Лемматизация итальянских слов

Азат Давлетшин

20 мая 2014 г.

1 Формулировка задачи

Имеется два множества $\mathbf{X}$ и $\mathbf{Y}$. $\mathbf{X}$ - множество слов итальянского языка. $\mathbf{Y}$ - множество пар (слово в начальной форме + часть речи). Имеются отображения $f : \mathbf{X} \to \mathbf{Y}$ и f^* - сужение f на некоторое подмножество $\mathbf{X}$. f - считается неизвестным, f^* - задано. Задача: экстраполировать f^* на все множество $\mathbf{X}$ как можно точнее. Критерий качества: доля правильных ответов на тестовой выборке.

2 Сведение к задаче классификации

Нужно выделить классы так, чтобы по классу можно было однозначно получить начальную форму и часть речи. Кроме того, желательно получить как можно меньше классов. В моем решении классом является следующий объект: (toDelete, toAdd, POS)

toDelete - целое число. Количество символов, которые нужно удалить с конца слова.

toAdd - строка. Что необходимо добавить в конец слова.

POS - символ. Часть речи

На исходных данных получилось примерно 250 классов.

3 Выделение признаков

Очевидно, что наиболее значимыми являются окончания. Для упрощения кодирования в качестве признаков были выбраны всевозможные суффиксы. Для того, чтобы получить матрицу признаков с помощью CountVectorizer, был сделан следующий хак: всевозможные суффиксы записывались в одно предложение. Например слово «Азат» $\rightarrow$ «Азат зат ат т». Далее запускался CountVectorizer в режиме «word», с параметром ngram_range = (1,1) - очевидно почему.

4 Классификатор

Поскольку фичи сильно разрежены, то в качестве классификатора был выбран LinearSVC со стандартными параметрами.

5 Результат

Время работы алгоритма составило примерно 15 минут. Результат на kaggle.com: **87,626%** качества.