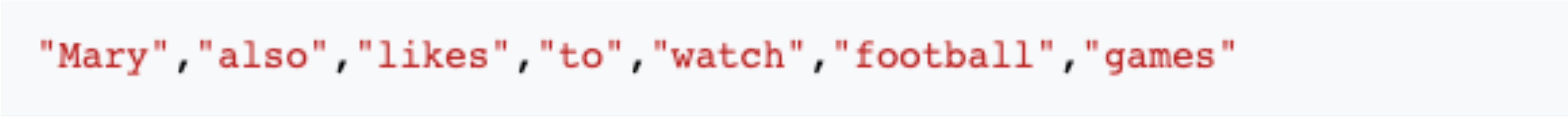
Bag-of-words

demo

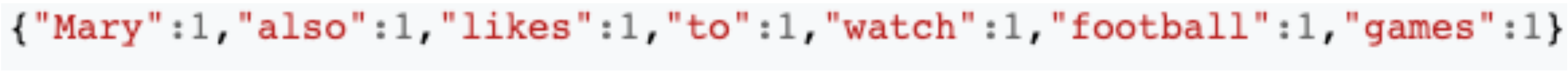
- Translation: represent text as a histogram of words
- Strips all syntactic and structural information from the text
- Examples:

Utterance



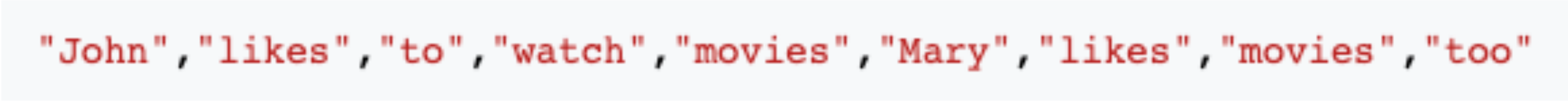
```
"Mary","also","likes","to","watch","football","games"
```

BOW



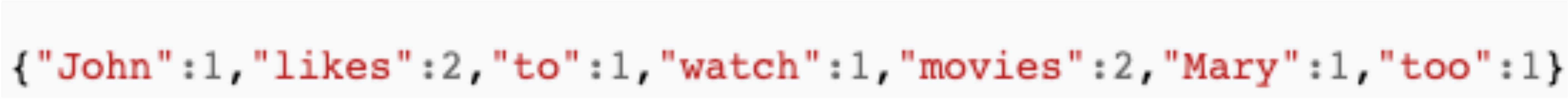
```
{"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1}
```

Utterance



```
"John","likes","to","watch","movies","Mary","likes","movies","too"
```

BOW



```
{"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1}
```

- Example taken from Wikipedia

Probabilistic text classification using BOW features

- Recall the principle of Maximum Likelihood Estimation (MLE):

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ P(\mathbf{X}, \mathbf{y}; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left(\prod_{i=1}^{M} P(\mathbf{x}_i, y_i\ ; \boldsymbol{\theta}) \right)^{\frac{1}{M}} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \frac{1}{M} \sum_{i=1}^{M} \log P(\mathbf{x}_i, y_i\ ; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} \log P(\mathbf{x}_i, y_i\ ; \boldsymbol{\theta})
\end{aligned}
$$

where

$\boldsymbol{\theta}$ = parameters to estimate

$\mathbf{X} \in \mathbb{R}^{M \times N}$ ← BOW features

$\mathbf{y} \in \{1,\ldots,K\}^M$

M = number of observations

N = vocabulary size

K = number of categories/classes