

## 文本语料

- 无结构文本
- 半结构文本 wikipedia

简单预处理

- 去除标点
- 去除停用词
- 去除低频文章

...

## 预处理后的文本语料

统计特征

- 词频
- TF-IDF

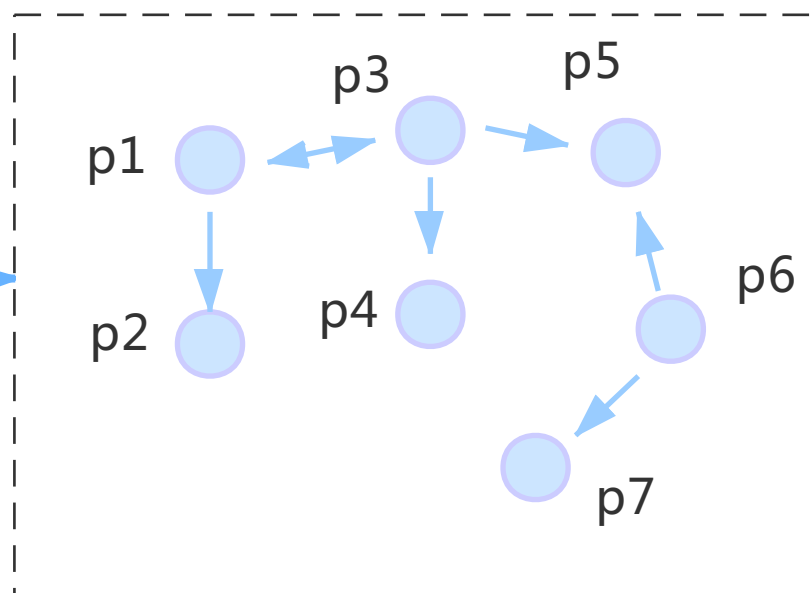
...

## 词典库

+

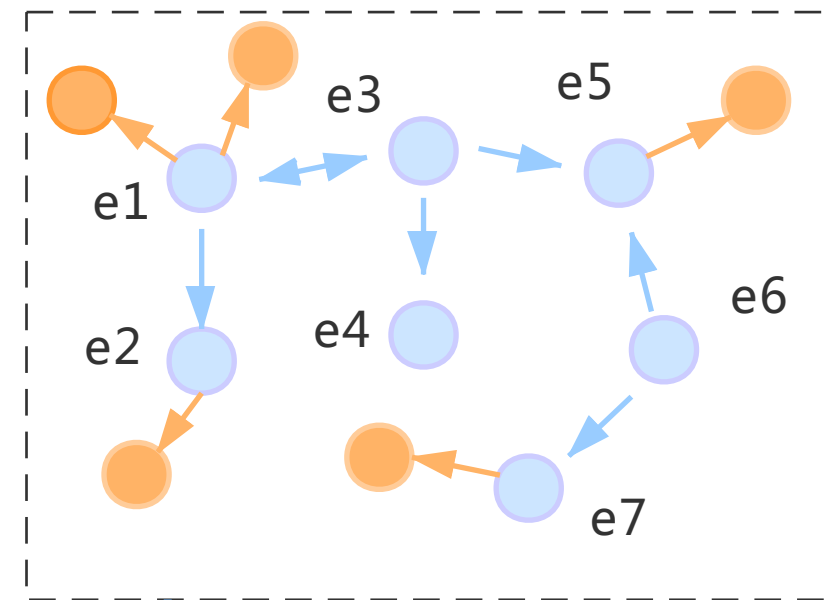
w1</br>  
w2</br>  
w3</br>  
...

信息抽取



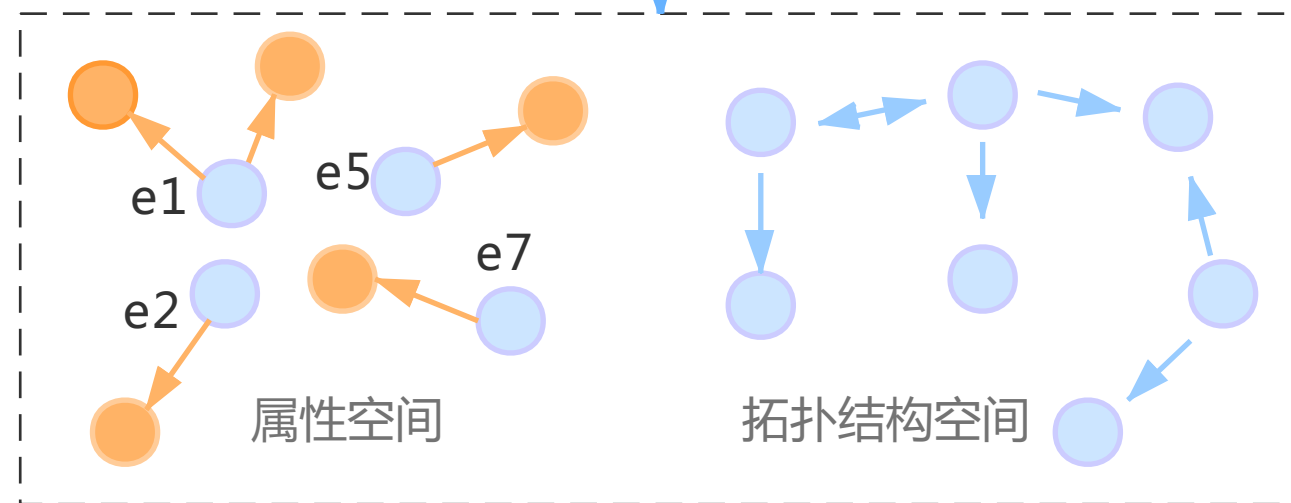
SPARQL

$$p_i = e_i \\ (2)$$



语义关联度  
计算

(3)



实体-实体

词语页面之间的  
对应关系

词语-实体

词语-词语

$\vec{v}_1$

$\vec{v}_2$