородными товарами и расположенных подряд, в одну линию в пределах специального здания. || *мн. ч.* (ряды́, -о́в). *Устар.* Торговое здание, где лавки, магазины расположены подряд.

6. пробор

(в ря́де). *Прост.* То же, что пробор.

7. агр.

Полоса шириной в один взмах косы, в один захват косилки; прокос. || Вал скошенной травы, хлеба и т. п.

8. матем.

(в ряду́). *Мат.* Совокупность величин, расположенных в определенной последовательности. || *Хим.* Совокупность соединений, из которых каждое находится в определенном отношении к предыдущему и последующему.

9. порядок

Прост. Последовательность, очередь. || Порядок.

Таким образом получаем следующий список значений. Размечаем 300 предложений из корпуса – на всякий случай, чтобы потом выбрать 200 и сделать выборку более сбалансированной. Количество вхождений того или иного значения указано в скобках:

1. предметы в ряду (100);
2. события в ряду (~3 → 0);
3. количество чего-либо (164);
4. группа людей (26);
5. магазины (5);
6. пробор (0);
7. агр. (0);
8. матем. (4);
9. порядок (0).

В группу 1 записывались идиомы типа *(поставить) в один ряд с, в одном ряду, другого ряда, первого ряда, из ряда вон.* Туда же записывались выражения типа *модельный ряд, синонимический ряд, звуковой ряд, ассоциативный ряд*, хотя, возможно, для них надо выделить новое значение.

В выборке оказалось сложно выделить значения типа 2:

() *Из оставшихся кораблей в Испанию прибыл после **ряда** злосчастных приключений лишь один—"Виктория".*

Главный признак выделения – упорядоченность событий или явлений. Однако часто определить, есть ли эта упорядоченность или нет, довольно сложно. Потенциальных значений типа 2 оказалось мало (меньше пяти), поэтому я записывала их в первую или третью группу. Предложение выше попало в группу 3, а, например, следующее в группу 1:

() *Шляпы он на этот раз каким-то чудом не потерял, и было похоже, что в **ряду** прочих обстоятельств его жизни это событие является одним из самых радостных.*

В группе 5 два вхождения – имя собственное *Охотный ряд*. Для интереса оставим эту группу, хотя можно было бы её присоединить к группе 1.

Для конечной выборки возьмем все вхождения групп 4, 5, 8 (всего 35 вхождений) и примерно поровну вхождений групп 1 и 3 (82 и 83). [Можно еще попробовать выкинуть идиомы, но не знаю, дойду до этого или нет.]

Фичи:

- предыдущее слово;
- следующее слово;
- часть речи предыдущего слова;
- часть речи следующего слова;
- число предыдущего слова;
- число следующего слова;
- число *ряда*;
- пять фичей наличия ключевых слов, условно отнесенных к каждой группе.

Вот такие получаются результаты (по F-мере) для разных моделей без ключевых слов:

class	**baseline**	**naive_bayes**	**logit**	**j48**	**random_forest**
1	0.617	0.805	0.820	0.696	0.869
3	0.365	0.913	0.886	0.824	0.874
4	0.000	0.807	0.655	0.667	0.579
5	0.000	0.000	0.200	0.000	0.333
8	0.857	0.000	0.857	0.000	0.400
av.	0.422	0.814	0.811	0.714	0.811

Для сравнения результаты этих же моделей на датасете interest:

interest	0.745	0.840	0.833*	0.763	0.826

* Эта модель обсчитывалась почти два часа))

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

Лучше всего в среднем оказался Naive Bayes (0,814). Logit и Random Forest не сильно от него отстают (0,811). Для датасета interest результат лучше по всем моделям, но не сильно.

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

Для групп 1 и 3 специально выбрано примерно одинаковое количество предложений (82 и 83 соответственно). Группа 4 в три раза меньше групп 1 и 3 (26), группы 5 и 8 – самые малочисленные (5 и 4). В общем выборка не сбалансирована, поэтому часто вхождения из более мелких классов попадают в два самых крупных. Хорошие модели (Naive Bayes, Logit) для первых трех групп показывают адекватные результаты (F-мера от 0,66 до 0,91). Для малочисленных групп результат может быть совсем плохим даже у них, хотя Logit дает F-меру 0,86 для 8 класса.

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

Все модели кроме baseline лучше всего различают группу 3. Потом идет группа 1, хотя отстает она не сильно. Класс 4 различается все еще лучше рандома, а классы 5 и 8 – чаще всего хуже. Естественно, группы, в которых больше вхождений, различаются лучше.

4. **Какие ключевые слова лучше всего указывают на то, какое значение слова должно встретиться? Попробуйте добавить в ваш датасет фичи "присутствие в примере ключевого слова". Улучшилось ли качество?**

Ключевые слова для групп:

1. предметы в ряду – *модельный, один, первый, второй, третий, четвертый, пятый, последний, предпоследний, верхний, задний, звуковой, ассоциативный, визуальный, вон, выходящий, синонимический, задний, длинный, стоять, стоявший, лежать, лежавший, сидеть, сидевший, двигаться, шагать, выстроиться, сесть, лечь, встать, между, театр, сцена*;
3. количество чего-либо – *целый, случай, страна, государство, параметр, мера, вопрос, ограничение, задача, преимущество, решение, причина, требование*;
4. группа людей – *вступление, вступить, трутень, поклонник, пополнить, пополнение, товарищ, член, партийный*;
5. магазины – *охотный, калашный, торговый*;
8. матем. – *натуральный, число, математический*.

Насколько целесообразно выделение ключевых слов, если большинство из них – предшествующие или последующие *ряда*, а в моделях эти два параметра уже учитываются? Результаты показывают, что целесообразно, но как так получается?

С ключевыми словами результаты такие:

class	baseline	naive_bayes	logit	j48	random_forest
1	0.617	0.894	0.847	0.790	0.854
3	0.365	0.923	0.916	0.898	0.889

4	0.000	0.909	0.745	0.714	0.615
5	0.000	0.750	0.727	0.000	0.000
8	0.857	0.571	0.889	0.000	0.400
av.	0.422	0.898	0.860	0.790	0.807

Для Naive Bayes, Logit, J48 качество ощутимо улучшилось. В Naive Bayes для каждого класса F-мера больше 0,5, что радует. Лидерство сохранилось за Naive Bayes: у него результат улучшился на 8 п.п. (с 0,814 до 0,898).