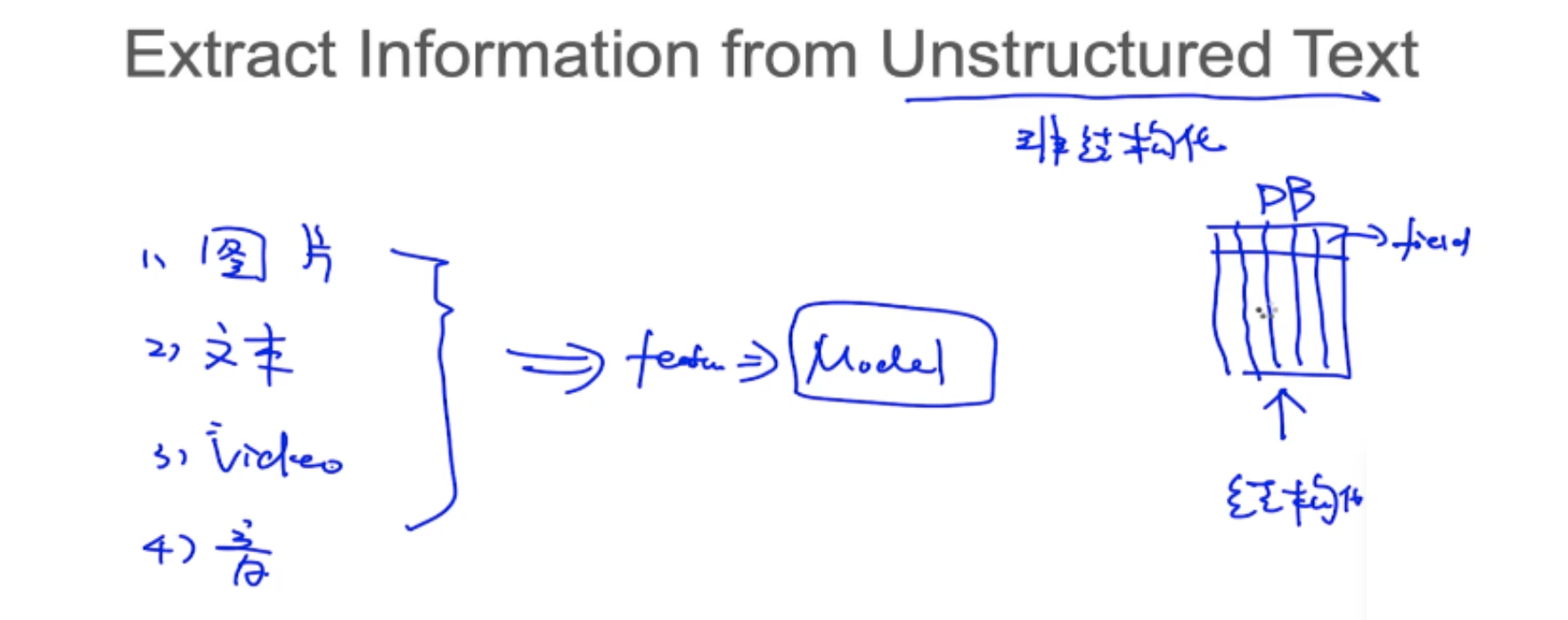
**1. Information extraction (IE)**

图片包含 屏幕截图, 文字

描述已自动生成



**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**

图片包含 文字

描述已自动生成

**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

图片包含 屏幕截图

描述已自动生成

**1.3 more applications**

知识库的搭建；

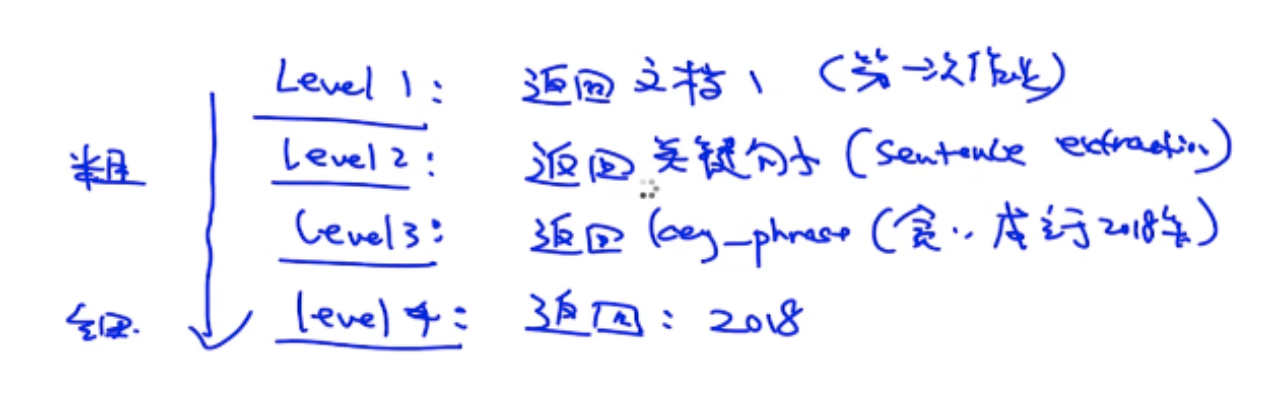
google scholar；（需要实体统一）

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



**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**

图片包含 屏幕截图

描述已自动生成

图片包含 文字, 收据

描述已自动生成

English Toolkits

NLTK NE、Spacy、Stanford Parser

Chinese Toolkits

HanNLP、HIT NLP、Fudan NLP

**2.2 Resume Analysis**

图片包含 文字

描述已自动生成

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)

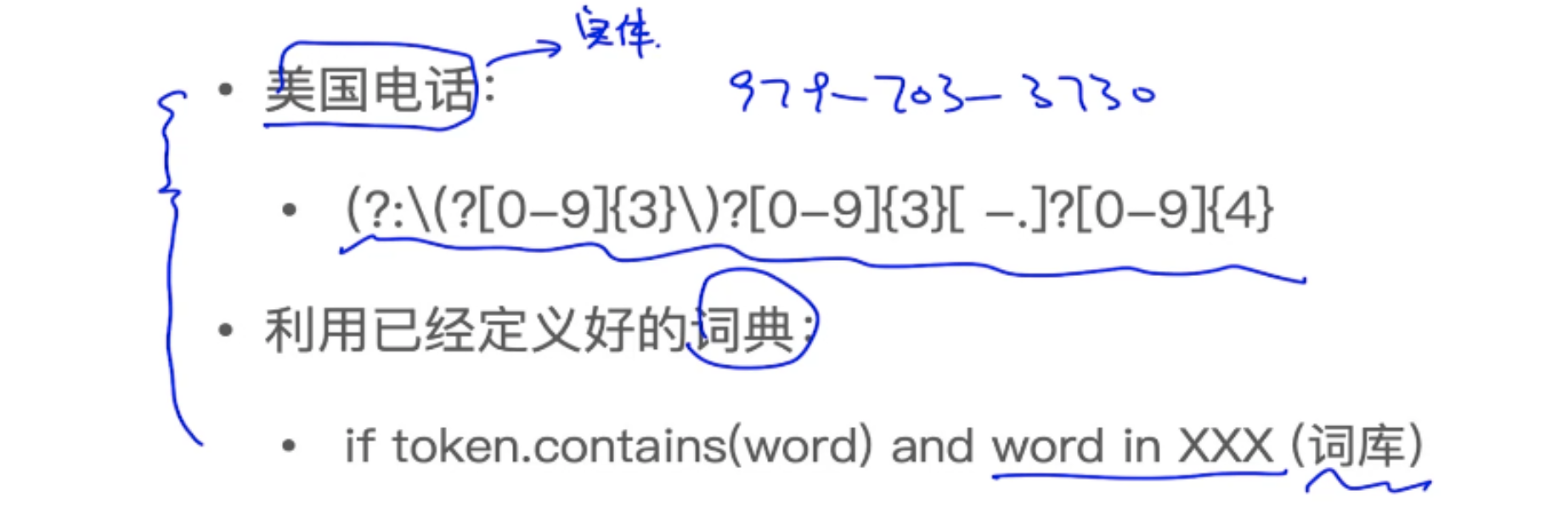
• 投票模型(Majority Voting)

• 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



3.4 majority voting

统计每个单词的实体类型，记录针对于每个单词，概率最大的实体类型。