Языковые модели

1. Предсказание следующего слова

Одной из задач обработки текстов на естественном языке выступает задача предсказания следующего слова по предшествующим. Примерами таких задач являются задача предсказания следующего слова при наборе текста в смартфоне или дополнение поисковых запросов.

Данная задача может быть сведена к оценке вероятностей встретить каждое из возможных слов после имеющегося. Соответствующая языковая модель примет вид:

$$w^* = \text{argmax}_{w_k \in V} P(w_k \mid w_{k-1}), \tag{1}$$

$$P(w_k \mid w_{k-1}) = \frac{P(w_k)\, P(w_{k-1} \mid w_k)}{P(w_{k-1})}, \tag{2}$$

где V – множество всех возможных слов, а $P(w_k \mid w_{k-1})$ – вероятность встретить слово w_k после w_{k-1}.

Для удобства полученную модель можно представить в следующем виде:

$$w^* = \text{argmax}_{w_k \in V} [\log P(w_k) + \log P(w_{k-1} \mid w_k)]. \tag{3}$$

Иногда по последнему слову оказывается невозможным определить следующее, поскольку в нем не учитывается контекст. Например, если последнее слово является союзом, то приемлемые варианты следующего слова определяет и предшествующее данному союзу слово. Оценка вероятности в данном случае проводится на основе m последних слов. Таким образом, модель (1)-(2) примет вид:

$$w^* = \text{argmax}_{w_k \in V} P(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-m}), \tag{4}$$

$$P(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-m}) = ((P(w_k)\, P(w_{k-1}, w_{k-2}, \ldots, w_{k-m} \mid w_k)) / P(w_{k-1}, w_{k-2}, \ldots, w_{k-m})). \tag{5}$$

Предположим, что текущее слово w_k зависит только от того, какие слова встретились перед ним и не зависит от того, в каком порядке они встретились. Тогда

$$w^* = \text{argmax}_{w_k \in V} \left[P(w_k) \prod_{i=1}^{m} P(w_{k-i} \mid w_k) \right] = \text{argmax}_{w_k \in V} \left[\log P(w_k) + \sum_{i=1}^{m} \log P(w_{k-i} \mid w_k) \right]. \tag{6}$$

Полученная языковая модель также используется для решения задачи классификации документов.

1.1. Наивный байесовский классификатор

Пусть стоит задача определить категорию новостей по их тексту. Тогда оценки вероятностей принадлежности новости к каждой $c \in C$ категории могут быть найдены по формуле:

$$c^* = \text{argmax}_{c \in C} \frac{P(c)\, P(d \mid c)}{P(d)}. \tag{7}$$

Здесь $P(c)$ – вероятность встретить новость данной категории, рассчитываемая по формуле:

$$P(c) = \frac{N_c}{N}, \tag{8}$$

где N_c – количество новостей в категории c, N – общее число новостей;

$$P(d \mid c) = P(w_1 \mid c)\, P(w_2 \mid c) \ldots P(w_m \mid c) = \prod_{i=1}^{m} P(w_i \mid c), \tag{9}$$

где w_1, ..., w_m – слова, встретившиеся в новости d (в предположении, что все слова независимы), а

$$P(w_i \mid c) = \frac{\nu_{i\,c} + 1}{\sum_{i' \in V} (\nu_{i'\,c} + 1)} = \frac{\nu_{i\,c} + 1}{|V| + \sum_{i' \in V} \nu_{i'\,c}} \tag{10}$$

после применения сглаживания Лапласа для устранения проблемы неизвестных слов;

$P(d)$ – оценка вероятности встретить новость, состоящую из данного набора слов. Поскольку $P(d)$ не оказывает влияния на результат, то итоговая модель классификации может быть представлена в виде:

$$c^* = \operatorname{argmax}_{c \in C} \left[\log P(c) + \sum_{i=1}^{n} \log P(w_i \mid c) \right]. \tag{11}$$

1.2. Оценка условных вероятностей в языковой модели на основе словаря

Условные вероятности в формуле (3) могут быть оценены непосредственно по имеющемуся словарю слов. Для этого необходимо рассмотреть имеющиеся в словаре биграммы, тогда:

$$P(w_{k-1} \mid w_k) = \frac{\nu_{(w_{k-1}, w_k)}}{\sum_{(w_i, w_j) \in V} \nu_{(w_i, w_j)}}, \tag{12}$$

где $\nu_{(w_{k-1}, w_k)}$ – частота встречаемости словосочетания $(w_{k-1},\ w_k)$.

Задание 1

Разработать систему предсказания следующего слова по предыдущим.

Варианты реализации:
1) загрузить любой текст и оценить вероятности на основе данного текста. При работе в Mathematica можно воспользоваться функцией ExampleData["Text",...];
2) воспользоваться данными с сайта http://norvig.com/ngrams/ или найти другие словари частот встречаемости слов.

Задание 2

Языковая модель (6) не учитывает порядок предшествующих слов. Информацию о порядке слов можно добавить путем оценки расстояния до каждого из предшествующих слов. Например, в предложении «Счастье есть удовольствие без раскаяния» расстояние между словами «счастье» и «раскаяния» равно 4.

Сравнить результаты предсказаний с учетом расстояний между словами и без него на том же тексте, на котором модель была обучена.