



CSRI

四川大学网络空间安全研究院

CyberSecurity Research Institute, Sichuan University

Spark 源码分析文档

版本号: v1.0

项目组: 大数据平台项目组

文档作者: 罗永刚

撰写日期: 2017 年 1 月 11 日

评审负责人: _____

评审日期: _____

四川大学网络空间安全研究院

目 录

第 1 章 带 English 的标题	1
1.1 封面相关	1
1.2 字体命令	1
1.3 表格样本	2
1.3.1 基本表格	2
1.3.2 复杂表格	3
1.3.3 其它	6
1.4 定理环境	7
1.5 参考文献	10
1.6 公式	11
第 2 章 中华人民共和国	13
2.1 其它例子	13
2.1.1 绘图	13
2.1.2 插图	13
第 3 章 Spark 应用启动过程分析	16
3.1 WordCount 程序	16
3.2 提交 WordCount 程序	17
3.3 准备 WordCount 提交程序运行环境	18
参考文献	20

第 1 章 带 English 的标题

这是 THUTHESIS 的示例文档，基本上覆盖了模板中所有格式的设置。建议大家在使用模板之前，除了阅读《THUTHESIS 用户手册》，这个示例文档也最好能看一看。

小老鼠偷吃热凉粉；短长虫环绕矮高粱^①。

1.1 封面相关

封面的例子请参看 `cover.tex`。主要符号表参看 `denation.tex`，附录和个人简历分别参看 `appendix01.tex` 和 `resume.tex`。里面的命令都很只管，一看即会^②。

1.2 字体命令

苏轼（1037-1101），北宋文学家、书画家。字子瞻，号东坡居士，眉州眉山（今属四川）人。苏洵子。嘉佑进士。神宗时曾任祠部员外郎，因反对王安石新法而求外职，任杭州通判，知密州、徐州、湖州。后以作诗“谤讟朝廷”罪贬黄州。哲宗时任翰林学士，曾出知杭州、颖州等，官至礼部尚书。后又贬谪惠州、儋州。北还后第二年病死常州。南宋时追谥文忠。与父洵弟辙，合称“三苏”。在政治上属于旧党，但也有改革弊政的要求。其文汪洋恣肆，明白畅达，为“唐宋八大家”之一。其诗清新豪健，善用夸张比喻，在艺术表现方面独具风格。少数诗篇也能反映民间疾苦，指责统治者的奢侈骄纵。词开豪放一派，对后代很有影响。《念奴娇·赤壁怀古》、《水调歌头·丙辰中秋》传诵甚广。

坡仙擅长行书、楷书，取法李邕、徐浩、颜真卿、杨凝式，而能自创新意。用笔丰腴跌宕，有天真烂漫之趣。与蔡襄、黄庭坚、米芾并称“宋四家”。能画竹，学文同，也喜作枯木怪石。论画主张“神似”，认为“论画以形似，见与儿童邻”；高度评价“诗中有画，画中有诗”的艺术造诣。诗文有《东坡七集》等。存世书迹有《答谢民师论文帖》、《祭黄几道文》、《前赤壁赋》、《黄州寒食诗帖》等。画迹有《枯木怪石图》、《竹石图》等。

① 韩愈（768-824），字退之，河南河阳（今河南孟县）人，自称郡望昌黎，世称韩昌黎。幼孤贫刻苦好学，德宗贞元八年进士。曾任监察御史，因上疏请免关中赋役，贬为阳山县令。后随宰相裴度平定淮西迁刑部侍郎，又因上表谏迎佛骨，贬潮州刺史。做过吏部侍郎，死谥文公，故世称韩吏部、韩文公。是唐代古文运动领袖，与柳宗元合称韩柳。诗力求险怪新奇，雄浑重气势。

② 你说还是看不懂？怎么会呢？

易与天地准，故能弥纶天地之道。仰以观於天文，俯以察於地理，是故知幽明之故。原始反终，故知死生之说。精气为物，游魂为变，是故知鬼神之情状。与天地相似，故不违。知周乎万物，而道济天下，故不过。旁行而不流，乐天知命，故不忧。安土敦乎仁，故能爱。范围天地之化而不过，曲成万物而不遗，通乎昼夜之道而知，故神无方而易无体。

有天地，然后万物生焉。盈天地之间者，唯万物，故受之以屯；屯者盈也，屯者物之始生也。物生必蒙，故受之以蒙；蒙者蒙也，物之穉也。物穉不可不养也，故受之以需；需者饮食之道也。饮食必有讼，故受之以讼。讼必有众起，故受之以师；师者众也。众必有所比，故受之以比；比者比也。比必有所畜也，故受之以小畜。物畜然后有礼，故受之以履。

履而泰，然后安，故受之以泰；泰者通也。物不可以终通，故受之以否。物不可以终否，故受之以同人。与人同者，物必归焉，故受之以大有。有大者不可以盈，故受之以谦。有大而能谦，必豫，故受之以豫。豫必有随，故受之以随。以喜随人者，必有事，故受之以蛊；蛊者事也。

有事而后大，故受之以临；临者大也。物大然后可观，故受之以观。可观而后有所合，故受之以噬嗑；嗑者合也。物不可以苟合而已，故受之以贲；贲者饰也。致饰然后亨，则尽矣，故受之以剥；剥者剥也。物不可以终尽，剥穷上反下，故受之以复。复则不妄矣，故受之以无妄。

有无妄然后可畜，故受之以大畜。物畜然后可养，故受之以颐；颐者养也。不养则不可动，故受之以大过。物不可以终过，故受之以坎；坎者陷也。陷必有所丽，故受之以离；离者丽也。

1.3 表格样本

1.3.1 基本表格

模板中关于表格的宏包有三个：`booktabs`、`array` 和 `longtabular`，命令有一个 `\hlinewd`。三线表可以用 `booktabs` 提供的 `\toprule`、`\midrule` 和 `\bottomrule`。它们与 `longtable` 能很好的配合使用。如果表格比较简单的话可以直接用命令 `\hlinewd{\width}` 控制。

首先来看一个最简单的表格。表 1.1 列举了本模板主要文件及其功能。请大家注意三线表中各条线对应的命令。这个例子还展示了如何在表格中正确使用脚注。由于 \LaTeX 本身不支持在表格中使用 `\footnote`，所以我们不得不将表格放在小页中，而且最好将表格的宽度设置为小页的宽度，这样脚注看起来才更美观。

表 1.1 模板文件。如果表格的标题很长，那么在表格索引中就会很不美观，所以要像 chapter 那样在前面用中括号写一个简短的标题。这个标题会出现在索引中。

文件名	描述
thuthesis.ins	L ^A T _E X 安装文件，DocStrip ^①
thuthesis.dtx	所有的一切都在这里 ^② 。
thuthesis.cls	模板类文件。
thuthesis.cfg	模板配置文。cls 和 cfg 由前两个文件生成。
thuthesis.bst	参考文献 BIB _T E _X 样式文件。
thuthesis.sty	常用的包和命令写在这里，减轻主文件的负担。

② 表格中的脚注

② 再来一个

1.3.2 复杂表格

我们经常会在表格下方标注数据来源，或者对表格里面的条目进行解释。前面的脚注是一种不错的方法，如果不喜欢脚注，可以在表格后面写注释，比如表 1.2。

表 1.2 复杂表格示例 1

x \ y	First Half		Second Half	
	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
East*	20.4	27.4	90	20.4
West**	30.6	38.6	34.6	31.6

注：数据来源《ThUthesis 使用手册》。

*：东部

**：西部

此外，表 1.2 同时还演示了另外两个功能：1) 通过 tabularx 的 |X| 扩展实现表格自动放大；2) 通过命令 \diagbox 在表头部分插入反斜线。

为了使我们的例子更接近实际情况，我会在必要的时候插入一些“无关”文字，以免太多图表同时出现，导致排版效果不太理想。第一个出场的当然是我的最爱：风流潇洒、骏马绝尘、健笔凌云的李太白了。

李白，字太白，陇西成纪人。凉武昭王暕九世孙。或曰山东人，或曰蜀人。白少有逸才，志气宏放，飘然有超世之心。初隐岷山，益州长史苏颋见而异之，曰：“是子天才英特，可比相如。”天宝初，至长安，往见贺知章。知章见其文，叹曰：“子

谪仙人也。”言于明皇，召见金銮殿，奏颂一篇。帝赐食，亲为调羹，有诏供奉翰林。白犹与酒徒饮于市，帝坐沉香亭子，意有所感，欲得白为乐章，召入，而白已醉。左右以水颰面，稍解，援笔成文，婉丽精切。帝爱其才，数宴见。白常侍帝，醉，使高力士脱靴。力士素贵，耻之，摘其诗以激杨贵妃。帝欲官白，妃辄沮止。白自知不为亲近所容，恳求还山。帝赐金放还。乃浪迹江湖，终日沉饮。永王璘都督江陵，辟为僚佐。璘谋乱，兵败，白坐长流夜郎，会赦得还。族人阳冰为当涂令，白往依之。代宗立，以左拾遗召，而白已卒。文宗时，诏以白歌诗、裴旻剑舞、张旭草书为三绝云。集三十卷。今编诗二十五卷。——《全唐诗》诗人小传

浮动体的并排放置一般有两种情况：1) 二者没有关系，为两个独立的浮动体；2) 二者隶属于同一个浮动体。对表格来说并排表格既可以像图 1.3、图 1.4 使用小页环境，也可以如图 1.5 使用子表格来做。图的例子参见第 2.1.2.2 节。

表 1.3 第一个并排子表格

111	222
222	333

表 1.4 第二个并排子表格

111	222
222	333

然后就是忧国忧民，诗家楷模杜工部了。杜甫，字子美，其先襄阳人，曾祖依艺为巩令，因居巩。甫天宝初应进士，不第。后献《三大礼赋》，明皇奇之，召试文章，授京兆府兵曹参军。安禄山陷京师，肃宗即位灵武，甫自贼中遁赴行在，拜左拾遗。以论救房琯，出为华州司功参军。关辅饥乱，寓居同州同谷县，身自负薪采椽，舖糒不给。久之，召补京兆府功曹，道阻不赴。严武镇成都，奏为参谋、检校工部员外郎，赐绯。武与甫世旧，待遇甚厚。乃于成都浣花里种竹植树，枕江结庐，纵酒啸歌其中。武卒，甫无所依，乃之东蜀就高适。既至而适卒。是岁，蜀师相攻杀，蜀大扰。甫携家避乱荆楚，扁舟下峡，未维舟而江陵亦乱。乃溯沿湘流，游衡山，寓居耒阳。卒年五十九。元和中，归葬偃师首阳山，元稹志其墓。天宝间，甫与李白齐名，时称李杜。然元稹之言曰：“李白壮浪纵恣，摆去拘束，诚亦差肩子美矣。至若铺陈终始，排比声韵，大或千言，次犹数百，词气豪迈，而风调清深，属对律切，而脱弃凡近，则李尚不能历其藩翰，况堂奥乎。”白居易亦云：“杜诗贯穿古今，尽工尽善，殆过于李。”元、白之论如此。盖其出处劳佚，喜乐悲愤，好贤恶恶，一见于诗。而又以忠君忧国、伤时念乱为本旨。读其诗可以知其世，故当时谓之“诗史”。旧集诗文共六十卷，今编诗十九卷。

不可否认 L^AT_EX 的表格功能没有想象中的那么强大，不过只要足够认真，足够细致，同样可以排出来非常复杂非常漂亮的表格。请参看表 1.6。

表 1.5 并排子表格

(a) 第一个子表格		(b) 第二个子表格	
111	222	111	222
222	333	222	333

表 1.6 复杂表格示例 2

Network Topology		# of nodes	# of clients			Server
GT-ITM	Waxman Transit-Stub	600	2%	10%	50%	Max. Connectivity
Inet-2.1		6000				
Xue	Rui	Ni	THUThESIS			
	ABCDEF					

最后就是清新飘逸、文约意赅、空谷绝响的王大俠了。王维，字摩诘，河东人。工书画，与弟缙俱有俊才。开元九年，进士擢第，调太乐丞。坐累为济州司仓参军，历右拾遗、监察御史、左补阙、库部郎中，拜吏部郎中。天宝末，为给事中。安禄山陷两都，维为贼所得，服药阳喑，拘于菩提寺。禄山宴凝碧池，维潜赋诗悲悼，闻于行在。贼平，陷贼官三等定罪，特原之，责授太子中允，迁中庶子、中书舍人。复拜给事中，转尚书右丞。维以诗名盛于开元、天宝间，宁薛诸王駉马豪贵之门，无不拂席迎之。得宋之问辋川别墅，山水绝胜，与道友裴迪，浮舟往来，弹琴赋诗，啸咏终日。笃于奉佛，晚年长斋禅诵。一日，忽索笔作书数纸，别弟缙及平生亲故，舍笔而卒。赠秘书监。宝应中，代宗问缙：“朕常于诸王坐闻维乐章，今存几何？”缙集诗六卷，文四卷，表上之。敕答云，卿伯氏位列先朝，名高希代。抗行周雅，长揖楚辞。诗家者流，时论归美。克成编录，叹息良深。殷璠谓维诗词秀调雅，意新理惬。在泉成珠，著壁成绘。苏轼亦云：“维诗中有画，画中有诗也。”今编诗四卷。

要想用好论文模板还是得提前学习一些 $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ 的相关知识，具备一些基本能力，掌握一些常见技巧，否则一旦遇到问题还真是比较麻烦。我们见过很多这样的同学，一直以来都是使用 Word 等字处理工具，以为 $\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ 模板的用法也应该类似，所以就沿袭同样的思路来对待这种所见非所得的排版工具，结果被折腾的焦头烂额，疲惫不堪。

如果您要排版的表格长度超过一页，那么推荐使用 `longtable` 或者 `supertabular` 宏包，模板对 `longtable` 进行了相应的设置，所以用起来可能简单一些。表 1.7 就是 `longtable` 的简单示例。

表 1.7 实验数据

测试程序	正常运行 时间 (s)	同步 时间 (s)	检查点 时间 (s)	卷回恢复 时间 (s)	进程迁移 时间 (s)	检查点 文件 (KB)
CG.A.2	23.05	0.002	0.116	0.035	0.589	32491
CG.A.4	15.06	0.003	0.067	0.021	0.351	18211
CG.A.8	13.38	0.004	0.072	0.023	0.210	9890
CG.B.2	867.45	0.002	0.864	0.232	3.256	228562
CG.B.4	501.61	0.003	0.438	0.136	2.075	123862
CG.B.8	384.65	0.004	0.457	0.108	1.235	63777
MG.A.2	112.27	0.002	0.846	0.237	3.930	236473
MG.A.4	59.84	0.003	0.442	0.128	2.070	123875
MG.A.8	31.38	0.003	0.476	0.114	1.041	60627
MG.B.2	526.28	0.002	0.821	0.238	4.176	236635
MG.B.4	280.11	0.003	0.432	0.130	1.706	123793
MG.B.8	148.29	0.003	0.442	0.116	0.893	60600
LU.A.2	2116.54	0.002	0.110	0.030	0.532	28754
LU.A.4	1102.50	0.002	0.069	0.017	0.255	14915
LU.A.8	574.47	0.003	0.067	0.016	0.192	8655
LU.B.2	9712.87	0.002	0.357	0.104	1.734	101975
LU.B.4	4757.80	0.003	0.190	0.056	0.808	53522
LU.B.8	2444.05	0.004	0.222	0.057	0.548	30134
EP.A.2	123.81	0.002	0.010	0.003	0.074	1834
EP.A.4	61.92	0.003	0.011	0.004	0.073	1743
EP.A.8	31.06	0.004	0.017	0.005	0.073	1661
EP.B.2	495.49	0.001	0.009	0.003	0.196	2011
EP.B.4	247.69	0.002	0.012	0.004	0.122	1663
EP.B.8	126.74	0.003	0.017	0.005	0.083	1656

1.3.3 其它

如果不想让某个表格或者图片出现在索引里面，请使用命令 `\caption*`。这个命令不会给表格编号，也就是出来的只有标题文字而没有“表 XX”，“图 XX”，否则索引里面序号不连续就显得不伦不类，这也是 \LaTeX 里星号命令默认的规则。

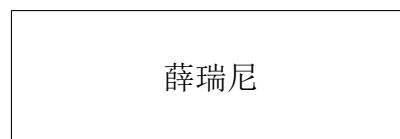
有这种需求的多是本科同学的英文资料翻译部分，如果觉得附录中英文原文中的表格和图片显示成“表”和“图”不协调的话，一个很好的办法就是用 `\caption*`，参数随便自己写，比如不守规矩的表 1.111 和图 1.111 能满足这种特殊需要（可以参看附录部分）。

如果的确想让它编号，但又不想让它出现在索引中的话，目前模板上不支持。

表 1.111 这是一个手动编号，不出现在索引中的表格。

THUThesis

Figure 1.111 这是一个手动编号，不出现在索引中的图。



最后，虽然大家不一定会独立使用小页，但是关于小页中的脚注还是有必要提一下。请看下面的例子。

柳宗元，字子厚（773-819），河东（今永济县）人^①，是唐代杰出的文学家，哲学家，同时也是一位政治改革家。与韩愈共同倡导唐代古文运动，并称韩柳^②。

① 山西永济水饺。

② 唐宋八大家之首二位。

唐朝安史之乱后，宦官专权，藩镇割据，土地兼并日渐严重，社会生产破坏严重，民不聊生。柳宗元对这种社会现实极为不满，他积极参加了王叔文领导的“永济革新”，并成为这一运动的中坚人物。他们革除弊政，打击权奸，触犯了宦官和官僚贵族利益，在他们的联合反扑下，改革失败了，柳宗元被贬为永州司马。

1.4 定理环境

给大家演示一下各种和证明有关的环境：

假设 1.1：待月西厢下，迎风户半开；隔墙花影动，疑是玉人来。

$$c = a^2 - b^2 \quad (1-1)$$

$$= (a + b)(a - b) \quad (1-2)$$

千辛万苦，历尽艰难，得有今日。然相从数千里，未曾哀戚。今将渡江，方图百年欢笑，如何反起悲伤？（引自《杜十娘怒沉百宝箱》）

定义 1.1：子曰：「道千乘之国，敬事而信，节用而爱人，使民以时。」

千古第一定义！问世间、情为何物，只教生死相许？天南地北双飞客，老翅几回寒暑。欢乐趣，离别苦，就中更有痴儿女。君应有语，渺万里层云，千山暮雪，只影向谁去？

横汾路，寂寞当年箫鼓，荒烟依旧平楚。招魂楚些何嗟及，山鬼暗啼风雨。天也妒，未信与，莺儿燕子俱黄土。千秋万古，为留待骚人，狂歌痛饮，来访雁丘处。

命题 1.1: 曾子曰：「吾日三省吾身——为人谋而不忠乎？与朋友交而不信乎？传不习乎？」

多么凄美的命题啊！其日牛马嘶，新妇入青庐，奄奄黄昏后，寂寂人定初，我命绝今日，魂去尸长留，揽裙脱丝履，举身赴清池，府吏闻此事，心知长别离，徘徊庭树下，自挂东南枝。

注释 1.1: 天不言自高，水不言自流。

$$\begin{aligned}\varphi(x, z) &= z - \gamma_{10}x - \gamma_{mn}x^m z^n \\ &= z - Mr^{-1}x - Mr^{-(m+n)}x^m z^n \\ \zeta^0 &= (\xi^0)^2, & (1-3) \\ \zeta^1 &= \xi^0 \xi^1, & (1-4) \\ \zeta^2 &= (\xi^1)^2, & (1-5)\end{aligned}$$

天尊地卑，乾坤定矣。卑高以陈，贵贱位矣。动静有常，刚柔断矣。方以类聚，物以群分，吉凶生矣。在天成象，在地成形，变化见矣。鼓之以雷霆，润之以风雨，日月运行，一寒一暑，乾道成男，坤道成女。乾知大始，坤作成物。乾以易知，坤以简能。易则易知，简则易从。易知则有亲，易从则有功。有亲则可久，有功则可大。可久则贤人之德，可大则贤人之业。易简，而天下矣之理矣；天下之理得，而成位乎其中矣。

公理 1.1: 两点间直线段距离最短。

$$\begin{aligned}x &\equiv y + 1 \pmod{m^2} & (1-6) \\ x &\equiv y + 1 \pmod{m^2} & (1-7) \\ x &\equiv y + 1 \pmod{m^2} & (1-8)\end{aligned}$$

《象曰》：大哉乾元，万物资始，乃统天。云行雨施，品物流形。大明始终，六位时成，时乘六龙以御天。乾道变化，各正性命，保合大和，乃利贞。首出庶物，万国咸宁。

《象曰》：天行健，君子以自强不息。潜龙勿用，阳在下也。见龙再田，德施普也。终日乾乾，反复道也。或跃在渊，进无咎也。飞龙在天，大人造也。亢龙有悔，盈不可久也。用九，天德不可为首也。

引理 1.1: 《猫和老鼠》是我最爱看的动画片。

$$\begin{aligned} \int_a^b \left\{ \int_a^b [f(x)^2 g(y)^2 + f(y)^2 g(x)^2] - 2f(x)g(x)f(y)g(y) dx \right\} dy \\ = \int_a^b \left\{ g(y)^2 \int_a^b f^2 + f(y)^2 \int_a^b g^2 - 2f(y)g(y) \int_a^b fg \right\} dy \end{aligned}$$

行行重行行，与君生别离。相去万余里，各在天一涯。道路阻且长，会面安可知。胡马依北风，越鸟巢南枝。相去日已远，衣带日已缓。浮云蔽白日，游子不顾返。思君令人老，岁月忽已晚。弃捐勿复道，努力加餐饭。

定理 1.1: 犯我强汉者，虽远必诛 ——陈汤（汉）

$$y = 1 \quad (1-9a)$$

$$y = 0 \quad (1-9b)$$

道可道，非常道。名可名，非常名。无名天地之始；有名万物之母。故常无，欲以观其妙；常有，欲以观其徼。此两者，同出而异名，同谓之玄。玄之又玄，众妙之门。上善若水。水善利万物而不争，处众人之所恶，故几于道。曲则全，枉则直，洼则盈，敝则新，少则多，多则惑。人法地，地法天，天法道，道法自然。知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者寿。

证明 燕赵古称多感慨悲歌之士。董生举进士，连不得志于有司，怀抱利器，郁郁适兹土，吾知其必有合也。董生勉乎哉？

夫以子之不遇时，苟慕义强仁者，皆爱惜焉，矧燕、赵之士出乎其性者哉！然吾尝闻风俗与化移易，吾恶知其今不异于古所云邪？聊以吾子之行卜之也。董生勉乎哉？

吾因子有所感矣。为我吊望诸君之墓，而观于其市，复有昔时屠狗者乎？为我谢曰：“明天子在上，可以出而仕矣！” ——韩愈《送董邵南序》 □

推论 1.1: 四川话配音的《猫和老鼠》是世界上最好看最好听最有趣的动画片。

$$V_i = v_i - q_i v_j, \quad X_i = x_i - q_i x_j, \quad U_i = u_i, \quad \text{for } i \neq j; \quad (1-10)$$

$$V_j = v_j, \quad X_j = x_j, \quad U_j u_j + \sum_{i \neq j} q_i u_i. \quad (1-11)$$

迢迢牵牛星，皎皎河汉女。纤纤擢素手，札札弄机杼。终日不成章，泣涕零如雨。河汉清且浅，相去复几许。盈盈一水间，脉脉不得语。

例 1.1：大家来看这个例子。

$$\begin{cases} \nabla f(\mathbf{x}^*) - \sum_{j=1}^p \lambda_j \nabla g_j(\mathbf{x}^*) = 0 \\ \lambda_j g_j(\mathbf{x}^*) = 0, \quad j = 1, 2, \dots, p \\ \lambda_j \geq 0, \quad j = 1, 2, \dots, p. \end{cases} \quad (1-12)$$

练习 1.1：清列出 Andrew S. Tanenbaum 和 W. Richard Stevens 的所有著作。

猜想 1.1： *Poincare Conjecture* If in a closed three-dimensional space, any closed curves can shrink to a point continuously, this space can be deformed to a sphere.

问题 1.1：回答还是不回答，是个问题。

如何引用定理 1.1 呢？加上 \label 使用 \ref 即可。妾发初覆额，折花门前剧。郎骑竹马来，绕床弄青梅。同居长干里，两小无嫌猜。十四为君妇，羞颜未尝开。低头向暗壁，千唤不一回。十五始展眉，愿同尘与灰。常存抱柱信，岂上望夫台。十六君远行，瞿塘滟滪堆。五月不可触，猿声天上哀。门前迟行迹，一一生绿苔。苔深不能扫，落叶秋风早。八月蝴蝶来，双飞西园草。感此伤妾心，坐愁红颜老。

1.5 参考文献

当然参考文献可以直接写 \bibitem，虽然费点功夫，但是好控制，各种格式可以自己随意改写。

本模板推荐使用 BIB_TE_X，样式文件为 thuthesis.bst，基本符合学校的参考文献格式（如专利等引用未加详细测试）。看看这个例子，关于书的^[1-3]，还有这些^[4-6]，关于杂志的^[7-9]，硕士论文^[10,11]，博士论文^[12,13]，标准文件^[14]，会议论文^[15,16]，技术报告^[17]，电子文献^[18,19]。中文参考文献^[20]应增加 lang="zh" 字段，以便进行相应处理。另外，本模板对中文文献^[21]的支持并不是十全十美，如果有不如意的地方，请手动修改 bbl 文件。

有时候不想要上标，那么可以这样 [12]，这个非常重要。

有时候一些参考文献没有纸质出处，需要标注 URL。缺省情况下，URL 不会在连字符处断行，这可能使得用连字符代替空格的网址分行很难看。如果需要，可以将模板类文件中

`\RequirePackage{hyperref}`

一行改为:

`\PassOptionsToPackage{hyphens}{url}`

`\RequirePackage{hyperref}`

使得连字符处可以断行。更多设置可以参考 `url` 宏包文档。

1.6 公式

贝叶斯公式如式 (1-13)，其中 $p(y|\mathbf{x})$ 为后验； $p(\mathbf{x})$ 为先验；分母 $p(\mathbf{x})$ 为归一化因子。

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \quad (1-13)$$

论文里面公式越多， \TeX 就越 happy。再看一个 `amsmath` 的例子：

$$\det \mathbf{K}(t = 1, t_1, \dots, t_n) = \sum_{I \in \mathbf{n}} (-1)^{|I|} \prod_{i \in I} t_i \prod_{j \in I} (D_j + \lambda_j t_j) \det \mathbf{A}^{(\lambda)}(\bar{I}|\bar{I}) = 0. \quad (1-14)$$

前面定理示例部分列举了很多公式环境，可以说把常见的情况都覆盖了，大家在写公式的时候一定要好好看 `amsmath` 的文档，并参考模板中的用法：

$$\begin{aligned} \int_a^b \left\{ \int_a^b [f(x)^2 g(y)^2 + f(y)^2 g(x)^2] - 2f(x)g(x)f(y)g(y) dx \right\} dy \\ = \int_a^b \left\{ g(y)^2 \int_a^b f^2 + f(y)^2 \int_a^b g^2 - 2f(y)g(y) \int_a^b fg \right\} dy \end{aligned}$$

其实还可以看看这个多级规划：

$$\left\{ \begin{array}{l} \max_{\mathbf{x}} F(x, y_1^*, y_2^*, \dots, y_m^*) \\ \text{subject to:} \\ G(x) \leq 0 \\ (y_1^*, y_2^*, \dots, y_m^*) \text{ solves problems } (i = 1, 2, \dots, m) \\ \left\{ \begin{array}{l} \max_{y_i} f_i(x, y_1, y_2, \dots, y_m) \\ \text{subject to:} \\ g_i(x, y_1, y_2, \dots, y_m) \leq 0. \end{array} \right. \end{array} \right. \quad (1-15)$$

这些跟规划相关的公式都来自于刘宝碇老师《不确定规划》的课件。

第2章 中华人民共和国

2.1 其它例子

在第1章中我们学习了贝叶斯公式 (1-13)，这里我们复习一下：

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \quad (2-1)$$

2.1.1 绘图

本模板不再预先装载任何绘图包（如 `pstricks`，`pgf` 等），完全由用户来决定。个人觉得 `pgf` 不错，不依赖于 `Postscript`。此外还有很多针对 `LaTeX` 的 GUI 作图工具，如 `XFig(jFig)`, `WinFig`, `Tpx`, `Ipe`, `Dia`, `Inkscape`, `LaTeXPiX`, `jPicEdt`, `jaxdraw` 等等。

2.1.2 插图

强烈推荐《`LaTeX 2ε` 插图指南》！关于子图形的使用细节请参看 `subcaption` 宏包的说明文档。

2.1.2.1 一个图形

一般图形都是处在浮动环境中。之所以称为浮动是指最终排版效果图形的位置不一定与源文件中的位置对应^①，这也是刚使用 `LaTeX` 同学可能遇到的问题。如果要强制固定浮动图形的位置，请使用 `float` 宏包，它提供了 `[H]` 参数，比如图 2.1。



图 2.1 利用 `Xfig` 制图

大学之道，在明明德，在亲民，在止于至善。知止而后有定；定而后能静；静而后能安；安而后能虑；虑而后能得。物有本末，事有终始。知所先后，则近道矣。古之欲明明德于天下者，先治其国；欲治其国者，先齐其家；欲齐其家者，先修其身；欲修其身者，先正其心；欲正其心者，先诚其意；欲诚其意者，先致其知；

^① This is not a bug, but a feature of `LaTeX`!

致知在格物。物格而后知至；知至而后意诚；意诚而后心正；心正而后身修；身修而后家齐；家齐而后国治；国治而后天下平。自天子以至于庶人，壹是皆以修身为本。其本乱而未治者否矣。其所厚者薄，而其所薄者厚，未之有也！

——《大学》

2.1.2.2 多个图形

如果多个图形相互独立，并不共用一个图形计数器，那么用 `minipage` 或者 `parbox` 就可以。否则，请参看图 2.2，它包含两个小图，分别是图 2.2(a)和图 2.2(b)。推荐使用 `\subcaptionbox`，因为可以像图 2.2 那样对齐子图的标题，也可以使用 `subcaption` 宏包的 `\subcaption`（放在 `minipage` 中，用法同 `\caption`）或是 `subfigure`、`subtable` 环境，像图 2.3，不要再用 `\subfloat`、`\subfigure` 和 `\subtable`。



(a) 第一个小图形



(b) 第二个小图形，注意这个图略矮些。如果标题很长的话，它会自动换行

图 2.2 包含子图形的大图形 (`subcaptionbox` 示例)



(a) 第一个小图形



(b) 第二个小图形，注意这个图略矮些。`subfigure` 中同一行的子图在顶端对齐。

图 2.3 包含子图形的大图形 (`subfigure` 示例)

古之学者必有师。师者，所以传道受业解惑也。人非生而知之者，孰能无惑？惑而不从师，其为惑也，终不解矣。生乎吾前，其闻道也固先乎吾，吾从而师之；生乎吾後，其闻道也亦先乎吾，吾从而师之。吾师道也，夫庸知其年之先後生於吾乎！是故无贵无贱无长无少，道之所存，师之所存也。

嗟乎！师道之不传也久矣，欲人之无惑也难矣。古之圣人，其出人也远矣，犹且从师而问焉；今之众人，其下圣人也亦远矣，而耻学於师。是故圣益圣，愚益愚。



图 2.4 并排第一个图



图 2.5 并排第二个图

圣人之所以为圣，愚人之所以为愚，其皆出於此乎？爱其子，择师而教之，於其身也，则耻师焉，惑焉。彼童子之师，授之书而习其句读者，非吾所谓传其道、解其惑者也。句读之不知，惑之不解，或师焉，或不焉，小学而大遗，吾未见其明也。巫医、乐师、百工之人不耻相师，士大夫之族曰“师”曰“弟子”之云者，则群聚而笑之。问之，则曰：彼与彼年相若也，道相似也，位卑则足羞，官盛则近谀。呜呼！师道之不复，可知矣。巫医、乐师、百工之人。吾子不齿，今其智乃反不能及，其可怪也欤！圣人无常师。孔子师郯子、苌子、师襄、老聃。郯子之徒，其贤不及孔子。孔子曰：“三人行，必有我师。”是故弟子不必不如师，师不必贤於弟子。闻道有先後，术业有专攻，如是而已。

如果要把编号的两个图形并排，那么小页就非常有用了：

李氏子蟠，年十七，好古文、六艺，经传皆通习之，不拘於时，学於余。余嘉其能行古道，作师说以贻之。

——韩愈（唐）

第3章 Spark 应用启动过程分析

3.1 WordCount 程序

阅读 Spark 源码以 Spark 集群中运行 WordCount 应用顺序为根据,理解出 Spark 应用运行的流程,从而达到阅读源码的目的。

本 WordCount 源码如程序3.1所示:

```
1 package org.apache.spark.examples
2 import org.apache.spark.{SparkConf, SparkContext}
3 object WordCount{
4     def main(args: Array[String]): Unit = {
5         var splits = 2
6         val conf = new SparkConf().setAppName("BigDataBench WordCount")
7         val spark = new SparkContext(conf)
8         val filename = args(0)
9         val save_file = args(1)
10        if(args.length > 2) splits = args(2).toInt
11
12        val lines = spark.textFile(filename, splits)
13        val words = lines.flatMap(line => line.split(" "))
14        val words_map = words.map( word => (word, 1))
15        val result = words_map.reduceByKey(_ + _)
16
17        result.saveAsTextFile(save_file)
18        println("Result has been save to: " + save_file)
19    }
20 }
```

程序 3.1 基于 Spark 的 WordCount 程序

将输入案例 input.txt 放入 hdfs 中,已搭建好环境的 Spark 集群,打包好的 WordCount jar 包且提交到集群中的任意一台机器(本例中为 master)中。

3.2 提交 WordCount 程序

在 master 中运行提交命令如下：

```
Shell 命令
/usr/bin/spark-submit
HelloWorld-1.0-SNAPSHOT-jar-with-dependencies.jar
/user/inputfile.txt 2
```

从这条命令可以看出使用的是 spark-submit shell 脚本，后面三个为参数，第一个为目录下的 WordCount jar 包，第二个为 hdfs 中的输入文件，第三个为 WordCount 中限定输出的 value 临界值。我们打开 spark-submit shell 脚本，脚本如所示。

spark-submit shell 脚本内容如程序3.2所示。

```
1 SPARK_HOME="$(cd "`dirname $(readlink -nf "$0")"`/..; pwd -P)"
2 export PYTHONHASHSEED=0
3 exec "SPARK_HOME"/bin/spark-class org.apache.spark.deploy.SparkSubmit "$@
  ↪ "
```

程序 3.2 spark-submit.sh 脚本

从程序3.2可以看出,spark-submit 执行流程是:首先找出 Spark 安装的根目录,然后再去执行 spark-class 这个 shell 脚本,参数为 org.apache.spark.deploy.SparkSubmit 以及提交命令后跟的所有参数。

spark-class 脚本先找到 master 的 java, 遍历 Spark 安装目录下 lib 目录下的 jar 包, 加入到 CLASSPATH 中, 具体内容如下:

spark-class 的 CLASSPATH 值

```
/usr/hdp/current/spark-historyserver/conf/
/usr/hdp/2.4.0.0-169/spark/lib/spark-assembly-1.6.0.2.4.0.0-169-
hadoop2.7.1.2.4.0.0-169.jar
/usr/hdp/2.4.0.0-169/spark/lib/datanucleus-api-jdo-3.2.6.jar
/usr/hdp/2.4.0.0-169/spark/lib/datanucleus-core-3.2.10.jar
/usr/hdp/2.4.0.0-169/spark/lib/datanucleus-rdbms-3.2.9.jar
/usr/hdp/current/hadoop-client/conf/
```

spark-class 最后通过 Java -cp <classpath> org.apache.spark.launcher.Main 参数列表, 开启 jvm 进程, 执行的主类为 org.apache.spark.launcher.Main, 传过来的参数有:

— org.apache.spark.deploy.SparkSubmi

- HelloWorld-1.0-SNAPSHOT-jar-with-dependencies.jar
- /user/inputfile.txt

最后启动的 Java 程序的命令如下：

spark-class 创建的 Java 虚拟机命令

```
java -Dhdp.version=2.4.0.0-169
-cp <classpath>
-Xms1g -Xmx1g
org.apache.spark.deploy.SparkSubmit
-master yarn
--deploy-mode client
HelloWorld-1.0-SNAPSHOT-jar-with-dependencies.jar
/user/inputfile.txt 2
```

3.3 准备 WordCount 提交程序运行环境

org.apache.spark.launcher.Main 代码片段如程序3.3所示。传入此类的 main 方法的参数有：

- org.apache.spark.deploy.SparkSubmit
- --master yarn-client
- HelloWorld-1.0-SNAPSHOT-jar-with-dependencies.jar
- /user/inputfile.txt

代码通过 SparkSubmitCommandBuilder 方法生成相应的 Java 命令，然后返回给 cmd，最后程序判断运行的主机是 Windows 还是 bash，打印到控制台，之后 spark-class 脚本重定位到这里并读取输出的字符流^①，通过 exec 执行 java 命令。

```
1 public static void main(String[] argsArray) throws Exception {
2     List<String> args = new ArrayList<>(Arrays.asList(argsArray));
3     String className = args.remove(0);
4     AbstractCommandBuilder builder;
5     if (className.equals("org.apache.spark.deploy.SparkSubmit")) {
6         try {
7             builder = new SparkSubmitCommandBuilder(args);
```

^① launcher.Main 函数在 spark-class 脚本中执行，spark-class 脚本会读取 Main 的标准输出，将标准输出解析为启动程序命令。

```
8   } catch (IllegalArgumentException e) { ... }
9
10  Map<String, String> env = new HashMap<>();
11  List<String> cmd = builder.buildCommand(env);
12  if (isWindows()) {
13      System.out.println(prepareWindowsCommand(cmd, env));
14  } else {
15      // In bash, use NULL as the arg separator since it cannot be used in
16      //    ↪ an argument.
17      List<String> bashCmd = prepareBashCommand(cmd, env);
18      for (String c : bashCmd) {
19          System.out.print(c);
20          System.out.print('\0');
21      }
22  }
```

程序 3.3 launcher.Main 程序代码片段

参考文献

- [1] Knuth D E. The \TeX Book. 15th ed., Reading, MA: Addison-Wesley Publishing Company, 1989.
- [2] Goosens M, Mittelbach F, Samarin A. The \LaTeX Companion. Reading, MA: Addison-Wesley Publishing Company, 1994: 112–125.
- [3] Gröning P, Nilsson L, Ruffieux P, et al. Encyclopedia of Nanoscience and Nanotechnology, volume 1. American Scientific Publishers, 2004: 547–579.
- [4] Krasnogor N. Towards robust memetic algorithms. In: Hart W, Krasnogor N, Smith J, (eds.). Recent Advances in Memetic Algorithms. New York: Springer Berlin Heidelberg, 2004: 185–207.
- [5] 阎真. 沧浪之水. 人民文学出版社, 2001: 185–207.
- [6] 班固. 苏武传. 见: 郑在瀛, 汪超宏, 周文复, 编. 传记散文英华. 武汉: 湖北人民出版社, 1998: 65–69.
- [7] Chafik El Idrissi M, Roney A, Frigon C, et al. Measurements of total kinetic-energy released to the $N = 2$ dissociation limit of H_2 — evidence of the dissociation of very high vibrational Rydberg states of H_2 by doubly-excited states. Chemical Physics Letters, 1994, 224(10):260–266.
- [8] Mellinger A, Vidal C R, Jungen C. Laser reduced fluorescence study of the carbon-monoxide nd triplet Rydberg series-experimental results and multichannel quantum-defect analysis. J. Chem. Phys., 1996, 104(5):8913–8921.
- [9] Shell M. How to use the IEEEtran \LaTeX class. Journal of \LaTeX Class Files, 2002, 12(4):100–120.
- [10] 猪八戒. 论流体食物的持久保存 [硕士学位论文]. 北京: 广寒宫大学, 2005.
- [11] Jeyakumar A R. Metamori: A library for incremental file checkpointing[D]. Blacksburg: Virgina Tech, June 21, 2004.
- [12] 沙和尚. 论流沙河的综合治理 [博士学位论文]. 北京: 清华大学, 2005.
- [13] Zadok E. FiST: A System for Stackable File System Code Generation[D]. USA: Computer Science Department, Columbia University, May, 2001.
- [14] IEEE Std 1363-2000. IEEE Standard Specifications for Public-Key Cryptography. New York: IEEE, 2000.
- [15] Kim S, Woo N, Yeom H Y, et al. Design and Implementation of Dynamic Process Management for Grid-enabled MPICH. the 10th European PVM/MPI Users' Group Conference, Venice, Italy, 2003.
- [16] Kocher C, Jaffe J, Jun B. Differential power analysis. In: Wiener M, (eds.). Advances in Cryptology (CRYPTO '99), volume 1666 of *Lecture Notes in Computer Science*. Springer-Verlag, 1999. 388–397.

- [17] Woo A, Bailey D, Yarrow M, et al. The NAS parallel benchmarks 2.0. Technical report, The Pennsylvania State University CiteSeer Archives, December 05, 1995. <http://www.nasa.org/>.
- [18] 萧钰. 出版业信息化迈入快车道 [EB/OL]. (2001-12-19)[2002-04-15]. <http://www.creader.com/news/200112190019.htm>.
- [19] Online Computer Library Center, Inc. History of OCLC[EB/OL]. [2000-01-08]. <http://www.oclc.org/about/history/default.htm>.
- [20] 贾宝玉, 林黛玉, 薛宝钗, 等. 论刘姥姥食量大如牛之现实意义. 红楼梦杂谈, 1800, 224:260–266.
- [21] 王重阳, 黄药师, 欧阳峰, 等. 武林高手从入门到精通. 第 N 次华山论剑, 西安, 中国, 2006.