自然语言处理

第2讲：自然语言处理概述

刘洋

内容提要

基本概念

典型任务

发展历史

相关资源

语言

- 语言是个体之间由于沟通需要而制定的指令。

汉语	语言	意大利语	linguaggio	蒙古语	Хэл
英语	Language	日语	言語	俄语	Язык
法语	Langue	韩语	언어	希伯来语	שפה
德语	Sprache	马来语	Bahasa	阿拉伯语	لغة
西班牙语	Idioma	泰语	ภาษา	卢旺达语	Ururimi
葡萄牙语	Língua	越南语	Ngôn ngữ	斯瓦希里语	Lugha

世界上具有丰富多样的各种语言

自然语言

- 自然语言：人类之间用于沟通交流的语言。

望庐山瀑布

（唐）李白

日照香炉生紫烟，
遥看瀑布挂前川。
飞流直下三千尺，
疑是银河落九天。

自然语言

```
#include <stdio.h>

int main()
{
    int x = 1;
    int y = 2;
    printf(“%d”, x + y);

    return 0;
}
```

机器语言

自然语言的特点

- 线性：自然语言呈现为一种线性的符号序列。

汉语　王教授昨天在北京做了一个演讲

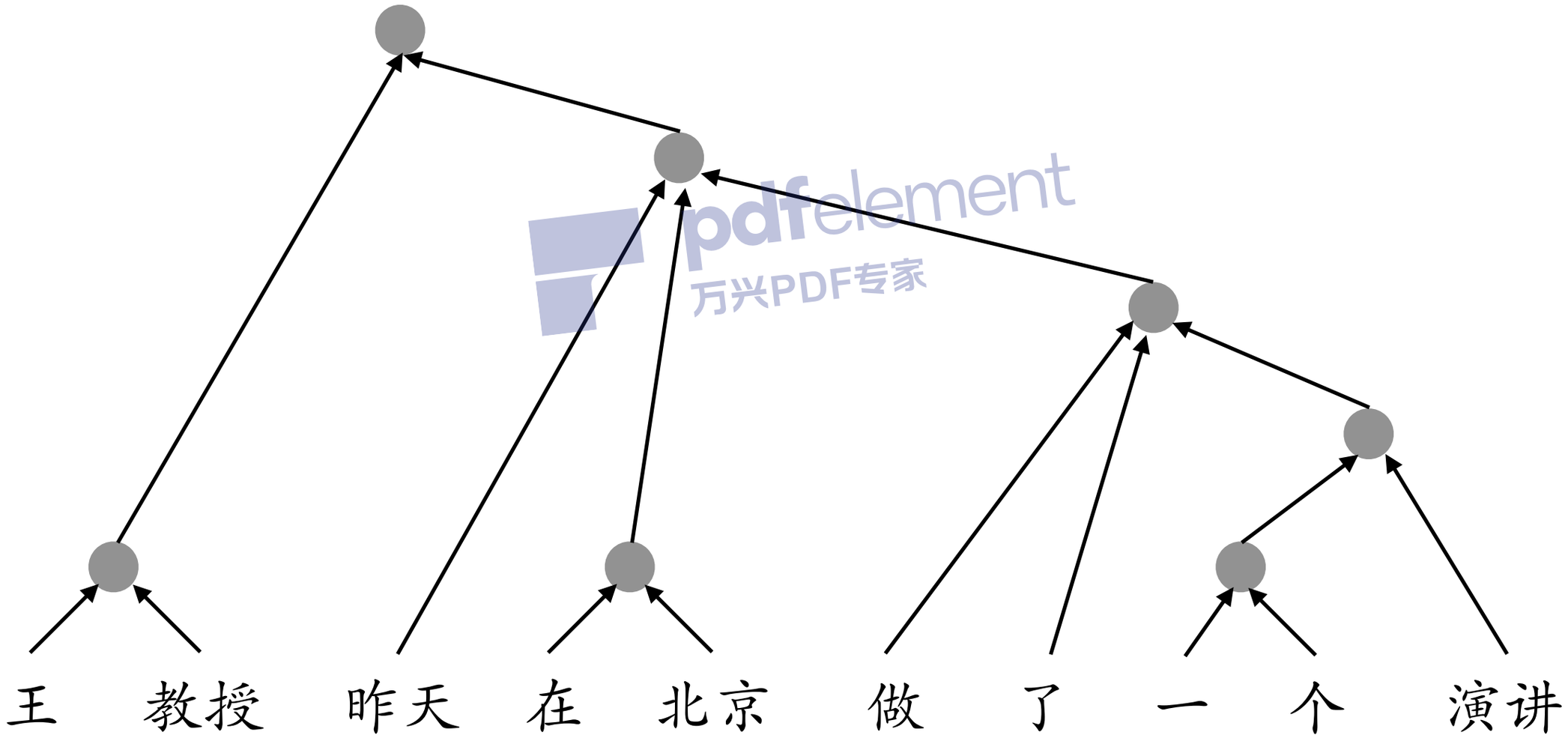
英语　Professor Wang gave a speech in Beijing yesterday

阿拉伯语　ألقى الأستاذ وانغ خطابا في بكين أمس

汉语和英语的语序是从左向右，而阿拉伯语的语序是从右向左。

自然语言的特点

- 层次性：自然语言内部存在层次结构。



自然语言的特点

- 歧义性：同一个自然语言句子存在多种不同的理解。

南京市长江大桥

咬死猎人的狗

他有两本鲁迅先生的书

The boy saw a girl with a telescope

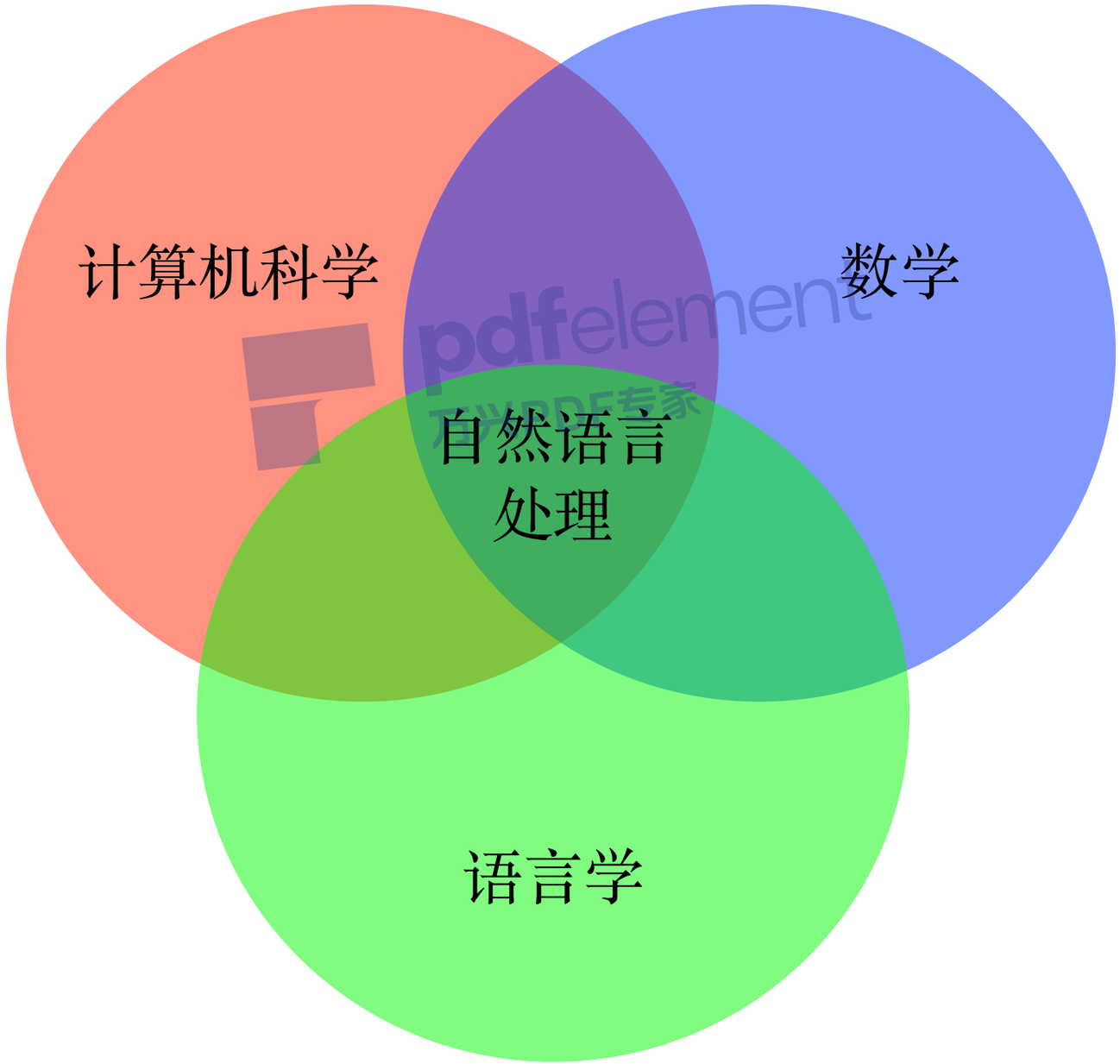
自然语言的特点

- 演化性：自然语言随着时代不断演化。

2000	I服了You	2010	神马都是浮云
2001	猪队友	2011	hold不住
2002	菜鸟	2012	十动然拒
2003	表酱紫	2013	逆袭
2004	偶稀饭	2014	也是醉了
2005	互粉	2015	重要的事情说三遍
2006	草根	2016	吃瓜群众
2007	Orz	2017	尬聊
2008	打酱油	2018	真香
2009	杯具	2019	柠檬精

自然语言处理

- 自然语言是一门利用计算机自动处理人类语言的交叉学科。



内容提要

基本概念

典型任务

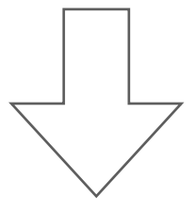
发展历史

相关资源

中文分词

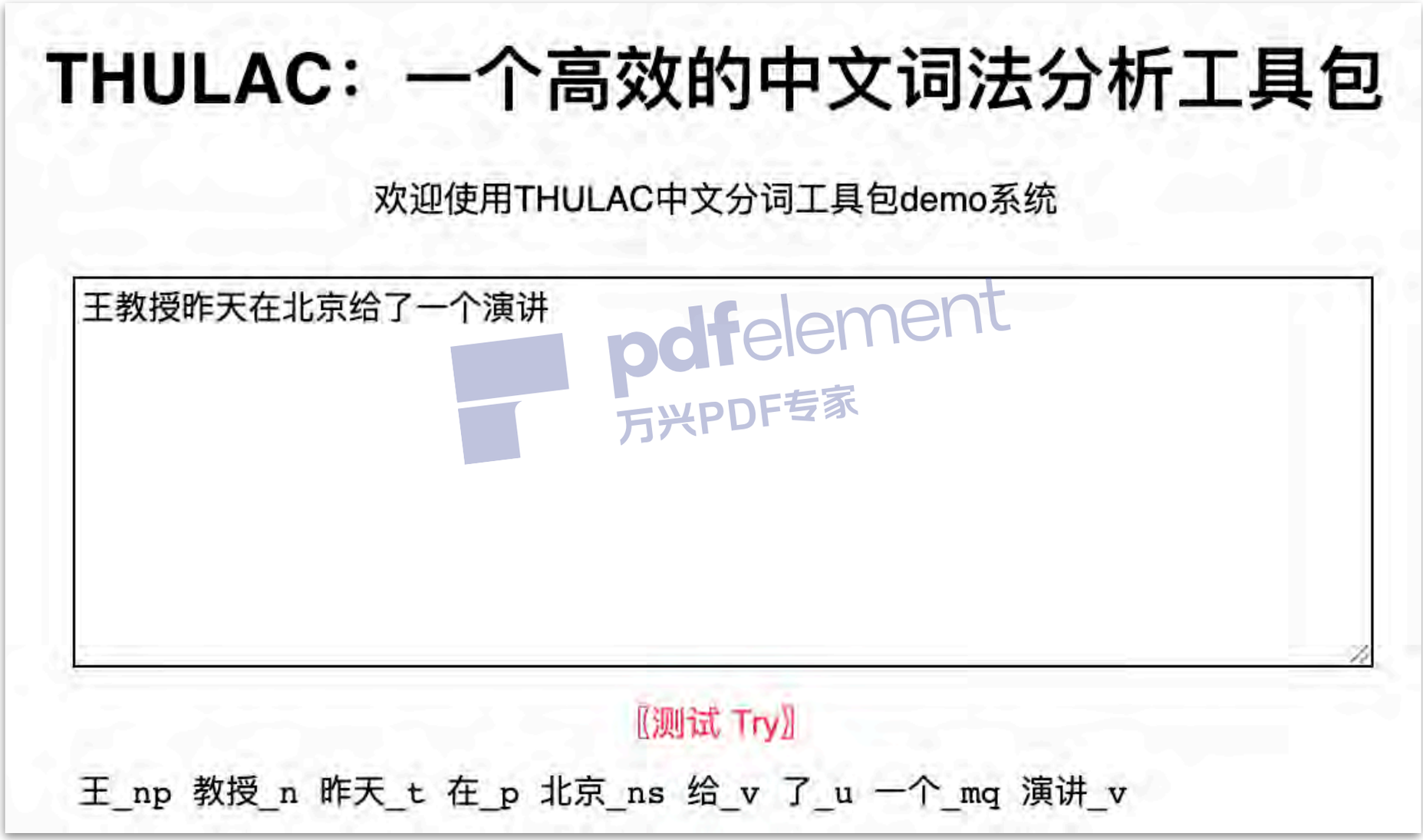
- 输入：一段不带空格的汉语文本。
- 输出：以空格隔开词语的汉语文本。

王教授昨天在北京做了一个演讲



王 教授 昨天 在 北京 做 了 一 个 演 讲

中文分词

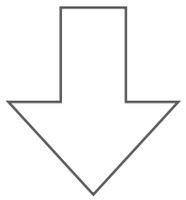


http://thulac.thunlp.org/demo

词性标注

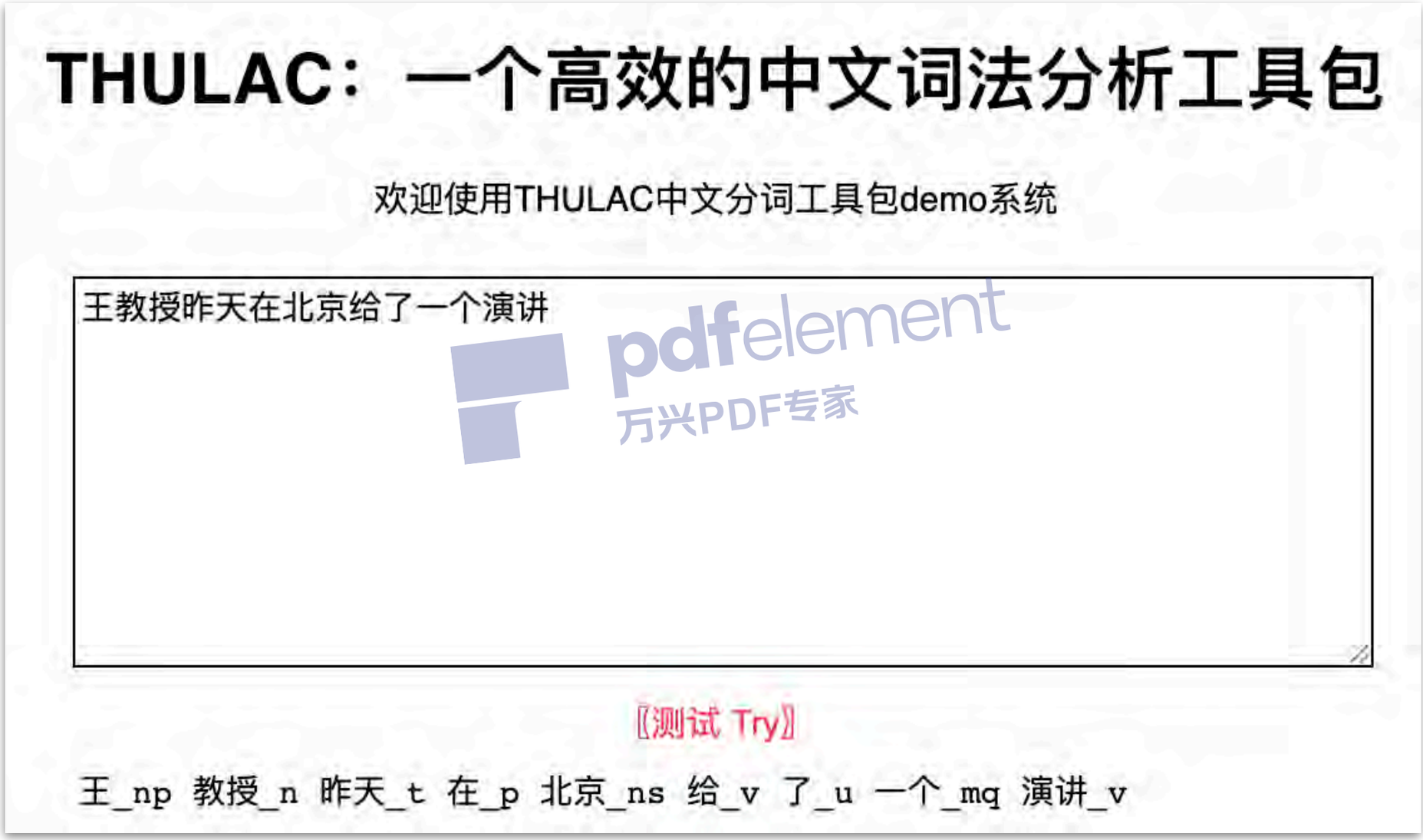
- 输入：给定一个词语的序列。
- 输出：输出一个对应的词性的序列。

王 教授 昨天 在 北京 做了 一个 演讲



np n t p ns v u mq n

词性标注

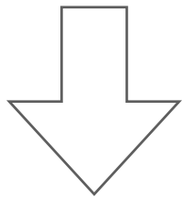


http://thulac.thunlp.org/demo

文本分类

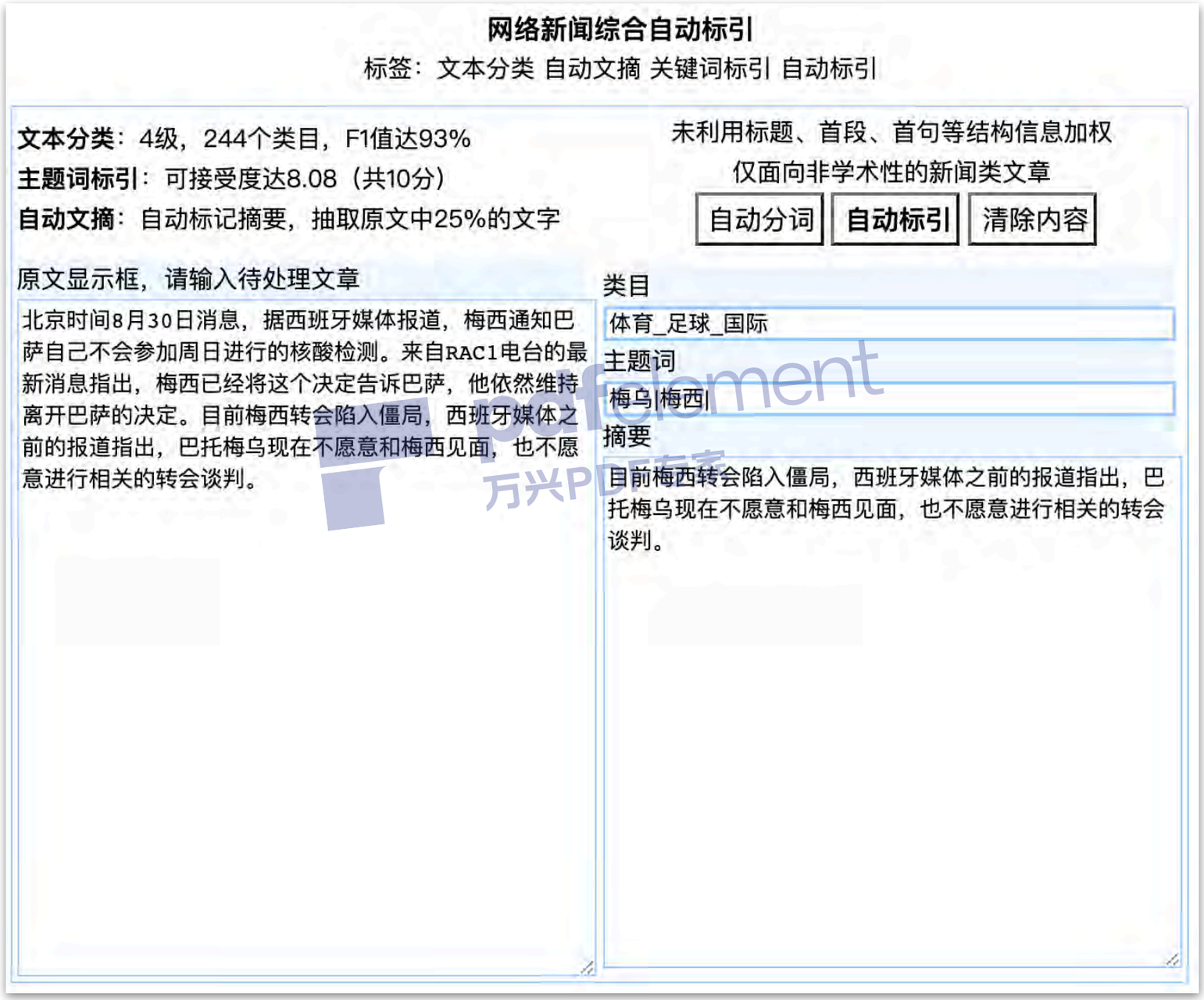
- 输入：一段文本
- 输出：该文本的类别。

北京时间8月30日消息，据西班牙媒体报道，梅西通知巴萨自己不会参加周日进行的核酸检测。来自RAC1电台的最新消息指出，梅西已经将这个决定告诉巴萨，他依然维持离开巴萨的决定。目前梅西转会陷入僵局，西班牙媒体之前的报道指出，巴托梅乌现在不愿意和梅西见面，也不愿意进行相关的转会谈判。



政治	军事	经济	教育	体育	科技	旅游	医疗

文本分类

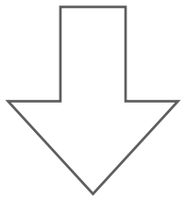


http://www.languagetech.cn/nlp/index/demo_class.aspx

语言模型

- 输入：给定一个词语序列
- 输出：预测下一个词

New York is the largest city in the United _____



…	country	Nations	State	States	books	trees	…

语言模型

Language Modeling

Language modeling is the task of determining the probability of a given sequence of words occurring in a sentence.

This demonstration uses the public 345M parameter OpenAI GPT-2 language model to generate sentences.

Provide some initial text, and the model will generate a list of the most-likely next words. You can click on one of those candidate words to choose it and continue, or you can keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:

New York City is the largest city in the United

Predictions:

99.0% States
0.6% Kingdom
0.1% Nations
0.1% S
0.1% State
← Undo

https://demo.allennlp.org/next-token-lm

语言模型

Masked Language Modeling

Masked language modeling is a fill-in-the-blank task, where a model uses the context words surrounding a [MASK] token to try to predict what the [MASK] word should be.

The model shown here is BERT, the first large transformer to be trained on this task. Enter text with one or more "[MASK]" tokens and the model will generate the most likely substitution for each.

Sentence:

The doctor ran to the [MASK] room to see [MASK] patient.

Mask 1 Predictions:		Mask 2 Predictions:	
31.1%	waiting	42.6%	his
11.6%	emergency	39.4%	the
6.2%	operating	5.2%	another
5.9%	next	5.1%	her
3.5%	other	4.4%	a

https://demo.allennlp.org/masked-lm

语法改错

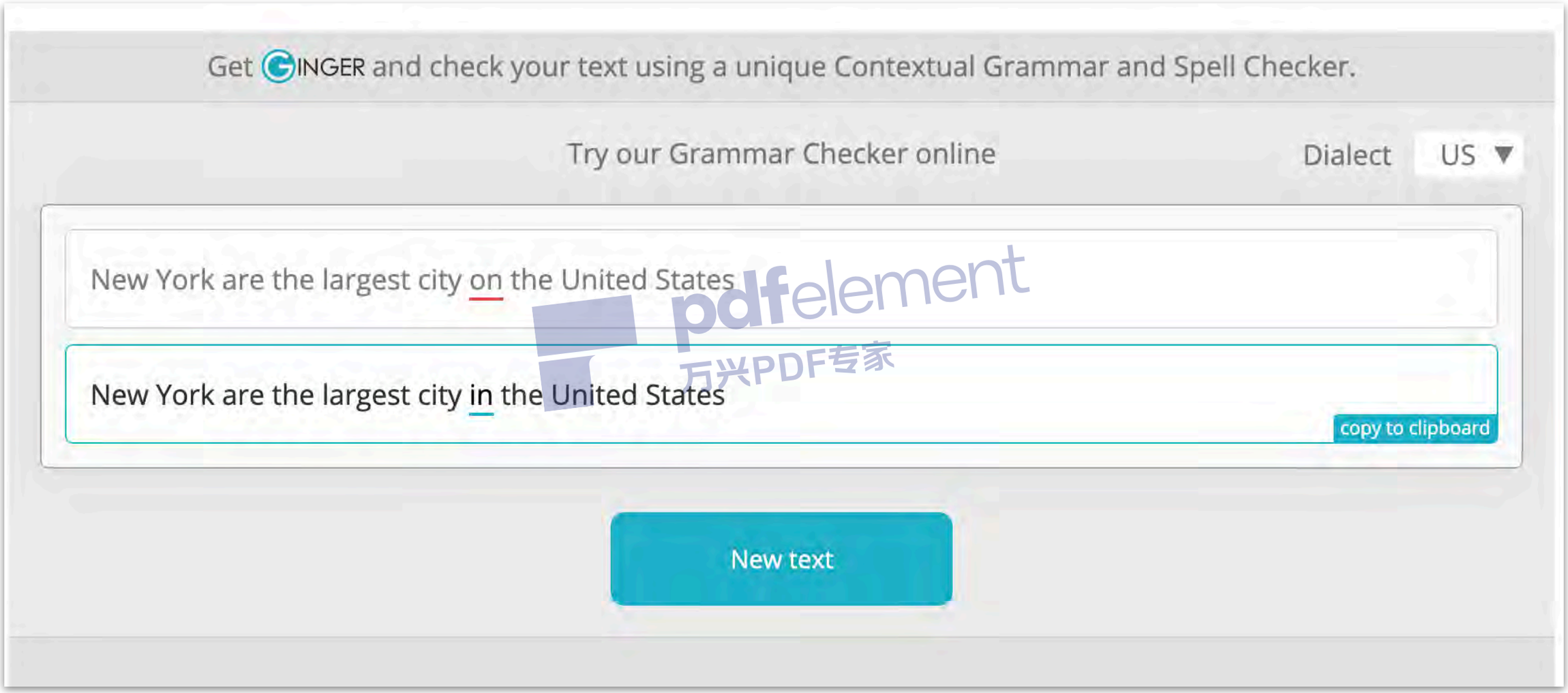
- 输入：一段可能包含语法错误的文本。
- 输出：识别出文本中的语法错误并进行修改。

New York are the largest city on the United States

New York are the largest city on the United States

New York is the largest city in the United States

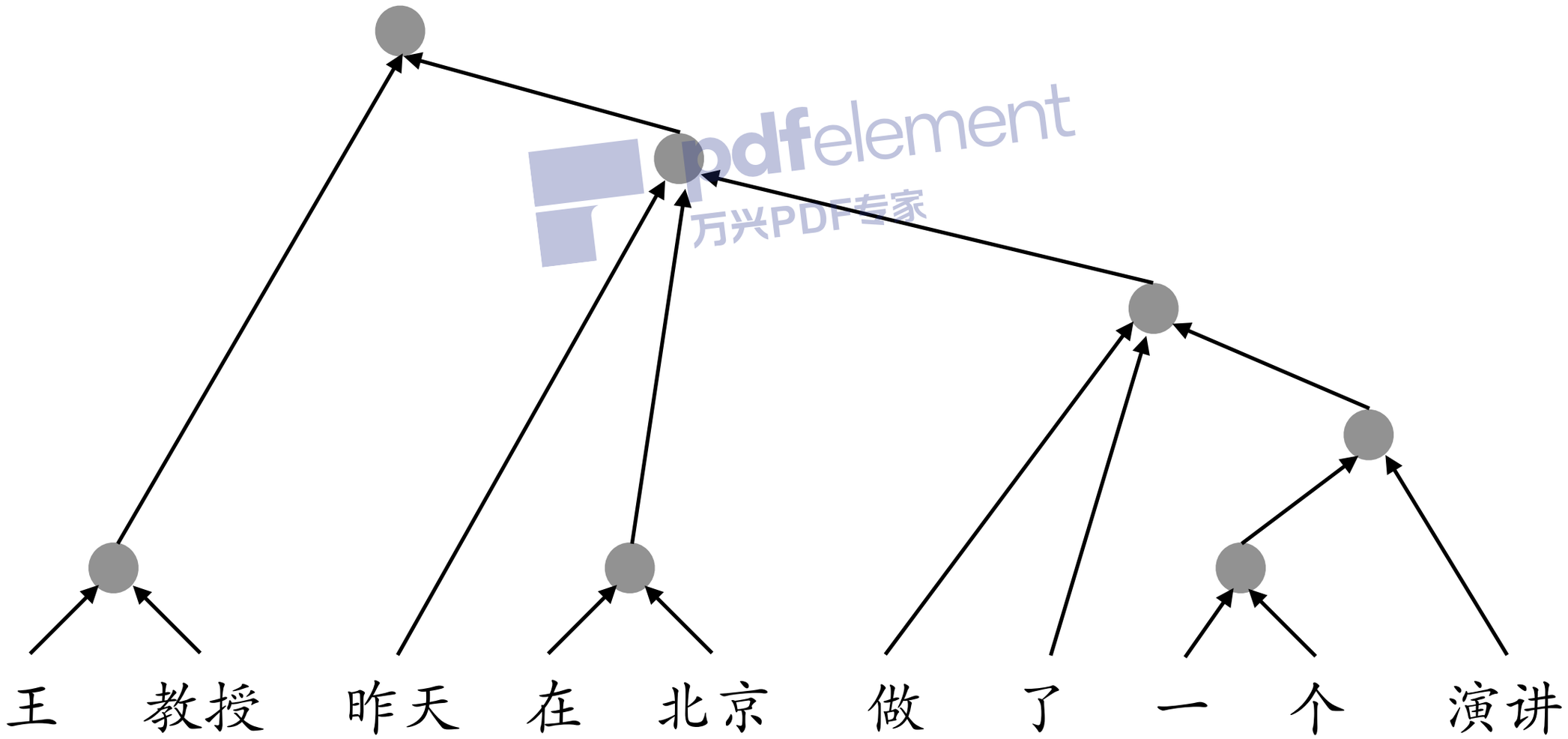
语法改错



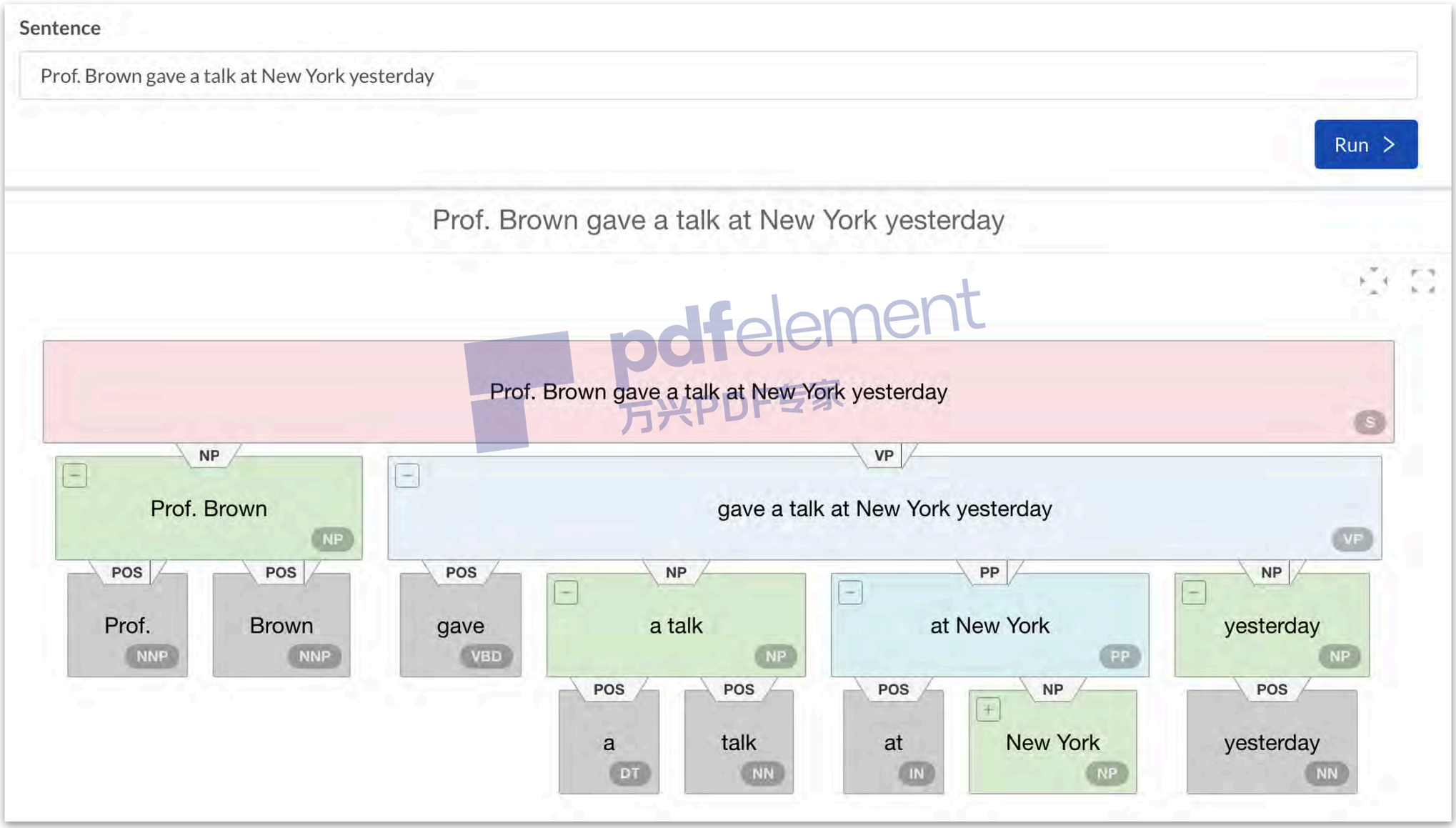
https://www.gingersoftware.com/grammarcheck

句法分析

- 输入：一个自然语言句子
- 输出：句子的句法结构（短语结构或依存结构）

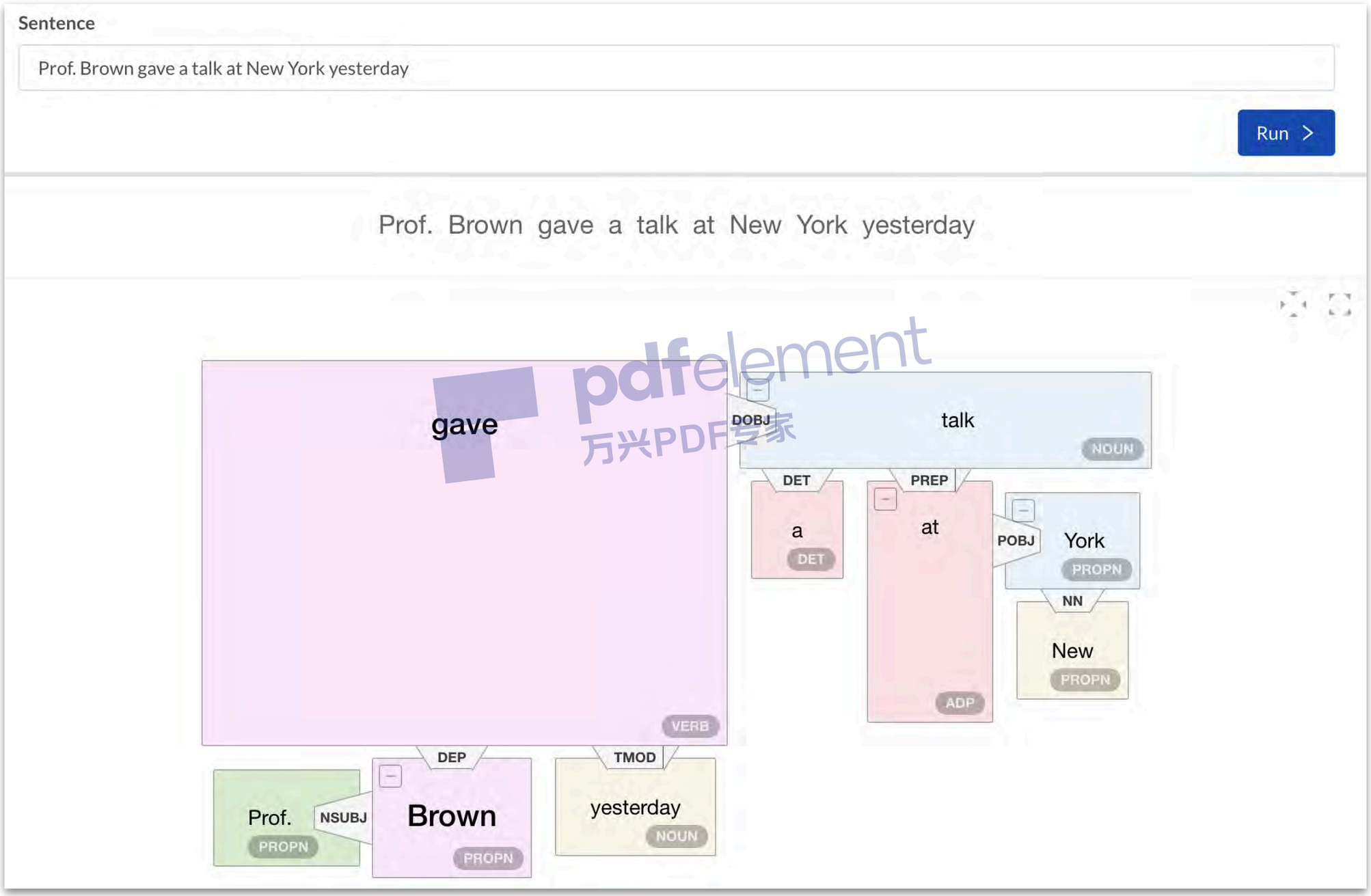


句法分析



https://demo.allennlp.org/constituency-parsing

句法分析

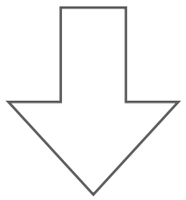


https://demo.allennlp.org/dependency-parsing

拼音输入法

- 输入：拼音符号的序列
- 输出：汉字序列

wangjiaoshouzuotianzaibeijingkaihui



王教授昨天在北京开会

拼音输入法

https://www.google.com/intl/zh-CN/inputtools/try/

情感分析

- 输入：一段自然语言文本。
- 输出：情感的类别（如正面、中性、负面）

> 陈可辛都拍不出这么做作的电影，明明煽情得要死还要搞出一副纪录片的质感，后半段整个崩掉，生怕观众get不到还要派黄晓明出来说教点题。有几个战斗场面本来拍得还行，但都抵不过杜淳那出戏的假方言。

正面	中性	负面

情感分析

Sentiment Analysis

Sentiment Analysis is the task of interpreting and classifying emotions (positive or negative) in the input text.

Model

GloVe-LSTM

This model uses GloVe embeddings and is trained on the binary classification setting of the Stanford Sentiment Treebank. It achieves about 87% on the test set.

Demo Usage

Enter text or Choose an example...

Input

I really hate this movie

Run >

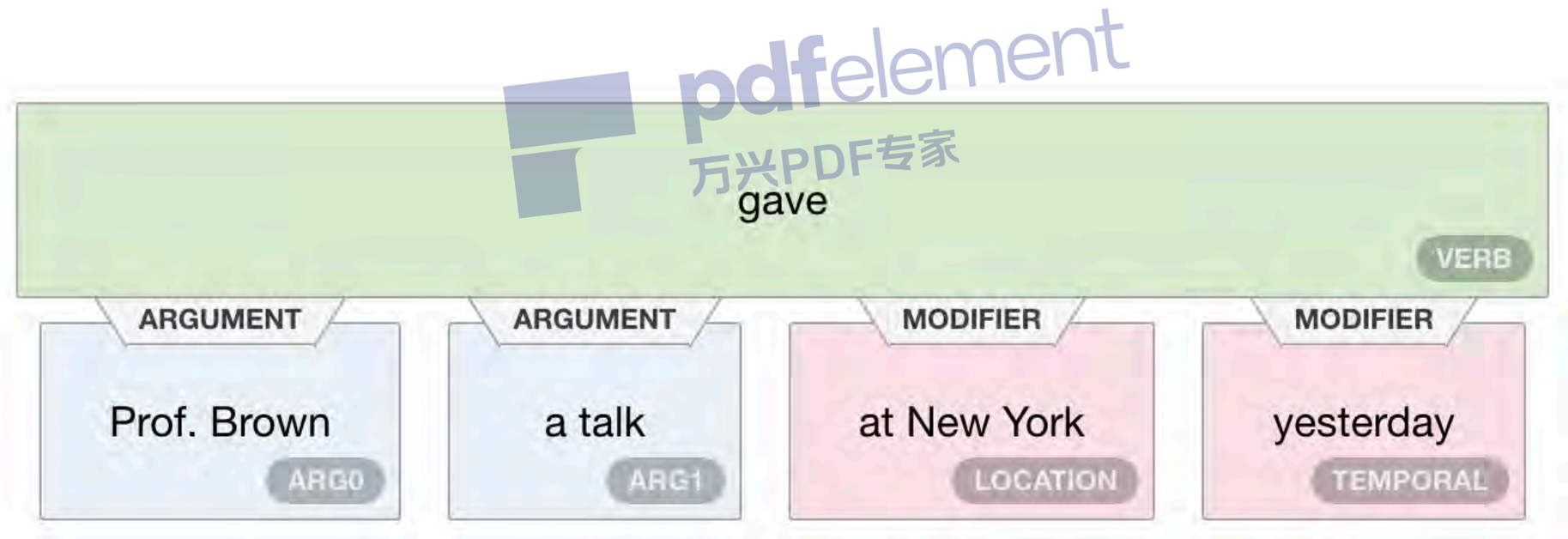
Answer

The model thinks the sentence is Negative. (64.9%)

https://demo.allennlp.org/sentiment-analysis

语义角色标注

- 输入：一个自然语言句子。
- 输出：标出句子的谓语及相关语义角色。



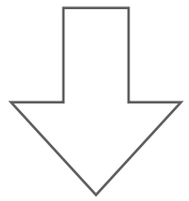
语义角色标注

https://demo.allennlp.org/semantic-role-labeling

语义分析

- 输入：一个自然语言处理句子
- 输出：该句子的语义表示形式

Every boy likes a star



$$\forall x(\text{boy}(x) \rightarrow \exists y(\text{human}(y) \wedge \text{pop}(y) \wedge \text{like}(x, y)))$$

语义分析

Utterance

Show me the flights from New York to Los Angeles

Run >

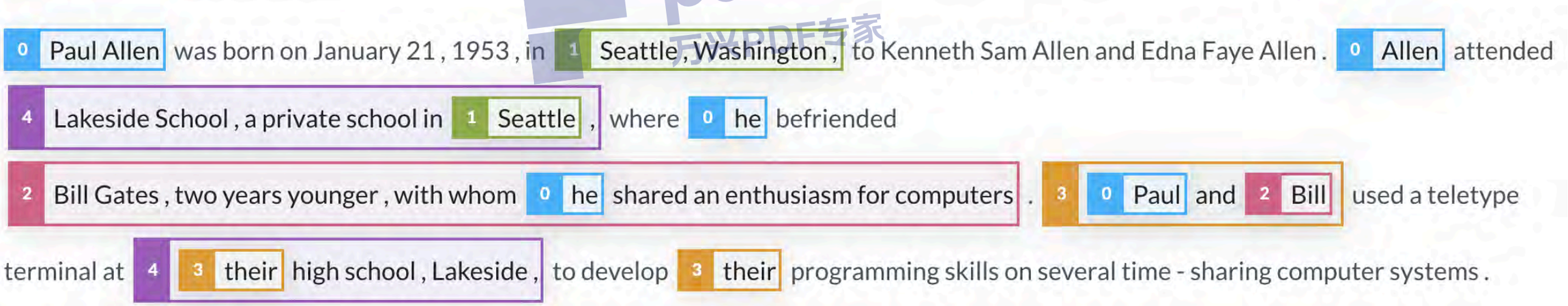
SQL Query

```
(SELECT DISTINCT flight . flight_id
 FROM flight
 WHERE (flight . from_airport IN
         (SELECT airport_service . airport_code
          FROM airport_service
          WHERE airport_service . city_code IN
              (SELECT city . city_code
               FROM city
               WHERE city . city_name = 'NEW YORK' ) )
       AND flight . to_airport IN
         (SELECT airport_service . airport_code
          FROM airport_service
          WHERE airport_service . city_code IN
              (SELECT city . city_code
               FROM city
               WHERE city . city_name = 'LOS ANGELES' ) )) ) ;
```


https://demo.allennlp.org/atis-parser

指代消解

- 输入：一段自然语言文本
- 输出：该文本中代词所指向的名词



指代消解

Document

Paul Allen was born on January 21, 1953, in Seattle, Washington, to Kenneth Sam Allen and Edna Faye Allen. Allen attended Lakeside School, a private school in Seattle, where he befriended Bill Gates, two years younger, with whom he shared an enthusiasm for computers. Paul and Bill used a teletype terminal at their high school, Lakeside, to develop their programming skills on several time-sharing computer systems.

Run >

0 Paul Allen was born on January 21 , 1953 , in 1 Seattle , Washington , to Kenneth Sam Allen and Edna Faye Allen . 0 Allen attended 4 Lakeside School , a private school in 1 Seattle , where 0 he befriended 2 Bill Gates , two years younger , with whom 0 he shared an enthusiasm for computers . 3 0 Paul and 2 Bill used a teletype terminal at 4 3 their high school , Lakeside , to develop 3 their programming skills on several time - sharing computer systems .

https://demo.allennlp.org/coreference-resolution

机器翻译

- 输入：一段源语言文本
- 输出：一段目标语言文本

王 教授 昨天 在 北京 做 了 一个 演讲

Prof. Wang made a speech at Beijing yesterday

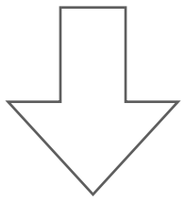
机器翻译

https://fanyi.sogou.com/

文本摘要

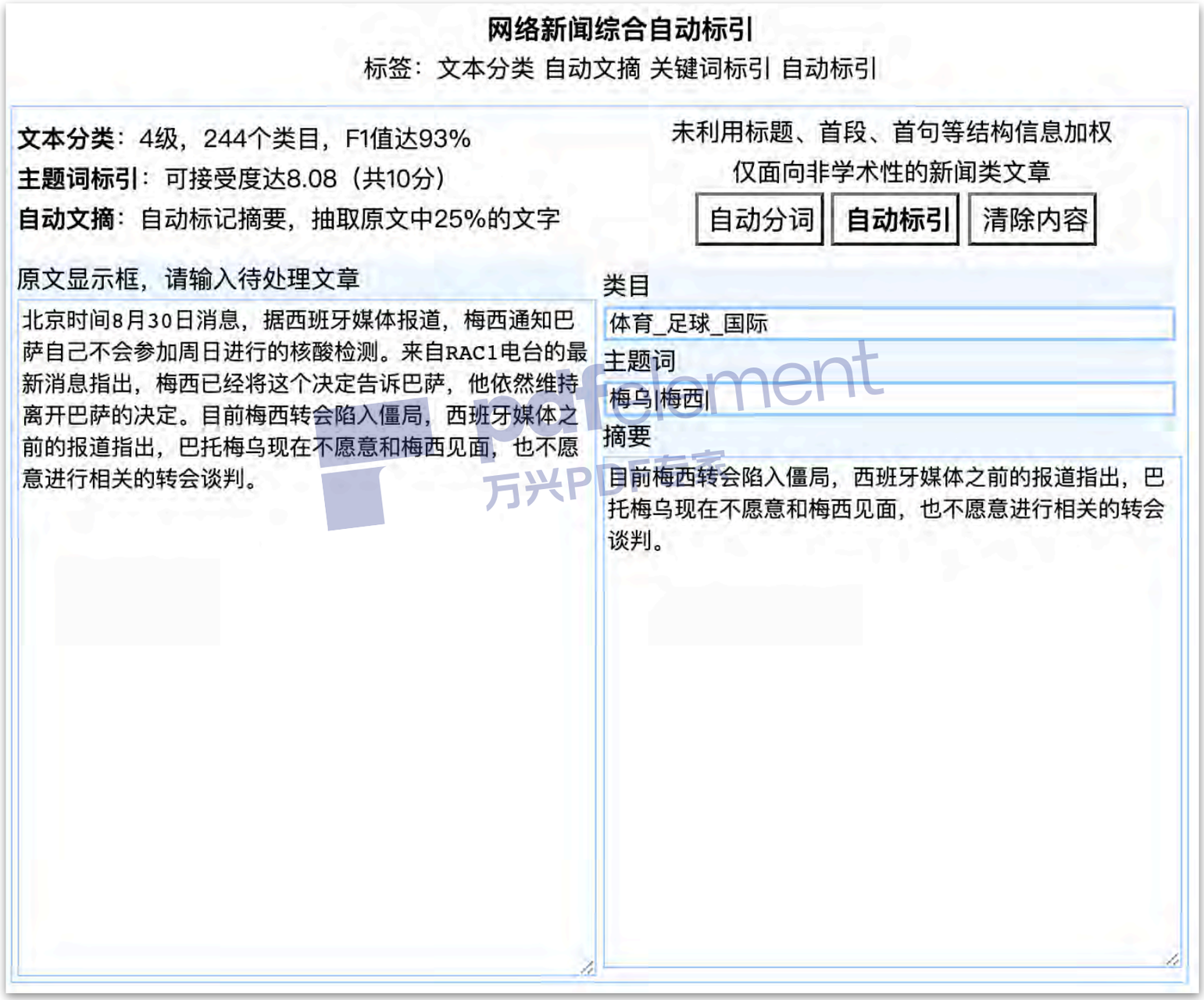
- 输入：一段自然语言长文本。
- 输出：一段能概括长文本核心意思的短文本。

北京时间8月30日消息，据西班牙媒体报道，梅西通知巴萨自己不会参加周日进行的核酸检测。来自RAC1电台的最新消息指出，梅西已经将这个决定告诉巴萨，他依然维持离开巴萨的决定。目前梅西转会陷入僵局，西班牙媒体之前的报道指出，巴托梅乌现在不愿意和梅西见面，也不愿意进行相关的转会谈判。



梅西执意离开巴萨，拒绝参加周日的核酸检测。

文本摘要

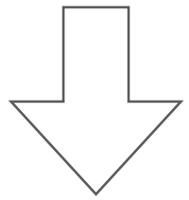


http://www.languagetech.cn/nlp/index/demo_class.aspx

对联生成

- 输入：对联的上联
- 输出：对联的下联以及横批

松 叶 竹 叶 叶 叶 翠



秋 声 雁 声 声 声 寒

对联生成

http://duilian.msra.cn/