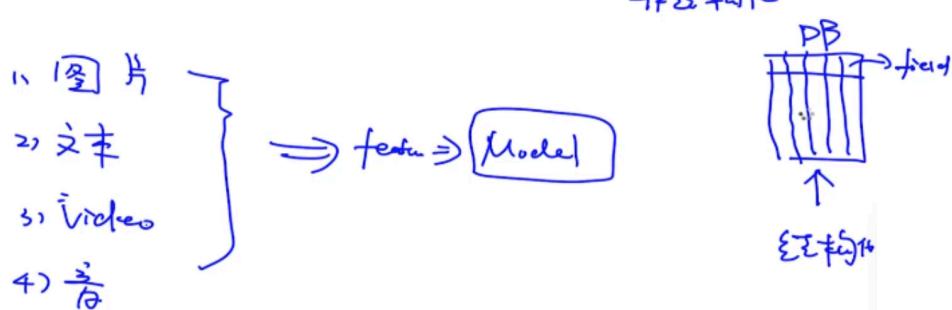


1. Information extraction (IE)

- 信息抽取概要
- 信息抽取应用场景
- 命名实体识别介绍
- 搭建命名实体识别分类器
- 文本的特征工程
- 特征编码
- 关系抽取介绍
- 基于规则的方法
- 基于监督学习方法
- Bootstrap方法
- Distant-supervision方法
- 无监督学习
- 实体消歧
- 实体统一
- 指代消解
- 句法分析
- CKY算法

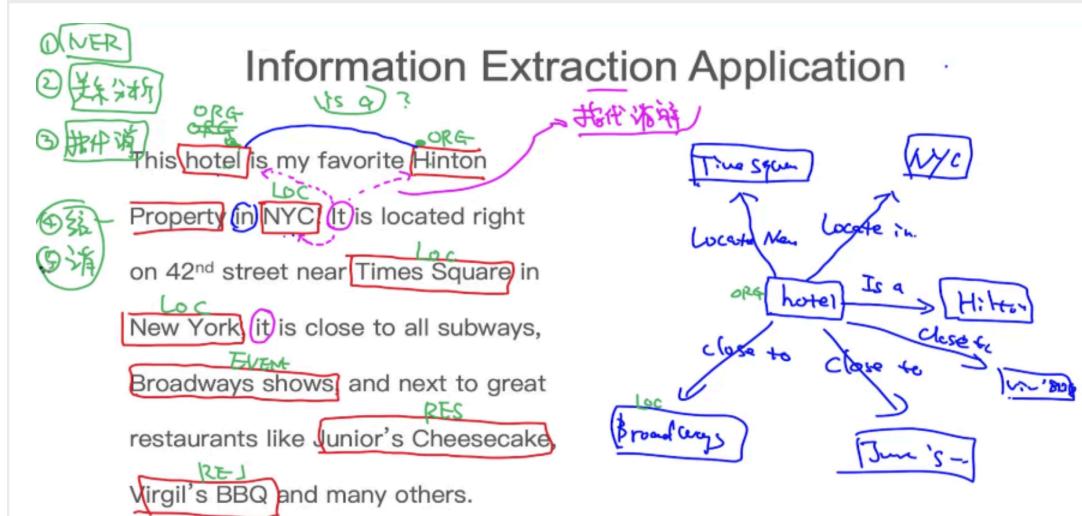
Extract Information from Unstructured Text



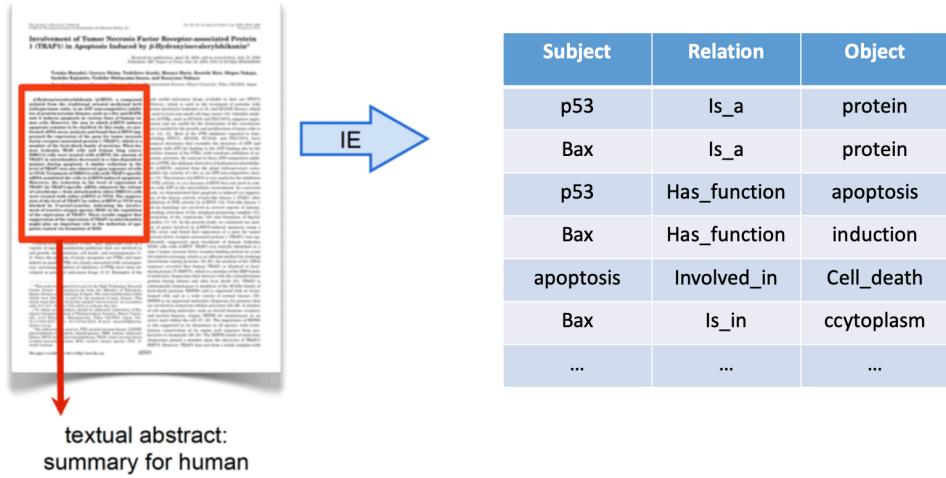
1.1 task

1. 抽取实体 (entities)
人, 地名, 时间
2. 抽取关系 (relations)
位于, 工作在, 部分

1.2 IE application



指代消解：it 到底是指哪个
在做关系抽取的时候，实体的类型是很重要的



1.3 more applications

知识库的搭建；

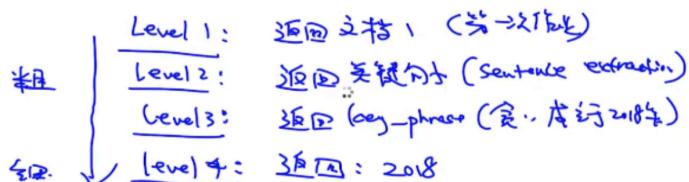
google scholar ; (需要实体统一)

用户库 : rapleaf, spoke、购物引擎、产品搜索、专利分析、证券分析、问答系统

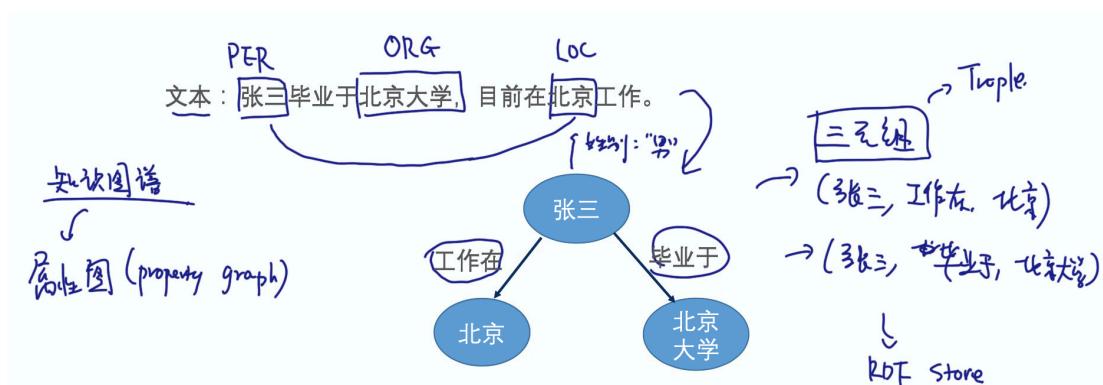
1.4 Search Engine vs Question Answering

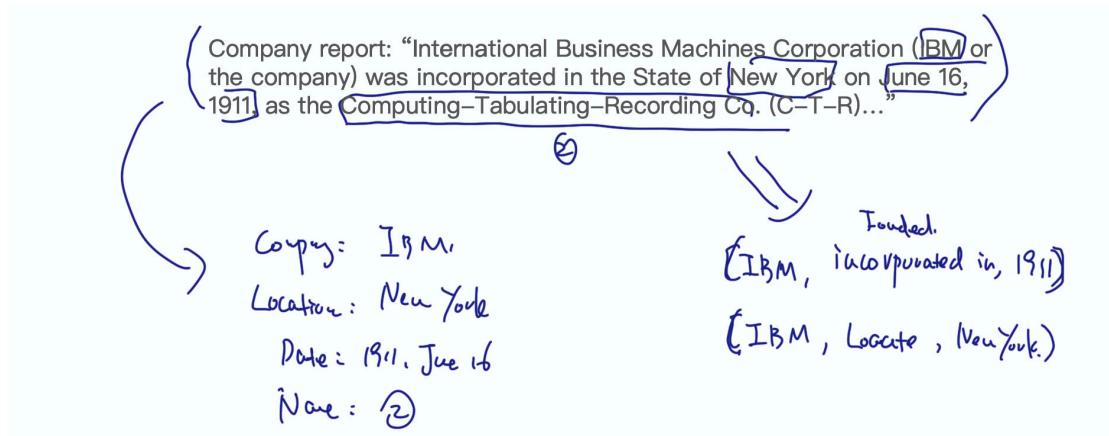
问答系统：

不需要筛选

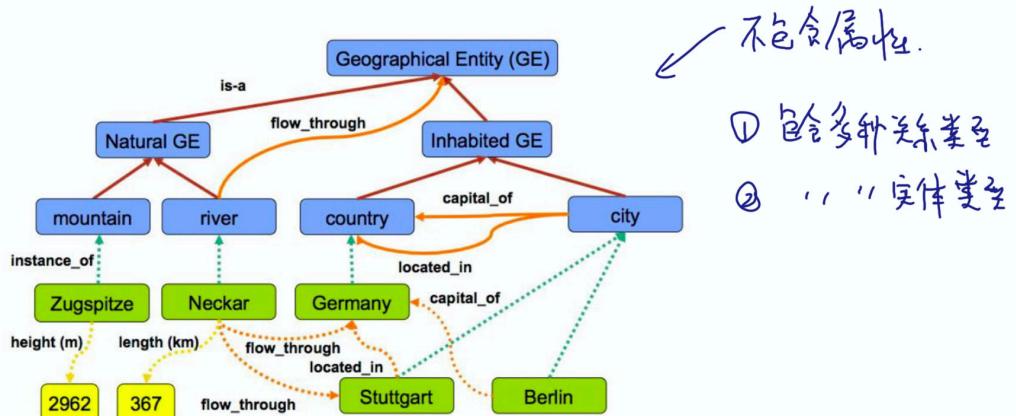
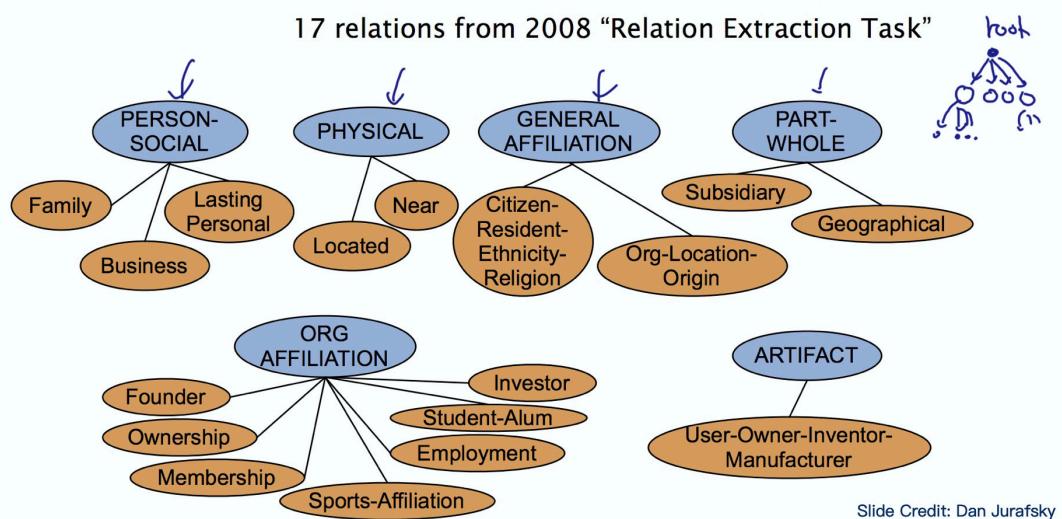


1.5 examples, 三元组





1.6 automatic content extraction(ACE)



2. Named Entity Recognition

命名实体识别(Named Entity Recognition, 简称 NER), 又称作“专名识别”, 是指识别文本中 具有特定意义的实体, 主要包括人名、地名、机构名、专有名词等。

2.1 cases

UIR Case 1: Chat bot → AIUL (AIchat 2004)
→ 规则 (规则)



Case 2: Extract from News

斯坦福全球AI报告：人才需求两年暴增35倍，中国机器人部署量涨500%。刚刚，斯坦福全球AI报告正式发布。从去年开始，斯坦福大学主导、来自MIT、OpenAI、哈佛、麦肯锡等机构的多位专家教授，组建了一个小组，每年发布AI index年度报告，全面追踪人工智能的发展现状和趋势。“我们用硬数据说话。”报告的负责人、斯坦福大学教授、前任谷歌首席科学家Yoav Shoham谈到这份最新的报告时表示。今年的报告，从学术、工业、开源、政府等方面详细介绍了人工智能发展的现状，并且记录了计算机视觉、自然语言理解等领域的技术进展。

产品名：AI Index
组织：斯坦福，MIT，OpenAI，哈佛
公司：麦肯锡，谷歌
人物：Yoav Shoham

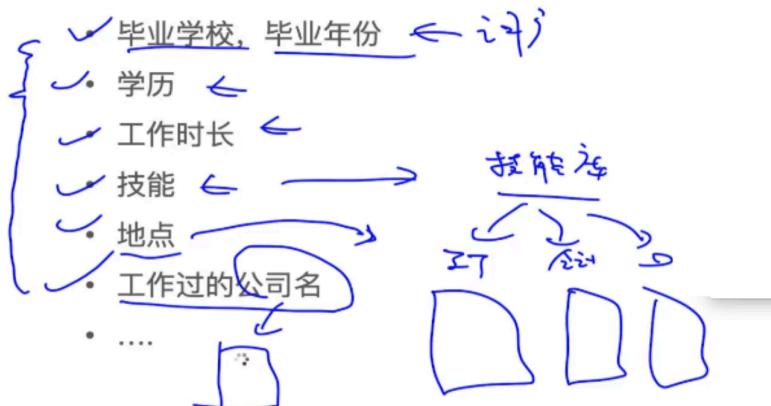
English Toolkits

NLTK NE、Spacy、Stanford Parser

Chinese Toolkits

HanNLP、HIT NLP、Fudan NLP

2.2 Resume Analysis



定义我们关心、需要的实体类别

3. create NER recognizer

定义实体种类、准备训练数据、训练 NER

见 jupyter11

3.1 evaluate

Precision/recall

F1-score

3.2 methods

- 利用规则(比如正则)
- 投票模型(Majority Voting)
- 利用分类模型
- 非时序模型:逻辑回归, SVM...
- 时序模型: HMM, CRF, LSTM-CRF

3.3 rule-based approach

- 美国电话:
- $(?:\([0-9]{3}\)\)?[0-9]{3}[-.?[0-9]{4}$
- 利用已经定义好的词典:
 - if token.contains(word) and word in XXX (词库)

3.4 majority voting

统计每个单词的实体类型，记录针对每个单词，概率最大的实体类型。

4. 特征工程 feature engineering

提取每个单词最简单的特征，比如单词长度

e.g. 随即森林见 jupyter

4.1 Feature Engineering for Supervised Learning

例如：

The professor **Colin** proposed a model for NER in 1999

我们可以获取哪些 feature ?

1) bag-of-word features

- 当前词：colin
- 前后词：professor, proposed
- 前前，后后词
- bi-gram
- trigram...

2) 词性

- 当前词：名词
- 前后词：名词、动词
- 前前后后.....
- ngram

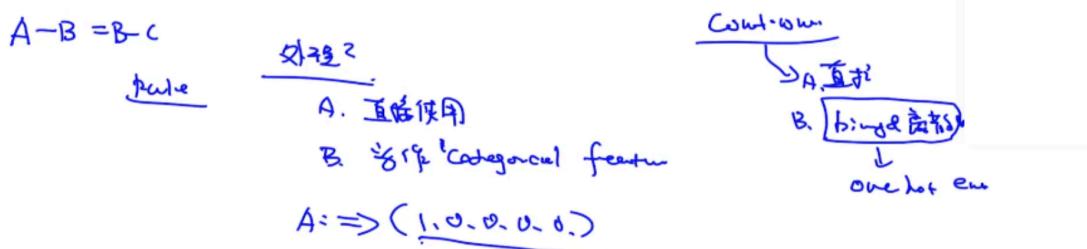
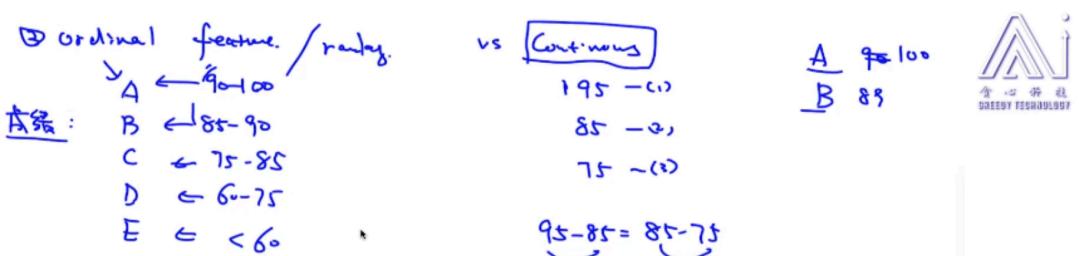
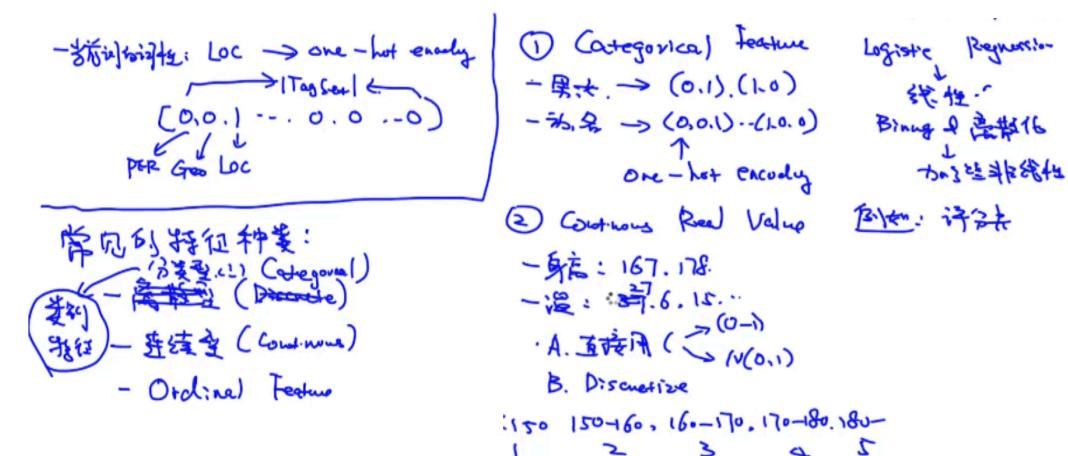
3) 前缀后缀

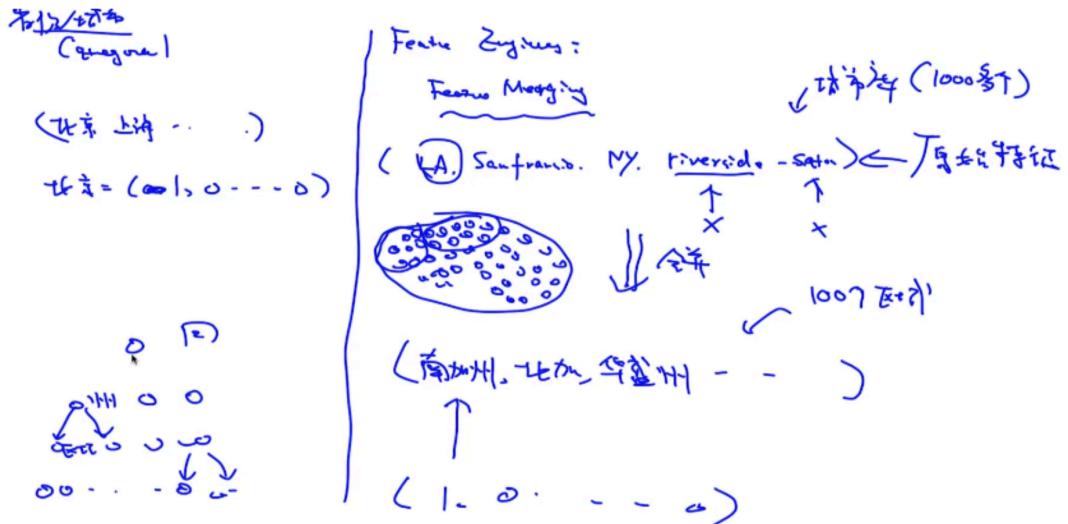
- 当前词：co, in
- 前后词.....

4) 当前词的特性

词长、包含多少个大写字母、是否大写开头、是否含有-、前面词是否包含大写、是否包含数字.....

4.2 Feature Encoding





5. relation

5.1 ontological relation

- IS-A (hypernym relation)
- instance-of

5.2 Open Source Knowledge Base

FreeBase
WorldNet
Yago
Dpedia
KnowledgeVault

5.2 relation extraction



5.3 rule-based method 基于规则的抽取

e.g. Extracting is-a

X	Y
apple	fruit
banana	fruit
civic	car

规则集：

- X (is a) Y
- Y (such as) X
- Y includes X
- Y, especially X
- X or other Y

文章：

- ... [apple] (is a) [fruit] ...
- ... [fruit] (such as) [banana] ...
- ... [fruit] (includes) [apple] ...
- ... [car] (such as) [civic] ...

规则集：人工定义的；工作的核心：定义规则集

可以对规则加限制：

X	Y
apple	fruit
banana	fruit
civic	car

规则集：

X such as Y \Rightarrow X [Fruit] is a Y

加实体限制后规则集

- ① 提升模型的准确率
- ② 只返回想要的结果

提升准确率；只返回想要的结果

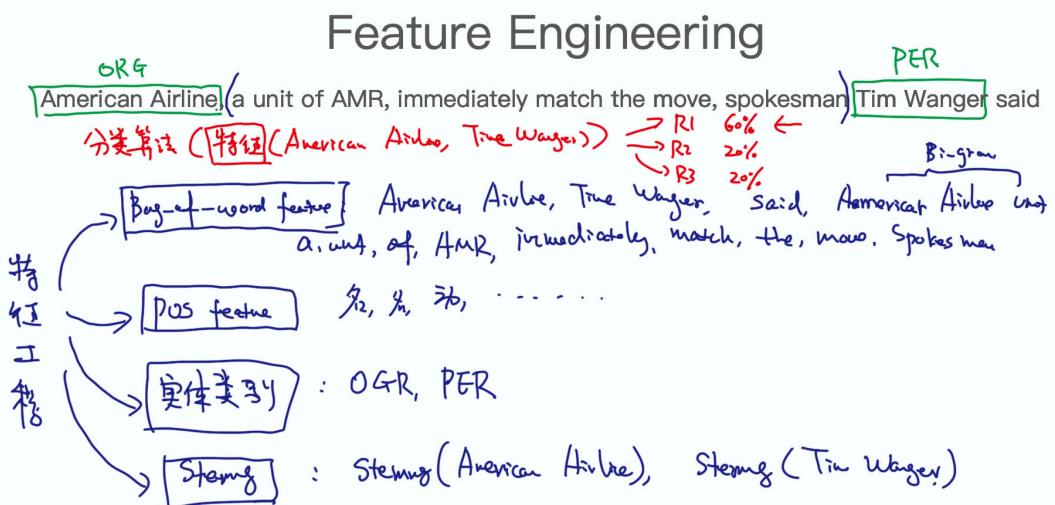
优点：准确；不需要训练数据

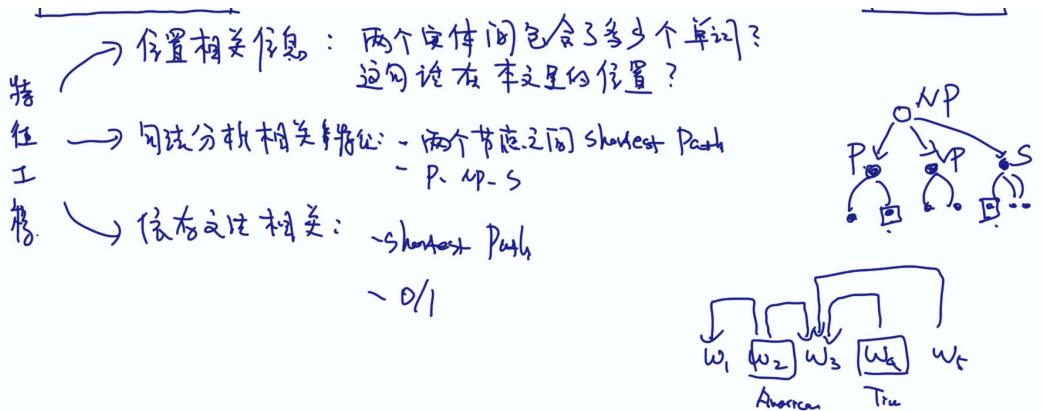
缺点：low recall rate；成本高；规则难设计

5.4 基于监督学习的方法

- 1) 定义关系类型
- 2) 定义实体类型
- 3) 训练数据准备
 - 实体标记好
 - 实体之间的关系

Example :





Classification Model

