# N-gram（n元组）语言模型

1. a large _____ (A. Monday B. screen C. water)

2. I eat a large _____ (A. salad B. screen C. water)

| | | |
|---|---|---|
| “a large” | : | 47965 |
| “a large Monday” | : | 0 |
| “a large screen” | : | 89 |
| “a large water” | : | 36 |
| “eat a large” | : | 15 |
| “eat a large salad” | : | 1 |
| “eat a large screen” | : | 0 |

a large water fountain
A large water tank
a large water bottle
a large water buffalo
a large water jug
a large water jar
a large water container
a large water tower
a large water garden
…

https://corpus.byu.edu/coca/

# N-gram（n元组）语言模型

n-gram是一种统计语言模型，根据前n-1个项来预测第n个项。

在NLP应用层面，项可以是音素（语音识别应用）、字符（输入法应用）、词（分词应用）……

一般来讲，可以从大规模文本或音频语料库生成n-gram模型。

习惯上，1-gram称为unigram，2-gram称为bigram，3-gram称为trigram……

# N-gram（n元组）语言模型

$$S = w_1 w_2 \ldots w_{n-1} w_n$$

$$P(S) = P(w_1)P(w_2 \mid w_1)\ldots P(w_n \mid w_1, w_2, \ldots, w_{n-1})$$

$$= \prod_{i=1}^{n} P(w_i \mid w_1 \ldots w_{i-1})$$

$$P(S) = \prod_{i=1}^{n} P(w_i) \qquad \text{unigram (n=1)}$$

$$P(S) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}) \qquad \text{bigram (n=2)}$$

$$P(S) = \prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1}) \qquad \text{trigram (n=3)}$$

# N-gram（n元组）语言模型

$$P(w_n \mid w_1 \ldots w_{n-1}) = \frac{\mathrm{count}(w_1 \ldots w_n)}{\mathrm{count}(w_1 \ldots w_{n-1})}$$

# N-gram（n元组）语言模型

| # | freq | verb | the | noun |
|---|---|---|---|---|
| 1 | 3446 | opened | the | door |
| 2 | 1889 | tell | the | truth |
| 3 | 1874 | telling | the | truth |
| 4 | 1813 | open | the | door |
| 5 | 1471 | opens | the | door |
| 6 | 1341 | closed | the | door |
| 7 | 1256 | solve | the | problem |
| 8 | 1238 | tell | the | story |
| 9 | 1168 | change | the | way |
| 10 | 1108 | use | the | word |
| 11 | 1077 | answer | the | question |
| 12 | 1068 | meet | the | needs |

…

https://www.ngrams.info/coca/v_the_n.txt