计算语言学 第一次作业 中文问答系统测试

请同学们利用如下给定的若干个代表性中文问答系统，设计若干个问题提交给该系统，得到系统输出的结果，并对该结果是否正确予以明确判断。

作业内容

问题设计

为测试问答系统性能，每位同学设计20个常识事实性问题。并提供这些问题的标准答案，答案需唯一（What, When, Where, Who, Which）而非开放性问题（Why, How）。为了保证问题的难度，每个问题应保证在系统测试中，至少有2个系统在该问题的至少一个问法上答错。

必须保证同学给出的标准答案的正确性。否则，作业成绩将按照答案错误数量扣分。为了保证正确性，每一个问题同学们需提供两个公开网页作为其答案的佐证，且需要在上交的文件中给出网址和答案对应段落。这两个网页应尽量来自权威来源，如维基百科、百度百科等。

问题分为如下5个类别，请同学们对每个类别提4个问题。

1. “是什么（What）”型问题，如“光合作用的产物是什么？”
2. “什么时候（When）”型问题，如“中华人民共和国是什么时候建立的？”
3. “什么地点（Where）”型问题，如“爱因斯坦是在哪里出生的？”
4. “是谁（Who）型问题”，如“中国历史上第一个皇帝是谁？”
5. “哪一个（Which）型问题”，如“哪座山是中国最高的？”“二氧化碳和氧气中哪一个能令澄清石灰水变浑浊？”

每个问题必须提供两种含义相同的问法，以测试问答系统的鲁棒性。比如问题“中国历史上第一个皇帝是谁？”的另一种问法可为“谁是中国第一个称帝的人？”，问题“爱因斯坦是在哪里出生的？”的另一种问法可为“爱因斯坦的出生地是哪儿？”。

系统测试及分析

将问题输入给如下中文问答系统中进行测试，得到系统回答。系统的输出与标准答案意思一致即可认为正确，不必文字上与标准答案完全一样。可测试的代表性问答系统包括：

1. 百度文心：https://wenxin.baidu.com/moduleApi/ernie3 中的自由问答
2. 智源悟道：https://open.wudaoai.com/ 需要进行注册，注册后在“体验中心”中选择“智能问答”即可
3. 源1.0：需要先访问https://air.inspur.com/home，注册用户并点击首页的“申请API”，填写相关信息后等待批准通过。批准通过后访问http://221.194.179.88:11016/?question=中国第一位女皇帝是谁&usrname=lwh&phone=1368888888，将自己的用户名，电话和问题填入其中，即可得到答案。**据测试，源1.0只在工作日批准申请且需要数个工作日，为避免赶不上作业截止日期，请同学们提前申请。**

在得到测试结果后，需要分别对每个问答系统的准确性（答对问题比例）和鲁棒性（两种问法答案一致比例）进行分析并得出高下结论。同时针对所提出的具体问题，对每个问答系统回答不同类别的问题的能力进行具体分析并举例说明。

提交内容

需要将所提不同问法的问题、类别、标准答案，标准答案佐证来源网址及对应段落，以及每个模型针对每个问法的回答填入所给的qa_submission.xlsx中，连同一份pdf格式的实验报告上交至网络学堂。表格中第0个问题为示例，仅用于提供填写范例，不计入问题数目。pdf格式的报告应包含上文要求的对测试结果的分析。