说话人1 00:01  
大家往前坐好吧？人应该没那么多，

说话人2 00:06  
有那么多吗？你。

说话人1 00:09  
不说37个吗？一个，

说话人2 00:14  
ok。

说话人1 00:23  
大家能听到吗？

说话人2 00:27  
能听到吧？

说话人1 00:29  
可以，这是讲话，那个同学把那个门关一下，后面那个门。

说话人2 00:34  
对，

说话人1 00:37  
我们就开始上课，然后我们这个课是叫自然语言处理，

说话人2 00:43  
然后。

说话人1 00:45  
我的名字你在屏幕上可以看到叫乔梁，然后我的邮箱就是 [Bridge@139.com](mailto:Bridge@139.com)，就是乔梁的英文，然后我的手机号13926026874有什么事情中有可能我们也会有些群像微信群或学习通上的一些东西，

说话人2 01:07  
然后。

说话人1 01:08  
我们也可以有什么课程的一些东西可以保持沟通。

首先一下介绍我自己，然后。

说话人2 01:22  
我。

说话人1 01:24  
我本科是学数学的，然后实际上是一直学数学，然后研究生也算是数学是计量经济学的。然后后来博士的话也是应用数学，就是做信息安全密码学这块东西的。

所以。

说话人2 01:43  
我。

说话人1 01:43  
去年大概是10月份就加入我们新华，实际上过去很多年都是一直在产业界做的，

做过还待过很多家公司，第一份工作应该是在国企在中国银行，

说话人2 01:59  
然后。

说话人1 02:00  
后面的很多经历的话都是在。

在外企，就是像Teradata或者SAS，我不知道你们有没有听过这些公司，这都是一般我们说的这种数据行业就比较就比较专业，或者说比较顶尖的一些global的公司，像Teradata是你们有没有学数据仓库，或者说这块东西基本上我们说是国内的做数据平台，数据仓库的黄埔军校大概是这样子，我很多同事在这个方向也是做得非常好， SAS的话它是做我不知道你们有没有学过，你们在学Python是吧？

 Sas实际上是之前的话是很流行的，他是在商业统计做比如像银行的一些风险分析，或者说一些生物统计，医学统计是一个很主流的工具，当然也做过其他的一些平台，我自己之前大概工作了有有20多年就一直做数据挖掘，机器学习人工智能这些东西，但我做的东西是比较杂，像一般的数据平台，数据挖掘，数据治理，还有一些分析的应用，包括后面就做管理做咨询的事情比较多一点。

结合这门课程的话，但我现在研究的话最主要是像AI for data或者说工业数字化的一些东西，大家如果有兴趣，我们也可以一起来做些东西，然后像 NLP相关的经验的话，

最近几年做的比较少，之前做的还做过一段时间，最早我们像所谓的某手机厂商，我等一下有几页 slide给大家介绍。

一个客户资深的项目，它就是一个很大规模的自然语言处理的，就是手机厂商是手机华为，手机刚刚推出来最早的时候。

说话人2 04:01  
做客户之声，

说话人1 04:03  
客户之声实际上是一种舆情分析，等一会我们可以看到一些细节。

另外比如说做银行的一些投诉，因为我们经常打比如说打他的call center的电话是吧？会有可能会咨询投诉等等之类的。当然现在很多都已经用机器人去替代了对吧？一些客服机器人，早些年的时候基本上你到语音打进去，然后有个后台的小姑娘来接电话，然后记录下来，然后给你一些这种feedback这样子对不对？

那里也有很多这种文本，当然还有其他的一些场景，像做欺诈，保险欺诈，我们现在比如说我们买车车险里边，

我们去报损的时候，会填大量的一些东西，你到4s店，比如说我的车被撞了，或者等等之类，你会填这些东西，你要说我要去理赔。

保险公司它有一块很重要的数据就是这种我们叫非结构化数据文本数据来，就是因为有很多保险欺诈的东西，如果你们不知道这些行业，保险欺诈实际上是每一年给保险公司，特别是财产保险造成的损失是非常巨大的，这些基本上是几百个亿，就是说所以像这种欺诈侦测对保险是非常紧要的一个一个业务流程，实际上是反过来说基于文本的这种欺诈对他们也是非常关键的，因为有些东西你在结构化数据里是看不出来的。

比如说我的一些地址，我的一些就是这些地点等等之类，你要把它关联起来去做分析，才能有可能找到一些欺诈的蛛丝马迹，

说话人2 05:46  
包括。

说话人1 05:47  
像反洗钱、智能投顾，智能投顾是证券行业在做的，包括内容推荐，包括像早些年还做过简历机器人，就好像大家现在是大三是吧？

很多时候比如说你要递简历的时候，现在都已经自动化了，你的简历很有可能这些公司会买一个相当于一个简历排查的机器人，它实际上是基于这种自然语言处理的一个自动化的程序，然后它会相当于一个初筛，然后第一次来判定你你符不符合我们公司的需要，这里面可以做很多的分析。

Ok这个是个基本的背景，我也是想就结合我之前的一些经历，然后再结合我们这次课程的要求，就给大家来上这门课这样子，这是我的一个。

简单的一个介绍这样子。下面我想请大家介绍自己，因为我们相互熟悉一下，因为时间我们还上课，然后我们选三位同学，一位就是我们班的班长是哪位或者学习委员有吗？能举下手吗？班长学习委员有吗？班干部有吗？不要怕，你们已经都大人了，哪位是班长？

说话人2 07:16  
你怎么称呼？

说话人1 07:20  
林彤，哪个童？儿童红彤彤的？彤，

说话人2 07:25  
 ok，

说话人1 07:27  
你就简单介绍自己，你们为什么会选这门课？

说话人2 07:34  
我不知道。 Ok。

说话人1 07:38  
那挺好的，答案不知道也挺好的。

我们再选另外两位同学之后，有什么事情我是不是可以直接找你联系，就是比如说就麻烦你有可能比如临时换课或等等之类的，你就通知一下大家这样子，你们是整个班都选了这门课吗？

说话人3 08:01  
领导就是说给我们开的，然后所有人都这不是我们自己选的，这好像不是自己选的，

说话人1 08:06  
这不是专业任选吗？所以我就很好奇，

说话人4 08:10  
任。

说话人1 08:11  
选课。

说话人5 08:14  
直接给我们安排，

说话人1 08:15  
你们必须要来选，我是。

Ok，我们再选两位同学，然后我们选选后面的同学，坐后面的穿红衣服的同学，你。

说话人2 08:28  
怎么称呼我？

说话人4 08:32  
对，陈

说话人1 08:34  
陈泽凯你们都是广东人吗？基本都是，你是广东什么地方的？我。

说话人4 08:40  
是福建的，我不是。

说话人2 08:42  
 Ok。

说话人1 08:44  
你是客家人吗那边？闽南不是客家人吗？我是。

说话人4 08:51  
我们不讲客家话，我们讲闽南话，ok，

说话人2 08:53  
你为什么。

说话人1 08:54  
你也是因为系领导的安排才不得不选这门课是吧？什么？我说是不是系领导的安排你就不得不选这门课，

说话人4 09:03  
我不知道，反正修够学分就好了。

Ok我们。

说话人1 09:08  
是考察课，所以。

应该不会难为大家，但是就是说还是希望能够帮到大家，或者大家在这个课程里面学到一些东西，你怎么称呼？张什么？我刚才忘了，

说话人2 09:27  
张广豪。

说话人1 09:29  
豪就是豪放的豪，你为什么选择也是一样的？就不得不选是吧？

说话人3 09:34  
对，

说话人2 09:36  
 Ok，

说话人1 09:36  
你们没有一个说有一个是主动发，自己发自内心的觉得这门课有意思，过来选的没有。

Ok。课后的话我做了一个excel，就是一个学生学情调查表，

说话人2 09:57  
然后。

说话人1 09:59  
我就请大家大概填一下，大概这个内容是比较简单的，你屏幕上可以看到大家的这个姓名、学号、性别，因为我现在跟大家不熟，所以就麻烦。

性别然后来自什么地方，然后这里有个组队，还有什么是否队长什么之类的，就是。

等一下我会介绍几页片子。我是希望利用这个课程，看大家有没有兴趣今年去组队去参加泰迪杯。

泰迪杯的话等一下我会介绍国内的一个针对大学生的一个a级的一个比赛，等一下我说我用教材是泰迪公司写的，所以如果大家有兴趣的话，就顺便把这个信息填在里面，但是有可能是因为太极杯它是要三个人组队的，所以肯定这里面有一个leader， Leader要先跳出来，就是说我想参加，然后有可能在班里找两个帮手这样子去搭配的去做好吧？

有可能是需要大家帮一下忙的。

另外一个我不知道大家有自己都有手提电脑，

说话人2 11:20  
对，然后。

说话人1 11:23  
方便的话可以把手提电脑的配置也大致填一下，最主要是内存跟 CPU，就是。

很有可能是我希望还是有的，因为我们光在课堂上去做所谓的实训，这个时间肯定是远远不够的，所以很多时候有可能需要大家在课后用用自己的手提，自己都不去跑一些东西，这样的话我会我会大概看大家的配置，然后觉得后面用什么样的一些包一些软件会更合适，我们可以灵活的调整好吧？

这是一块信息。

还有一块的话，我看到的所谓的一些选修的课程，

我不是太肯定，比如说我们有一门高级激励学习，大家都有选吗？还是说有的人会不会选这门课？所有的课都选了，那一有一个文本挖掘你们也全选了吗？我上面这4门课你们全部所有人都上了是吧？

说话人3 12:39  
对。

说话人1 12:41  
知识图谱的话你们也会全部人去上，

说话人2 12:45  
ok。

说话人1 12:47  
这块我改一下，大家可以不用填了好吧？

说话人2 12:51  
然后。

说话人1 12:53  
当然还有其他一个我也想去了解一下，大家如果可以的话也帮忙去填一下，比如说毕业后的一些职业的意向或者选修这门课的一些想法，

说话人2 13:08  
包括。

说话人1 13:10  
之前大家在线上的学习的一些资源的占比，包括大模型，包括大模型好吧？

说话人2 13:20  
我。

说话人1 13:21  
晚一点我看看班长林彤是吧？课间的时候我们加一下微信好吧？

说话人2 13:30  
我们。

说话人1 13:31  
就继续，这个是我们的课程框架，就是我们整个学期18次课，两个学时，两个学分，36个学时，每一次上两节课对吧？我们整个分的话，我觉得大家这个时间点学自然语言处理的话，

坦率的说是你可以说是既是即使有幸的很幸运，也是很不幸的，所谓幸运的话，看这个框架就是我们说的所谓，因为大家是人工智能专业，人工智能的话只有两个框，只有两个支柱对吧？

一个是CV计算机视觉这一块的图像这一块的，还有一块就自然语言处理，所以这两块的东西的话，

我印象中我们自然语言处理学的东西应该是比较少的，所以我觉得还是有必要是把这两个框架的内容都能够。

都能够比较好的去把握一下，这是另外一个的话，我们整个框架里面还有底下面的一些基础的东西，实际上是 Ai或者说人工智能的话，实际上是对数学对对计算机，还有对数据库对等等之类的这些要求还是挺高的。

所以我们这门课的话，

 focus在自然语言处理这一块，我刚才说的为什么不信，不信的话，最近两三年甚至说三四年自然语言处理大家应该知道是吧？0应该是01年还是02年， chat GDP出来之后，整个自然语言处理发生了非常大的一个变化，所以这个时候实际上我们这个课定位成是一个基础课，实际上自然语言处理课程非常多的非常杂，

说话人2 15:30  
这门课。

说话人1 15:32  
不是很好学，实际上也不是很好教，所以的话我们这个课是定位在一个基础的课程来的，实际上后面还有一些高级的课程，或者说一些专题的课程，所以这里的话有可能对我也是一个很大的挑战，就怎么样把我们一些基础的东西跟现在大模型的最新发展能够糅合在一起，这个是有可能是我们要共同努力的好吧？

我们在这个基础课程就是一个基本的框架，我们大概分成几块，首先的话我们有一个基本的简介，我们先把整个课程后面的一些基本概念能够铺垫出来，

一第一排的那几个4个列的，实际上就是我们说的自然语言处理的一些基本的任务，最底层的一些任务就是你做后面一些上游的任务，高级的任务，你有可能都会涉及到的，比如说我要做分词，我要做一些词性标注，或者说我要做一些向量化表示等等之类的，所以它是一个很基础的任务，我不知道你们学文本挖掘的时候这块有学吗？也有学吗？所有人都学了，包括从分时到向量化，ok我们有可能要去做一些调整，然后再往上的话就有可能是一般我们说的一些高级的任务。如果我对这个文本做完预处理的话，比如说我去做一些分类聚类，就常见的一些对吧？或者说做一些情感的分析，或者说做一些这种深度学习的东西。那么再往下实际上是现在就跟大模型结合的比较紧的就是做qa这一块。好好没事。做q这一块的q这一块应该是整个自然语言处理里面一个最高级的任务，而且也是现在基于这种所谓的大模型，

说话人2 17:43  
它所最。

说话人1 17:44  
重要的一个应用场景，所以标红的这几块，比如说像向量化，比如说ARP跟深度学习的一些关系的一些处理，包括一些相关的内容，

说话人2 17:56  
包括。

说话人1 17:57  
 Qa这一块，我们很有可能会结合目前大模型的一些东西去。

说话人2 18:04  
讲好。

说话人1 18:06  
整个课程的话就是考核方式的话，在右边这个饼图里边，因为我们是考察课，所以基本上我是希望是以这种项目的方式去推动我们整个课程，所以最最终我们60%是希望通过这种课题报告的方式来做期末的考察。

当然中间的话有可能比如说我们有百10%的一个日常的考察，另外有10%做理论部分的一些日常的作业，还有20%就是我们实训部分的一些日常的作业，通过这个作业的一个质量的话，来考察大家日常的一些学习的状况，这个是我们这个课程的一个基本的简介。

我们的参考资料的话最主要的就是用的这一本就是 Python中文自然语言处理基础与实战这本。大家如果方便的话可以去买一下，当然我也会把电子版提供给大家。我为什么选这本书？是因为我们是有要求一定要用近三年的一些所谓的教材，但是我刚才谈到的我们现在这个时间点学自然语言处理，实际上还有很大一个挑战，因为现在基本上大家都是用用大模型，用预训练模型，用深度学习的方式去讲，但是如果这样去讲的话，在我们所谓的36个课时里面是很难去把这个内容去完全覆盖的。

另外一个的话我还有其他的一些参考书，大家如果可以的话也可以去熟悉一下，而且我会在每一章的时候会指明，就是说这一章有可能建议大家是结合另外一本参考书去看，最主要是有5本，第一本是比较强烈建议的，一个是去年二三年出来的张琪写的一本是写的比较新的，应该他是复旦的本科生跟研究生用的一本自然语言处理的一本教材，应该还是很不错的，这个也是我们一个主要的参考书。

第二本是日本的一个斋藤康一写的一本就是面向自然语言处理的一个深度学习进阶的一本书，实际上它有两本书，这本书的话大家可以去豆瓣上去看看这点评。

他写的两本书，一个是深度学习的一个入门，还有一个是深度学习for Lp的都是有很好的一个评价，大家也可以去以它作为重要的参考书。

第三本是刘志远他们写的一本，这本书最主要是我觉得可读性比较强，因为很多时候我们想快速的去了解专业的一些东西的时候，因为很有可能有的书有可能里面的公公式非常多，对吧？

但是有可能我想先把一个它的基本的思想给把握住，所以我们同时可以去参考刘志远的这本书就是数据驱动的自然语言处理，它也是比较新的，是220年出版的还是不错的。

后面两部书就是大部头如果有兴趣的大家我刚才说了AI只有两个方向，如果大家日后想从事AI相关的职业的话，你不可能说我什么都做的，没有人这样子，没有人有这个能力，就好像是比如说如果你们问我图像相关的东西，我很有可能就会say no，因为我对图像不是太了解，因为这两个不同的这种方向差别，实际上是即便是自然语言处理里面的一些这种细分的一些方向也会很深，里面每一个主题都有可能会写一本专著出来。

所以的话如果大家对自然语言处理有进一步的兴趣的话，我们可就可以参考另外两本，一个就是自然语言处理综论，

这个也应该是西安交大他们的人工智能专业选用的一本教科书。

第五本就是钟诚庆是我们国内做中文自然语言处理是一个。

说话人2 22:28  
一个。

说话人1 22:29  
老先生说或者说是一个大家，所以我们说的中文自然语言处理就跟其他的有可能不一样，等一下我们会看，所以的话而且它里面它13年的它里面实际上是涉及到一些很基础的东西，是因为之前的话我们叫统计自然语言处理，我们在谈自然语言处理的时候，用的是很多是统计跟传统的继续学习的方法，但是现在整个范式是完全改变掉了，深度学习都已经被盖在底下，完全是预训练的这种模式了，所以这个是主要的参考资料，就是我们所选用的这一本的话，我建议大家都可以去买一下，好吧，我晚一些，我也会把电子版给到我们班长，你就帮忙给转发一下好吧？

另外一个就是教学设计这一块的话，我刚才谈到的我们这个时间点讲自然语言处理，而且你们刚才说你们里面有可能很多东西在上一门课里面都已经接触过了，因为我不知道接触的深浅，所以我们会灵活的调整。

我能想到的就是4个方面，第一个就是我会尽可能的把当前的大模型的一些最新的进展。

说话人2 23:50  
能够。

说话人1 23:51  
融到我们这个课程里面来，尽管我们还是有一个大概的框架，刚才说的按照这本书的框架把它走完，但是里面有的时候讲的时候，我会依据大家的表现或者说反应，我们可以去灵活的调整，主要的目的还是。

说话人2 24:09  
要把。

说话人1 24:10  
基础的一部分跟现在最新的发展能够衔接在一起，这是一个。

第二个就刚才谈到了，为什么让大家有没有兴趣去参加泰迪杯，我想因为泰迪杯里面等一会我会谈它里面一般每次考试有三道题，然后其中有一道就会是自然语言处理相关的题目。

我是希望就是说从我们这个课程班里面，如果有一些队想参加泰迪杯的话，我们可以去做自然语言处理的那道题，然后同样的话能够帮助我们一起。

说话人2 24:43  
就是。

说话人1 24:44  
边同步学习，然后边去进入实战。

第三个的话。

说话人2 24:51  
大家。

说话人1 24:52  
可以看到我这里是有一个有一个耳麦对吧？

说话人2 24:56  
我。

说话人1 24:57  
讲课的一些录音我会录下来，录下来之后的话，我也会转成这个文本，然后我们会把这些文本包括其他的一些文本，我们可以给一个大模型对吧？

我们可以通让大模型来看，因为我还在设计，比如说我们期末考试的时候，很有可能我是给你一份试题，但是这个试题不是让你做，这个试题是让你用你自己训练的大模型来做，然后根据你你自己用你自己训练的大模型，做这份试卷所获得的打分，作为你这次期末考试的打分。

但这里面有一个很大的问题，大家应该可以想到是什么问题，对吧？就是显卡的问题，但是应该我们很快可以解决，因为我们应该还是有两张4090的卡，应该可以很快把这一块给做起来，因为大模型加red这一块，是现在刚才我谈到做qa这一块，

说话人2 26:02  
目前。

说话人1 26:03  
最主流的方式，我想我会把主要我们这个课程的重点会放在放在这一块上好吧？

第四个的话就是与这个行业的实物相相衔接，我想在我们这个课程的时间里边应该会分两次，应该就相当于我们把我们的人分两拨，去两个不同的公司去调查访谈一下，我觉得有可能大家可以实际的看在实际的公司，而且是做自然语言处理这一块的，他们在干嘛，他们需要什么样的人，他们怎么样去管理的，就是组织这种技术团队的等等之类，当然你也可以根据你课程自己的一些感悟，也可以跟那些实物的人去做沟通，基本上是两家，一家是做知识图谱的，因为知识图谱你们也在学那门是吧？

知识图谱是自然语言里面一个很重要的一个应用方向或者说一块东西，当然尽管预训练模型对他影响很大，但是它的地位还是在那个地方。还有一个就是。

说话人2 27:19  
做。

说话人1 27:21  
自然语言查询的，什么叫自然语言查询？就是你们有没有学学过数据库？有学 Ok对。自然语言查询就txt c口，就是我给一个自然语言的话，比如说请统计一下，比如说我们这个班，你们上个学期比如说你们学了自然语言处理，你们自然学了机器学习，你们机器学习的平均分数是多少？我这个自然语言处理对吧？一个自然语言，但是我要把它丢给数据库，要去做一个select，对吧？做一个average什么东西，然后from什么东西，甚至说我说女同学的还要做一个while条件等等之类的，这是一个。

我作为有一个朋友，他是他们公司专门做这种有一块业务做这种开出c口的，但是有可能我们要分成两个不同的队，因为你人太多了，别人公司装不下，

说话人2 28:18  
这个。

说话人1 28:19  
是大概的想法好吧？

说话人2 28:21  
当。

说话人1 28:22  
还有一些约定是因为这是我回校园里面的第一门课，我这个学期只上这门课，我希望就刚刚说了，无论你们是怎么想的，我自己会尽力能够把这门课上好，然后也实际的我也不知道你们怎么看自己未来自己的发展，还是希望能帮到大家多一些选择。

旁边这几句话，这6个字实际上也不是给你们说的，是给我儿子说的，我儿子跟你们一样，现在在大学生比你们低一年，他现在二年级，他当时读大学的时候我给他说的，第一个要自律，实际上就是大学自我管理，像时间课堂纪律或者说有可能分团队的话不同的题目，因为在实务中做事情全是团队合作的，因为大家已经大三了，你很有可能就是下半个学期或甚至暑假就开始去找实习单位了，对吧？很多东西都是很现实的，这里我也推荐一本书，有兴趣大家可以去看。奈飞文化手册如果你们没有的话，可以把电子版给你们奈。

飞文化手册应该有8个这种企业文化的原则，其中第一条就是只招成年人，你大概知道什么意思？只招成年人，即便是在职场里面，大家要用一个成人的方式来做一些事情。第二个当然就是做就是自学，自学并不是说如果你让我们自学，那要你来干嘛的对吧？要我们这老师干嘛，不是这个意思，就是。

大学跟因为已经三年你们大学都快毕业了对吧？这是当时我跟大学跟跟中学跟高中是很不一样的，很多时候要靠自我去学习，你完全在课堂上是远远不够的。

另外一个就是刚才我推荐的几本书是吧，实际上大学里没有所谓的教材，这个概念没有教科书这个概念，我们只有参考书，我们只有主要参考书跟其他的参考书，是吧？同样一本书不同的老师去讲，他有可能讲法他的着重点，或者说他的这种思路是完全不一样的，因为大家是完全利用自己对这个东西的理解，或者自己的这种专长来看这个东西的。

所以在这一块的话我还是比较强调的，大家一定要我们以前在职场里面一句话叫 Learning by doing，一定要干中学，实际上你以后工作你都知道了，自学能力为什么很重要？

就是因为你在职场里面经常做这种事情，你的领导告诉你，就是说你去做个什么事情，你绝对不可能是说领导不好意思，我没学过，我不会对吧？

绝对不会这样，你必须是说当你可以说我有可能这个东西这里面我有可能哪一块不行，有可能你协调一个其他的资源来支持我一下，或者是说我还不是很熟，我有可能要花点时间把这东西捡起来，你能不能多给我点时间把袋子烂放缓一点，对不对？但是你绝对不可能说我不会我做不了，你把这个任务给其他人吧。

所以很多时候我们在职场里面大量的去碰到新东西新问题，都是要 learning by doing去做的，所以这一块我觉得大家一定要有这个sense。

当然第三个就是自驱，自驱最主要是因为大家是人工智能专业的，

而且不仅仅你们在学习，我们也在学习，应该是我们是下周下周我们老师还要培训，也是培训 disc，因为这次AI的东西。

说话人2 32:00  
会。

说话人1 32:01  
削减我们所有的人，而且但是你们本身就是学AI的，对吧？

如果你们被浪潮抛到所抛下去的话，我觉得是很不应该的，很不应该的，所以我觉得大家也可以多去自己去了解这一块AI，包括现在都不谈AI了，现在都谈agi，ASI都已经这样子了，你都可以知道这个东西是发展的有多快，好吧？

好了，废话大概是这样子，然后我们就转入正式的课程，我分两节，这次都是一些概念，压力不是很大，我们是不是下课了？没有吧？Ok。一个是LP简介，还有一个就是语料库。 Lp简介的话就是一些基本的概念，我们就大概过一下，我是希望就是说。

说话人2 32:57  
通过。

说话人1 32:58  
这一讲的话，大家能够把 LP它研究内容的框架主要是框架，因为有些东西你没必要去了解细节，但是这个框架一定要有，包括它里面的一些子任务一定要。

说话人2 33:14  
就。

说话人1 33:14  
像我刚才说的里面每一块东西都可以写一门专著的，比如说我们讲到情感分析的时候，我可以给你一本参考书，那个人写做情感分析的，写了一本几百页的书，那个书就是叫情感分析，就每一块都可以写一本书出来，所以你比如机器翻译我们不会讲的，因为翻译这个问题基本上已经是解决了一个问题，大家现在都不会谈机器翻译了，而且因为现在机器翻译更多的是在一些实用中的东西，或者说一些小语种的东西，像之前我们有个项目做这种东南亚语言怎么样子跟中文之间的呼应都是那种东西对吧？

所以这个大家一定要有框架，另外一个就要有流程，就是LP一个任务，如果给你去做，给你个事情让你去做，它的一般流程是什么？里面会有一些哪些子任务，我自己擅长什么东西，我自己不擅长什么东西，就像刚才说的，如果我擅长的，我去own这个事情的时候，我怎么去领导一个团队，甚至说用我自己的其他东西，我去查漏补缺，把它给做出来。

这个框架性的东西比细节的内容要重要，大家一定要记住这句话，特别是对自然语言处理，自然语言处理是相对于图像或者相对于其他的东西，它是非常杂的，非常杂，而且它的变化是非常快的，像我自己现在也在学，是因为我原先积累了很多，像我做华为的项目，10年之前的项目，那套知识体系已经完全被淘汰掉了，等一会你就可以看得到，所以大家要学这块东西的话，一定是要有一个很open的心态去做这块东西。

第一个知识点，自然语言处理的概念，这里面实际上就是两个东西，一个叫什么叫自然语言，一个就是什么叫为什么加处理，或者什么叫自然语处理，因为就多了个处理自然语言大家应该比较清楚，我们现在说的都是自然语言对吧？没有什么特别的当然但是自然语言的话，它相对的一个概念你应该要知道，相对于自然语言就是人造语言，像我们编程Python，我们写c我们anyway。

说话人2 35:21  
包括。

说话人1 35:22  
我记得好像是之前有语言学家还创造了一门叫世界语的东西，我不知道你们有没有听过对吧？

那些都是人造语言，不是自然语言，另外我们自然语言。

分层我们都知道像口我现在说的是口头语言对吧？有可能让你们交作业，那个作业那是个书面语言对吧？所以语音分析跟自然语言处理它基本上合在一起的，基本上是合在一起的，我们最近楼下在搞实验室，你们知道吗？我们跟科大讯飞合作，你们的师弟就学弟有可能会有一些新东西去搞科大讯飞不就是做语音的吗对吧？他做语音起家的，这是一个。

第二个自然语言它的几个特点大家要去了解一下，实际上自然语言绝对不是我们从面上看起来就是那么一个很简单，自然语言是很复杂的，或者说它是很深深邃的一个东西，就是一个鸡生蛋蛋生蛋生鸡的问题，你就可以你可以问自己先有自然语言还是先有智能，你去想一想这个问题都很难去回答。

所以这一块大家如果要学自然语言的话，特别是如果有机会以后，你去在实务中去做自然语言的话，就一定要对自然语言这个方向或者说这个专业问题充满了敬畏，我做了20多年我还是学数学的，我最怕的就是两个方向，一个是自然语言，我是很谨慎的，如果有这种项目这种题目，我一定要问清楚你要干嘛，你期望是什么？

还有一个是什么问题大家能猜到吗？还有一个也很难，但你看起来很简单，就是时间序列做数据预测也很难，你不要以为你学过时间序列，你们学了统计学是吧？统计学里面什么阿瑞玛那套东西，因为时间序列这个东西是立马就是你做了一个预测值数字预测是吧？比如像股价我明天预测它会到多少点，它立马就可以去检验你预测的好坏了，对不对？而你真的预测的很准是很难的，自然语言也是因为自然语言谁都懂，对吧？你要把这个东西给搞准的话，用一些数学模型的机器学习模型的东西，别人一眼就看出来了，你分类分的不准，对吧？

你情绪判定的准不准，你回答问题你是个弱智这样子来回答，还是说你真的是我靠很牛逼这种多人会话，什么意图理解都很准，一一试就试出来了，因为所有人都懂自然语言交互界面是自然语言，所以这两类问题特别自然语言就是一定大家要对它充满了敬畏。

第二块当然就是自然语言处理，当然就是我们这门学科就为什么加了处理这两个字，对吧？为什么不叫自然语言分析？你们是学的文本挖掘对吧？为什么不叫自然语言挖掘？对吧？为什么叫自然语言处理？是因为自然语言就是从头到尾整个环节非常庞杂，而且里面很多环节并不一定就一定要用到数学的东西，用一些模型的东西它是很复杂的东西，

从你接入这个语料到你一直到最下游的任务，比如说做这种智能问答对吧？

这个是个很复杂的一个流，所以它里面有很多的这种任务，所以的话基本上我们是笼统的去把它定义成这个叫处理，而不是分析或者说挖掘。

所以在这里的话我们也可以去辨析一下，就是说比如说我们在谈文本分析的时候，我们很多的时候是强调落在分析上，分析，当分析的话就包含了一般的，比如说我们说不用到那种数据挖掘和机器学习模型的那种分析，那个是叫我们有挖掘，因为我们有挖掘的话，实际上就是把数据挖掘平推到这种文本数据上，对不对？它基本上还是一些核心是那些传统的那种自然语言数据挖掘的那些算法。

所以这里面这个概念可以从这个角度去理解，所以我们说的 Ta文本分析，它肯定是小于 LP的，就是它里面的分析技术不一定是数据挖掘，我做一个持平分析，然后通过一些可视化的手段，比如说一些热力图或等等之类的，我讲完这一页我们要加快文本挖掘的话比 ta还要小，因为它主要是focus在数据挖掘的算法上对吧？

当然就是说还有计算机语言学计算机语言是语言学他们山头提出来的一个概念，他跟我们比如从计算机，因为人工智能是从计算机这个角度切进来的，它还是有一些不一样的地方，它主要是用数学的方法去解决它语言学里面的一些问题。我们主要是用计算机结合了语言学的一些东西，解决我们在实际中碰到的一些这种自然语言的一些应用问题，好吧？我们先讲到这一页，我们休息10分钟好吧？

原来没有。

说话人4 41:00  
意思，这是第一个问题，

说话人1 41:29  
我们是40分钟一节课还是50分钟一节课？4那就是40分那很短，我以为是50分钟一节课。你们后面还有其他课吗？

说话人2 44:29  
没有了。

说话人1 44:32  
知识图谱是明天也是这个时间吗？知识图谱谁哪个老师给你们上？文老师你们文本挖掘用哪本书？忘记了什么时候学。

说话人3 45:15  
的，没有。

说话人2 45:16  
书。

说话人1 45:21  
几年级学的？是上个学期还是上。

说话人4 45:24  
二。

说话人1 45:25  
年级入学我天。

说话人3 45:44  
当时有打竞赛，然后是搞我们的话文本处理的，然后有拿奖，然后老师其实让我们几个免修了。

说话人1 45:53  
免修什么就是课。

说话人3 45:55  
课上都没有意见，这是我们做的内容，跟他交流已经就差不多了。

说话人2 46:01  
 Ok。

说话人1 46:04  
那也是一个选修课，对吧？但是我记得这个是任选跟必选有什么区别？

说话人3 46:12  
主要是我们学校主要是我们院系的选修课太少了，但是我们学分要求也比较高，然后所以就。

说话人1 46:20  
拿这个学分是吧？Ok，你买了这本书是吧？是最近买的吗？但是你已经翻成这样子了，这是。

说话人2 46:40  
二手的 Ok。

说话人1 46:44  
我这个PPT会转成PDF，然后你们有去用那个叫什么？学习通吗？对，你在学习通上应该可以发现我那门课程，我稍后有些东西也可以放在那儿，但是这个教材的电子版我就给到林林童是吧？你就转发给大家好吧？因为这里面有个版权的问题，我害怕老师直接给学生又被别人怎么样，但是这些这几本书的电子版我都有，就这种书的电子版我就给到你一个人，你们学生之间转是没问题的，好吧？

说话人3 47:23  
说明什么？这个钱不够吗？

说话人4 47:26  
我大一点工资钱，然后你是通过把的情况你说谁知光多少次，谁知谁知我们的思想还是创新的情况。

说话人1 47:51  
林彤你加我一下微信好吧？就有些东西有可能就需要你帮忙。

说话人4 48:06  
对您把钱收回来。

说话人2 48:08  
好好。

说话人4 48:13  
什么你特别给我，你告诉的原则告诉人家告诉我，谁知这种指标谁知。

说话人1 48:25  
你没有人买过游戏本吗？

说话人2 48:29  
没有，

说话人1 48:30  
你打游戏用什么？用普通本。

说话人4 48:33  
轻薄本玩的都是439。

说话人1 48:35  
轻薄本打游戏，

显卡够吗？

说话人2 48:41  
不够吗？

说话人1 48:42  
黑神话打过没有？跑不起来。能跑起来吗？

说话人4 48:47  
跑不起来我不知道，

说话人1 48:50  
我是买了个游戏本，但我不是打游戏的，我是要它显卡很好。

说话人4 48:59  
还有你看我如果不要提了，你自己会把你所有的造成下去的活动了，然后把它理解我们说的我们的生意不一样的。

说话人1 49:12  
你们的一般打什么游戏，

说话人4 49:16  
一般什么什么东西，法网什么？

说话人2 49:22  
网。

说话人5 49:22  
游戏。

说话人4 49:24  
电脑为什么有？

说话人1 49:25  
我看一下你们跟我儿子是不是打的相同的游戏，我靠。

说话人4 49:29  
跟CS一样的。

说话人1 49:32  
那种打枪的 Ps的。

说话人4 49:36  
前站吗？

说话人1 49:38  
用手柄玩还是用鼠标点？

说话人4 49:41  
用。

说话人2 49:43  
鼠标的，

说话人1 49:46  
就那种组队的那种是吧？

说话人2 49:49  
对，

说话人1 49:50  
我知道我儿子一晚上也是组队，我靠太好。

说话人4 49:53  
了。因为。

说话人1 49:56  
透着门就听到他在那喊什么快点什么的。

现在的孩子你们这代人都喜欢打游戏了，没办法不。

说话人4 50:08  
理解。

说话人1 50:19  
好了，我们开始了，上课了，好吧？

我们下面内容还很多，然后我们就快速的过一下，刚才。

说话人2 50:26  
大家坐好好。

说话人1 50:28  
同学史努比坐好，然后我们就继续然后有什么问题大家可以打断我，另外一个这一页是很关键的，大家了解一下就LP的研究内容。

我觉得最主要就是说我刚才说了它是。

说话人2 50:48  
非常。

说话人1 50:49  
庞杂，而且变化很快的，如果你想在这一块里面，你以后做这个方向的话，如果你一开始不建立好好的框架的话，很容易会迷失掉的。

所以我是很希望大家一开始的时候就死活要在自己脑子里要印一个很很深的一个框架，然后以后遇到什么新的一些知识概念模型，特别是深度学习的模型非常多，你自己有可能会晕到里面的，所以大家一定要建立一个框架，这个框架的话大概三个维度，一个是力度，一个是层次，还有一个是范式，力度是很好理解的，对吧？

力度我们都是都是中学学语文过来的对吧？我们的词、句子、篇章、段落甚至说一本书对吧？这个是比较好理解的。

范式也比较好理解，因为我们本身还学了那么多数学的东西，对吧？你是基于规则的，基于机器学习的，是基于深度学习的，或目前的基于预训练模型和大模型的，这个是很容易看清楚的。

就是另外一个维度就是基于这个层次，大家有可能要把这几个概念先敲到自己脑子里就是说什么叫语，用什么叫形态，什么叫语法，大概跟我们小时候学中文是一样的，我们也是从拼音字母到认字到词，然后到句子，讲句子的时候涉及到语法对吧？

然后到后面阅读理解，阅读理解实际上就相当于语用了，就说自然语言在一个特定的上下文里面或者场景里面怎么去用，基本上也是分成这几块对不对？

所以要从这里面去看这些任务，所以我们基本上会涉及到比如像形态里面就像我们做一些词形的分析，或者说一些语法里面做这种词性标注或者语法分析，那么语义的话像做一些这种语义的分析或者说语义的消息，实际上在我们最终所服务的都是一些所谓的语用的分析，就是我们在真正的应用的或使用的场景，我到底用它来解决什么问题的，我是要做应答，还是做分类，还是做情绪的这种辨别，还是做翻译是吧？

还是做摘要等等之类的，就是属于语用的范畴，所以大家一定要建立起这么一个一个学习的框架。

说话人2 53:17  
当然另外。

说话人1 53:18  
一块就是ARP的挑战，就我刚才说了，我们要很敬畏他，

光敬畏还不够，是因为实际中它的确是很复杂，最主要就是歧义的问题。你像左边这个段子，大家应该都见过这种段子是吧？原先喜欢一个人，现在喜欢一个人，像里面有一个我看了半天我都不知道什么意思，我靠。第九个。

说话人2 53:44  
我。

说话人1 53:46  
大然后父亲说你大我都不知道什么意思，因为我知道大大在北方里面应该是伯父大伯的意思，但是实际上在他那个话里面还是有的东西是我作为一个一个所谓的到这个年我都不知道他怎么去识别它，或者说它到底是什么意思。

说话人2 54:10  
你。

说话人1 54:10  
要像第八个对吧？像。

说话人2 54:13  
有。

说话人1 54:13  
可能我们仔细看还能够看明白对吧？

说话人2 54:16  
明明。

说话人1 54:20  
明明白白。明白，白白喜欢他，实际上这里面明明跟白白是两个人，对吧？中间有个明明白白，另外一个也是有所谓的明明或者说明白这样子去切，然后我们有可能把这个句子的含义给捕捉到，对不对？

所以自然语言处理最大的问题就是这些，就是右边所谓的这一堆的歧义，包括语音的歧义，好像是他们当时有个笑话，就是说我们的国家领导人这个去我忘了是哪一任总统，是克林顿还是奥巴马是吧？这谁是who？这是谁？这是胡书记是吧？这是 This is who。 Who。在英语里面你到底是who是谁，还是说这是胡先生对吧？

所以这是语音的歧义，当然也包括其他的像切分词义，词义也比较常见，你比如说打鼓，打酱油，我不希望大家上课来打酱油，类似于他相同的一个形状，但是有不同的含义。

比如说像下面这些大家有可能比我要了解的最新的一些词汇，我们在比如说在社交文本里面经常遇到最新的词汇，对吧？比如说躺平，比如说yyds这些都是我们新一代的，像你们都是00后对吧？

00后所弄出来一堆这种文本，一般的人，特别是年纪大的根本看不明白什么意思，

比如说我们做产品评价的时候，如果说什么华为的屏幕，什么yy DS或者小米的什么轻薄本什么，你怎么去分析客户对这个产品的态度，如果你没有把新的词汇考虑进来，所以这个也是 ARP里面它所面临的一个挑战，最关键的就是歧义要靠我们后面的一些一些方法一些数学模型的去加以解决，这个是自然语言处理它的一个发展的简史，我觉得大家可以把4个阶段大致的能够区别开来。

就是1个早期的时候大概是。

70年就是70 60年代，一直到60年代，基本上就是所谓的一个初创时期，到80年代就是90年之前，大概是所谓的理性主义时期到2000年，大概是叫经验经验主义时期，到2010年就2020年之前大概是深入学习的时期，然后到现在我们叫预训练模型或者大模型，当然我们说的大语言模型不是那种多模态的或者说图像的大模型，我们一般说的大模型的大语言模型，每个时期实际上他的思路还是比较它的特点还是比较好分辨的，比如初创期，我们先不管它有两个流派，一个是符号学派，一个随机学派。

符号学派就是形式通过一些数据逻辑形式的去做一些语言的分析的工作，当然水利学派就是把统计学引进来了，而当最终统计学在一定的时期是占据自然语言处理最核心的方法。

所以我刚才一般我们书原先的书叫统计自然语言处理，比如说我们当时学的叫统计自然语言处理，那么理性主义时期当然就是基于系统的研究，有大量的系统，经验主义时期就是数据驱动，就是积极学习统计学的方法进来，去解决相应的自然语言处理的问题。

那么再往后的话，因为跟大家的生活是有瓜葛的，应该很清楚的深度学习，因为深度学习。

先是在这个图像里面取得李飞他们取得了很大的成功是吧？

然后就开始横扫自然语言这个领域，到现在我22年还是21年那个时候欠的GDP出来之后，基本上整个的范式完全转到大圆模型这样来，所以基本上我们可以从这几个角度去看，在这里我会推荐一本书，这本书我不知道大家有没有看过，就尼克写得非常好，就是叫人工智能简史。当他最近也出了本书叫理解图灵，我最近也在看，你们是学自然语言，你们是学所谓AI专业的，我建议大家去看一下这本书，我有电子版，我也可以给到大家有兴趣可以去看一看。

他写的就跟刘志远那本书很像，写的比较轻松，没有公式，没有公示，因为我们知道大家有可能都很怕公示没有公示，但是你可以先了解一下这个行业它的一些人物，他的思想的发展的脉络，一些趣事，然后你看看能不能就说实际上是这个方向还是很有意思的一个方向。

结合刚才的发展阶段，就是说我们说整个LP就是所谓的四大范式，就是基于规则的，基于激励学习的或者当激励学习。

早些年的时候机器学习是属于统计学的，我是学统计的，学的话只是后来这个山头搞大了，所以机器学习的话就从统计学里分出来了，就好像统计学从数学里分出来了，对不对？

我们也可以比较理性的去看，所以很多时候我们会会把统计跟激励学习有可能会画个等号或者约等号，但是到现在应该分得很清，就是说你搞统计的搞统计，你搞经济学习就搞通搞搞激励学习，但是但是在这个阶段的时候，我刚才说了很多书都是叫统计自然语言处理他用的这些统计的方法那些方法，你比如说聚类，凭什么说聚类是个机器学习的算法对吧？

聚类很早的时候就在统计学里得到了发展了，对不对？然后再到深度学习，这个是看得很清楚的，对吧？应该是去年三个图灵奖的得主，就是所谓的搞深度学习的三驾马车，他们都同时拿了这个图灵奖，对不对？

然后再到221年不是19年或者是20年那个阶段疫情的时候，欠的GDP就是横空出世，我想大家可以分辨的清，但是我这里要强调一下，即便到现在这个阶段，他们有这样一个发展的脉络，但是你并不能够说谁好谁坏，你不能说预训练模型就一定比机器学习要好，或者说你一定说深度学习的针对于特定的场景任务，就一定要比这个基于规则就很low，逼格很low，我现在搞个高逼格的，我靠任何问题都扯出一个所谓的深度学习模型出来，没必要。

我们在实务中是讲讲求的是这个目标导向的，你只要有方法，你即便是用规则的方法把这个问题达到最终的应用目标的这种要求，就ok不一定要用深度学习的方式，因为有些东西你没办法用，比如说在实时场景之下对文本做拦截，你觉得用深度学习来得及吗？来不及，你想都想得到，你光靠深度学习拦不住的，因为在实时场景下做拦截的话，那个时间非常短，在某些特定场景之下都有可能不是到毫秒了，有可能到微秒了，哪个大哪个小，反正就我忘了，它是很小的一个时间窗口的。

这个时候你要在系统里面，你要去通过这种在线的深度学习或机器学习模型去做一个判断，做一个推断根本来不及。所以很多时候在那种场景之下，比如做欺诈或者做一些电信欺诈，我运营商为什么把叫什么地方？缅甸的那个地方，缅北的短信给你，欺诈短信给拦截拦截住，我靠模型来得及吗？来不及了。

我很多时候要靠关键词靠规则，因为我可以很快的做判断，所以这些方法你千万不要觉得就是说我一堆数学就很牛逼，不一定所有东西都是应用导向的，目标导向大家一定要记住这个东西。

说话人2 01:03:07  
好，

说话人1 01:03:09  
它的另外一个就是它的发展现状，发展现状你要问我，你说乔老师 Arp现在的发展现状是什么？我肯定会说我不知道，尽管我自己是搞自然语言处理，这个东西很复杂，因为我句话说nrp太泛了，你到底问什么对吧？你是问机器翻译吗？机器翻译问题我刚才说了基本上解决了，还是说语音识别语音识别也基本解决了，你还是说问智能问答还是问什么什么？

有一个叫我们最近国内的中文有个什么信息处理的一个委员会，他们编的21年的一个所谓中文信息处理发展报告，21年是最新的，他所谓做这个发展现状是什么？他结果我大概看了一下，我靠写了500多页，将近500页把这个现状给勾画出来，而且分了大概是15个专业领域，这个还是挺难回答的问题的。

但是如果我们要去理解这个问题的话，并不是没有途径的，我们可以把它简化一下，是因为现在自然语言处理就是预训练模型，就是大模型现在大模型发展的状态是什么？我们去理解大模型，或者理解这种人工智能的这种模型的时候，我们是有一套框架的，是吧？你们有可能看新闻都知道4个要素是吧？数据算法、算力还有场景，每一个你可以去大概想一想数据现在是个什么状况，算法是个什么状况，算力是个算力，大家都看得很清楚了，对吧？卡脖子，我们没有好的显卡，这是现实，因为显卡要把芯片给搞出来，我们没有光刻机对吧？这个就不用说了，这个现状你大概就知道了，算法。

算法的确大家有各种各样的说法，对吧？最近像图灵奖的之一，我一下忘了名字了，年龄大了，是因为我们现在所有的预期的模型都是基于全是方法机制的，但是他的说法就是说忘了诠释方法，为什么？因为它是个概率推断，你的生成的下一个字我是通过概率猜出来的，然后给你狂喂语料，这个东西不是我们人类不是这样子，你跟我对话，乔老师你好，然后我说你好，我跟你说你好的时候我是靠猜吗？不是，我是靠什么？我就靠逻辑推，因为我知道这是一个人给我打招呼，我说您吃了吗？吃了，我是靠逻辑来给你把吃了两个字弄出来的，我不是靠概率猜出来的，对不对？

这里面是有一个很深的大家有兴趣可以看。

最近几个视频他们在质疑，就是说目前的大模型的算法的基础有可能出了偏差的，所以现在大家研究很多做研究因果推断，因为因果在我们人类思维里面，因为我们人类思维也是用语言式思维的，是一个很深的一个东西，当然包括数据也包括场景我就不多说了。

当然这里我们就不能不说dpc对吧？现在你们不可能不知道dpc对吧？

说话人2 01:06:22  
寒假。

说话人1 01:06:23  
除了哪咤。

说话人2 01:06:25  
二我。

说话人1 01:06:26  
也去看了100，现在我不知道现在一百几了，120有没有120 140了。140150可以到前三了好像是，我记得好像150可以到前三了。

这个暑假并不仅仅是哪咤二暑假，最疯狂的事情就是引起，而且挪贷才100多亿，就是dpc DC的那一天造成了美国股市它暴跌是多少？4.26万亿，那个是100多亿，那个是4万多亿，你想想差多少倍？人民币都是人民币是disc。梁文聪他们湛江人，我们有湛江人吗？你是很牛逼，湛江人，是湛江人吗是吧？我记得他是湛江人，搞出来的非常厉害，因为我刚刚谈到了不是我们有4个框架，算力我们说不要说了，因为涉及到芯片，但是他们团队在这个方面，因为大概训练一次这种大模型，你基本上跑一跑，大概是我记得那个数字。

现在我不知道原先像opinion大概是100多美元，一一百多还是1000多美元就跑一次训练，很大概是一个亿人民币跑一次，原先最早的时候，所以他们在这里是。

说话人2 01:08:00  
做了。

说话人1 01:08:01  
很多突破，非常轰动，以至于就两个东西大家知道谈的最多的就是所谓国运级的技术创新，什么叫国运级？

现在一直在持续，所以大家也是有幸学人工智能，什么叫国运级的，然后谁什么意义？

他是说是从鸦片战争以来，中国对人类最大的科技震撼就是disc，这个是饶毅是大家要知道这里面是有很深的东西的，因为刚刚说了人工智能已经在朝agi就是通用人工智能到ASI就是超级人工智能，超级人工智能，就是说它已经超过人类的智能水平了，而且一旦是哪个国家掌握了这个东西的话，要远远超越某个国家掌，你有没有原子弹比威力还要大，有可能大上千倍上万倍。

是非常恐怖一个事情，就是我们小时候看的所有的所谓的科幻电影都有可能成为现实，我印象中是施瓦辛格演的，我不知道你们看过没有，那个叫什么名字，从未来回来的，然后他们到地堡里去什么之类的天网，这个是非常深刻的一个事情，我希望大家能够意识到后面的一些东西，我想给大家特别说，当然他们有很多这个技术上的突破，我们也会利用这个课程，有机会可以去体会一下别人是怎么样去做，而且他们都是一帮土博士，没有一个是从海外回来的，都是国内培养的一帮博士很年轻，基本上都是30岁左右，另外一个就是 Arp的应用场景，这里的话我也是希望大家。

能够建立起这个框架，从这个基础任务一直到后面的所谓的行业应用，因为我们讲场景的时候，我们是讲行业场景，是讲行业场景，你是哪个行业的场景？

比如说银行，像我做过银行、做过保险、做过证券，做过电信、做过制造业、做过零售，做过互联网，做过好多乱七八糟的东西，但是它有个场景是通用的，你比如说像运营商，电信跟保险都有可能做客户管理对吧？

客户管理是再往下抽象一层的一一一个场景。

那么在做客户管理里面有可能又有一个细分的场景，比如说客户营销，我们经常做客户营销对吧？客户营销里面有不同的营销的方法，有可能对不同的行业的这种客户都可以借助于比如说我们叫内容营销的方式，不是打电话，是基于你的内容偏好去给你推荐营销的信息，那个就是我们自然语言处理的一个场一个下游任务了，对不对？

你同样去做内容推荐，你可以基于情绪判断去做，你可以基于分类去做，你找到一个主题，你甚至基于主题模型去做，甚至说我基于关键词基于规则去做都可以，那是我们分析算法那一层的东西，当然前面的话你都要做分词，都要去你可以去爬，或者说你可以去获取你的语料等等之类，大家一定要把这个框架给建立起来，它中间是个网状结构，是个网状结构，你不是说我这个问题就一定要用那个方法去做，不一定的，我们学过机器学习是我们的分类，分类模型应该有十几种模型对不对？

你从决策树到罗伯斯基回归到神经网络到知识现场及你即便是决策树，决策树我我记得应该大概有20多种树，对吧？你用哪种树，你是用卡的还是用气的，是吧？

还有一种C4.5还是用C5.0，你光觉得树都有20多种树都是分类模型，所以我们一定要是建立起这样一个框架的意思，解决问题的时候，我们是从后往前倒的，我们要看就是场景问题，然后另外一个就是从前往后倒，然后再看我们有什么数据再把中间的东西给推导出来，千万不要抓住一个环节就死命去搞，没含义的。

因为这是个系统工程，你不能说我算法做牛逼，我就可以把这个问题给别人解决，不一定的，你没有数据，巧妇难为无米之炊，你没有好的数据，你数据质量很差，你又不去做数据质量的整改，你怎么能够把这问题解决好呢？对不对？当你说你有好的数据，你也有好的应用场景，你算法你不懂，你也没办法，你只能用规则规则还是比较粗放的，对不对？所以大家一定要有意识好吧？

另外是关于流程，流程的话我不知道你们之前有没有学过，就是说我们做分析的话经常谈，比如像数据挖掘的流程或KTV的流程，或者说数据平台建设的流程，我们 LP的流程跟流程都是一样的，一我们有个字眼叫分析闭环，这个闭环我们做所有的分析，无论是结构化数据还是非结构化数据，都一定要形成一个分析的闭环，就是我去做数据获取，

说话人2 01:13:38  
然后做。

说话人1 01:13:40  
做一些预处理，做一些清洗，然后去做特征工程跟建模，然后做评价做部署，然后做反馈。

这个闭环我一定要形成对数不同类型的数据是结构化的数据还是文本的数据，甚至说图像的数据都是用这种方式去做的，只是说在里面某一个方块里面，因为这种数据它有自己的不同的规格，因为文本跟一些数量的数据解卦肯定是不一样的，对不对？我可能会有一些额外的任务出来，对不对？

你一定要建立你看到一个需求任务的时候，你要首先把这个需求任务分到不同的处理环节里去，就说这个环节我要做什么，这个环节的任务要做什么，然后再去做进一步的细分，这个是大家需要建立的一个ARP的实施流程的，而且不同的方法它可以在哪些环节去用，比如基于规则我做清洗可以用可以用规则去清洗，但是我做一个基于规则的，比如说我做这个刚才说了这种欺诈短信的拦截，我也可以用规则去做，因为那是个实时和准实时的场景对不对？

所以不同的方法你不能说我基于规则的方法，你只能做预处理或做一些清洗的东西，不一定的，我也可以做模型部署，我用规则引擎部署到生产系统，我们经常做一些所谓规则引擎的部署，做一些这种实时准实时场景的一些逻辑都是没问题的，所以每一种方法你要搞清楚，它能是你说这种激励学习的东西就只能用到这种建模模型构造环节吗？不一定有一些很复杂的数据质量的问题，你不得不解决的话，有可能你要用机器学习的方式去解决，明白吗？它中间是个网状结构，刚才说大家说了大家一定要有这个sense，好吧？我要快一点，还有15分钟。

我们这个课程的开发环境我就不多说了，大家可以看看书，也可以准备一下，我在这里应该是有准备的，到时候我再去check一下，用and called用Python。用spider，当然你说你熟其他东西没问题，只要你能够用你自己的熟悉的方式，把代码跑通都没问题，但是我们教材提供的代码是基于spider的，我就不想去改了，好吧，我就用spider是吧？

然后我们会涉及到分词是用j8，但是你说有更好的没问题，你可以用更好的开源的东西去试。

比如说老师是用j8做的分词，但是我自己去做 home work的时候，我用了其他的一些我记得有什么发射test或什么之类的，我用另外一个开源包去做分析，没有问题。我非常而且你可以告诉我，你说乔老师我这个比你上课教了分析的效果，你看分析的要好，

说话人2 01:16:32  
没问题。

说话人1 01:16:34  
我是鼓励大家去研究开源的东西，研究开源社区，这个是非常关键的， Disc就开源了，这是disc之所以轰动的另外一个问题。

它开源了，开源非常重要，我们深入学习是用test flow，你要说我用 pen touch也没有问题好吧？因为我不知道你们之前用的是哪个框架好吧？当唯一我现在不能定的就是说我们做rag，我们用啥也可以这个设计的显卡，但是我觉得dvc肯定要去用的，但dvc是哪个版本？不好说，因为这是跟显卡有关，我们尽可能用一个就是高精度的版本。

说话人2 01:17:27  
另外一个。

说话人1 01:17:27  
就是大模型的开发框架用低费，但是说其他的东西也有很多，

因为你们在学知识图谱是不是我们也有可能会同时介绍有一个 for graph的一个是因为做rag里面有一块是做那种 graph的rag的把把知识要把它转化成知识图谱的，基于知识图谱来做问答，

说话人2 01:17:52  
我们也会。

说话人1 01:17:53  
考虑，但这一块我现在没办法给你答案，我要看我们的资源，大家我希望大家是因为基本上我想大家notebook都可以support，所以你可以在自己电脑上先去装上，都可以去先因为你们都学过拍摄对吧？

都可以先慢慢回忆起来，然后我们可以慢慢去试这样好吧？我就举个例子，有可能讲了半天，你不知道是干嘛的，对吧？这东西有什么意思，然后我们这10年之前应该是13还是14，我忘了。

还是1415就给华为手机，当时我在SARS公司做的还。

说话人2 01:18:33  
叫。

说话人1 01:18:34  
 VOC就voice of cast客户资深的一个项目，你看他干嘛，他从各种各样的社交数据的发源地，比如说像rkrks是它自己的一个call center的系统，因为它call center会有产生很多语音的一些文本，800热线pm pm是个售后，就是比如我们手机要维保，要维修，送到厂家去维修，它，它在售后会形成一个流转的环节，有专门的系统去管理里面会有大量的非结构化的文本，因为我要去维修，比如工程师要给意见什么那的，里面是很多文本，那么还有一块像像微博什么，大家都知道这种公开的市场，它把它抓过来做什么东西做，比如说你可以看到这些字眼，做产品的问题识别，做竞品的分析，做危机事件管理，做正负舆情分析，那么他还能做什么？

新技术的挖掘，做需求的挖掘，做消费者的特征分析，做什么社会社媒影响力分析，anyway它可以做很多东西，这个东西我们叫叫手写，我不知道你们有没有知不知道CRM系统就是客户管理客户关系管理系统CRM这个东西就是叫手写的CRM社交化的客户关系管理系统，但现在已经很成熟了，现在已经很纯，这是我们10年之前做的，我们怎么做就是这么做的，我们是用SARS的，SARS就是sas是美国的一家公司，是全球最大的没有上市的一个软件公司，而且在谷歌出来之前，全球的最佳雇主就这家公司很厉害，他老板叫good night，

我们就是用SARS的工具做，它最主要有几个模块，当然现在全部用Python的包去搞了是吧？

说话人2 01:20:35  
做垃圾。

说话人1 01:20:36  
的信息过滤，做分观点的识别分类，这里就是基于规则立体，就SARS自己搞的一套规则引擎，做情感分析，情感分析就是这种我们用hybrid一样，就是用混合模型，又会用机器学习的模型，分类模型又会用基于立体的规则去提取，你千万不要是把用模型把自己给定死，然后会做一些竞品相关的一些分析。

我举个例子，你大概什么叫观点识别，就是分成这个品牌领域跟评价，比如说mate7外观太丑了，或者说柜员语气很那是我们后面做的银行的一个例子，都很类似的，语气很不耐烦，它会有品牌就是mate7。你是mate7是吧？外观是这个产品的一个很重要的属性，对不对？

说话人2 01:21:35  
然后。

说话人1 01:21:37  
不好看是你的一个评价，你对他的一个情绪，对不对？我们做情感分析就是要把这东西给弄出来的，所以就说这个声音就代表了这个客户对你的产品的某一个方面的产品的这种飞车的一个情绪的判断，那有可能你很多这种客户的数据弄在一起，你就会形成消费者市场对你这个产品的一些态度，然后你是不是有可能真的要去看我的外观相对于我的竞品，如果很多人都说 Mate7的外观很糟糕，有可能我在下一代产品里面，我怎么样去提升我的外观的研发，或提供是吧？

或者我的摄像头，或者说我的电池发热等等之类，就类似这种东西，这里面都是用到我们后面一些分类的模式，但这里面还是挺复杂的，我们会谈到它用要构建一个很复杂的分类树，我们当时构建了大概是4个层级，总共是1000多个分类，就是你做分类的时候，你的分类目标，特别做文本的时候，你要预先去构建的，你不可能你给我个文本，你告诉我怎么去分类，你的分类目标是什么？

要人为去构建的，对这种很复杂的很大概一一千多个分类，去做基于架构去做相应的分类，对吧？这一些例子就是在一段话里面，我们可以把不同的这个观点等等之类的可以把它提取出来。

当时是投产之后还是帮到华为很多，而且华为当时。

说话人2 01:23:09  
要做。

说话人1 01:23:11  
多语言，因为当华为当时是在其他国家的德语英语，包括西班牙语它都有卖的，到了后来就没做了，原因你们应该很能猜得到对吧？因为SARS是个美国公司，因为他抓了孟晚舟之后，基本上美国公司跟华为之间的业务就断掉了。

最后一个讲一下ARP的重要意义，大家你可以去问一下大模型，我这里也是问的大模型，而且你看它左边叫大模型，会形成自己的思维链，就是cot他的思维链他怎么样的来回答你的问题，就像我刚才说了就一句话这个是公认的，不是我说的是公认的，就是自然语言处理是人工智能皇冠上的明珠，你就知道自然语言处理在整个人工智能里面大概是什么样的一个重要意义。

Ok最后的话我们就把语料库给大概快速过一下，你可以把语料库就看成一个数据库，就像我们做传统的结构化数据的继续学习的时候，我们有数据集对吧？我们有data赛，或者说我们有data base，语料库实际上就跟那个东西是类似的，我要有数据，只不过语料库这个概念，他最早就是没有自然语言处理这个东西的时候，别人语言学里面都已经在用语料在做，它是语言学过来的一个概念，就直接借用了，就仅此而已，没有什么特别的。

说话人2 01:24:49  
你。

说话人1 01:24:50  
这样大概去理解就。

说话人2 01:24:51  
好了，好吧？

说话人1 01:24:54  
为什么在ARP里面大家经常提到语料库，是因为对ARP来说的话，数据这个语料是越来越关键，特别是在预训练模型这个阶段，因为我们大家知道预训练模型像opa，他们基本上把我们人类的已知的所有的知识都已经消化掉了，你大概知道什么概念，就我没有数据可用了，我没有数据可用了，这是个很要命的问题，但是它的能力也只到了这样，当然已经很厉害了，但是基本上也是天花板我们能看得到的，对不对？所以如果我继续往下推进到ag或aasi的时候，其实里面很一个很重要的东西，就是我要去补充数据要补语料，对我们做这块是非常重要的。

常见的语料库的话，大家一看这个名字大概就知道就平衡语料库或者说随机语料库里面你针对一些特定的分类是不是它属于一种统计上的这种平衡的或者是随机的对吧？

实际上我们经常谈的语料库是通用语料库跟特定领域的语料库或者说专用语料库。

因为现在大模型，比如说我有可能跟大家一起做一个智能课堂的项目，我要把我讲课的东西，把我的教案，把我们的参考书把我们所有东西丢到这里面，

跟大家做一个数字教师，然后有可能服务大家日常的复习或者提问。

这种专业领域的知识有可能大模型是没有的，基座模型它是没有的。

所以我们在谈这个语料的时候，很多的时候就会说是有很多情况之下的缺损的是指我们叫专业领域的，专域的或者说领域的或者说专有的语料库，这个是这个不一定有的，或者说私域的我们叫私域的这个语料好吧，共识跟历史的话当然跟时间有关比较。

说话人2 01:27:01  
好好说，

说话人1 01:27:03  
语料库很关键，我这里就不多说了。

这里最主要就是说。

这个概念我要提一下，就数据标注，我大家有可能觉得我不知道你们有没有有没有了解过数据标注，这里有两个新闻，我希望大家去看一下，下面给了两个link的，一个是 AI虚虚假的语料可以去污染污染大模型的，另外一个就是 AI语料的迎来一个风口，很多上市公司就纷纷来做语料，实际上我想说的是说我们现在在新华学院学人工智能，坦率的说坦率的说，如果比如说我们毕业之后我们要就业，我们要直接去做人工智能相关的东西的话，比如算法工程师应该还是有蛮大一块距离的，不仅仅是你们其他的，比如原先我们经常合作的，像华师那些211或985那些学校的本科生毕业去做都有可能是有个概括的。

虽然很多的时候大家一开始做的时候，都围绕着一些边缘的任务切进去做的，这些任务有一些我们可以去做，而且可以很快在这个行业里去打开一条路的就是做标注，因为你有这种比如说我学了还至少是我们学了很多年的这种机器学习，就人工智能相关的一些核心的东西，我去再做一个特定领域的标注的时候，我有一些先天有些先天优势，大家一定要特别留意这样一个发展方向，而且现在是非常火，就我刚才说了，因为预训练模型已经发展到瓶颈了，其中造成瓶颈一个很大的原因就是数据的原因，数据不够用了，没数据了，我就要人为的去标注一些数据出来，去提升它的能力，这个是一个很大的东西，后面就是antkantk这个包你们之前用过没有？

上次你们之前学文本挖掘的时候用了art k吗？不知道。

Ok，这个是一个自然语言的工具包，它实际上什么都。

说话人5 01:29:11  
有的，自然语言。

说话人1 01:29:13  
核心的一些任务包括一些语料，当然它主要是英文的，但是我们主要用它来上传一些中文语料来学习它的函数。

大家课后可以到官网去看一下，它当然是英文的，但是我是非常鼓励大家去看这些开源软件的官网，因为这些官网里面有最新的一些信息，最新的一些future，我们只能去看官网，去看给哈巴的，它原始的官网才能捕捉得到的，你才能把它给用好的。

大家可以到他的官网去看一下他做文本的一些处理，包括一些语料库，我们主要用它的语料库当然也可以做一些基本的文本可视化。

当然有些函数我列在这儿了，但是这些函数的话我下面绿色的做了归纳，就是这些函数比如做搜索做对比什么那的，

这些函数很多，说实话有的时候我都不会记那么多函数，我当我用熟了我就知道，但我也很多年没有具体的扣定了，大家也不用记，你用多了自然就熟了，然后但是你要知道这个函数的一般的写法要有，这个要有是因为它毕竟你一个对象，你做这种对象的调用方法或者做一些其他的一些处理，它有你只要把一般的形式到时候用到之后去大概查一查，立马就你知道这个逻辑就可以去表达了，好吧？

最后一个简单的复习就是画这个图，大家可以去温习一下，就是我们这节课主要讲的概念，然后讲的语料讲的ltk的一些东西，然后有可能我40分钟真的很短，我要想想怎么去优化一下我的PPT，然后。

说话人2 01:30:56  
然后。

说话人1 01:30:57  
为我们下面讲具体的任务，然后去做一个基本的铺垫。作业的话大概是这些。

温习的作业刚才谈到的就是说大家有可能自己手提要装个环境，然后上nltk的官网去了解一下它的资源，然后另外的话就是我们书的第12章以及我指定的就是张琪那本书的第一章一定要去看，因为我讲的内容是依据于张琪的第一章，我会交叉的去用不同的参考书去讲的好吧？

提交的作业注意了，提交的作业因为我们有10%是日常考勤的，就基于这种提交的作业大家注意，第一个就是学情都不会为难大家学情调查表。

好，下周三之前林彤帮忙收一下好吧？另外一个就是我会给一份梁文峰的采访实录，你不用自己写读后感，你自己用大模型，我现在不会告诉你，反正你写一个500字，结合你对自然语言这门课的一些期望，写一个500字的读后感电子版交给我就好了，好吧？用大模型去写，

说话人2 01:32:11  
对，

说话人1 01:32:12  
我对，但是怎么样做提示词，ok我们下节课再去评判对吧？我也不是因为我就想知道你们现在用提示词大概是用了什么水平，好吧？你可以用你自己，你可以用gpck但gpc经常是连不上，你可以用Kimi，你可以用豆包都可以，没问题，好不好？最后几页大家自己去看一下好吧？

泰迪杯是a类比赛，我们已经在申请资金了，就有可能这次去应该是3月份报名，然后应该是4月份还是5月份就比赛，我想利用这个课程，有兴趣的我会搞一些队，然后我们去参加比赛好吧？它里面是有三个题，但是我是希望我们这个课程的同学参加的是自然语言处理的好吧？是下课了对吧？

Ok这节课就到这里，

说话人4 01:33:12  
然后。

说话人1 01:33:13  
作业怎么交，作业可以交到。

说话人5 01:33:17  
你可以发一个作业，然后。

说话人1 01:33:19  
我对这种不熟，你可以告诉我。

说话人5 01:33:21  
它上面你那边然后发布一个作业，然后大家就你那里可以设置时间，然后你就直接我们。

说话人2 01:33:30  
全部好好，

说话人1 01:33:34  
你们基本上都注册了，学习通了是吧？那就说我的课件也可以丢在里面，你可以去你们可以看到的对吧？

说话人5 01:33:41  
它有一个可以在资料里面上传，上面也可以说上传什么，

说话人2 01:33:46  
好好，我先关掉这个。

说话人1 01:33:51  
好谢谢都可以。