Natural Language Processing

自然语言处理

教师: 陈 震

单位: 清华大学基础工业训练中心

自然语言处理

- 自然语言处理

- 主题模型

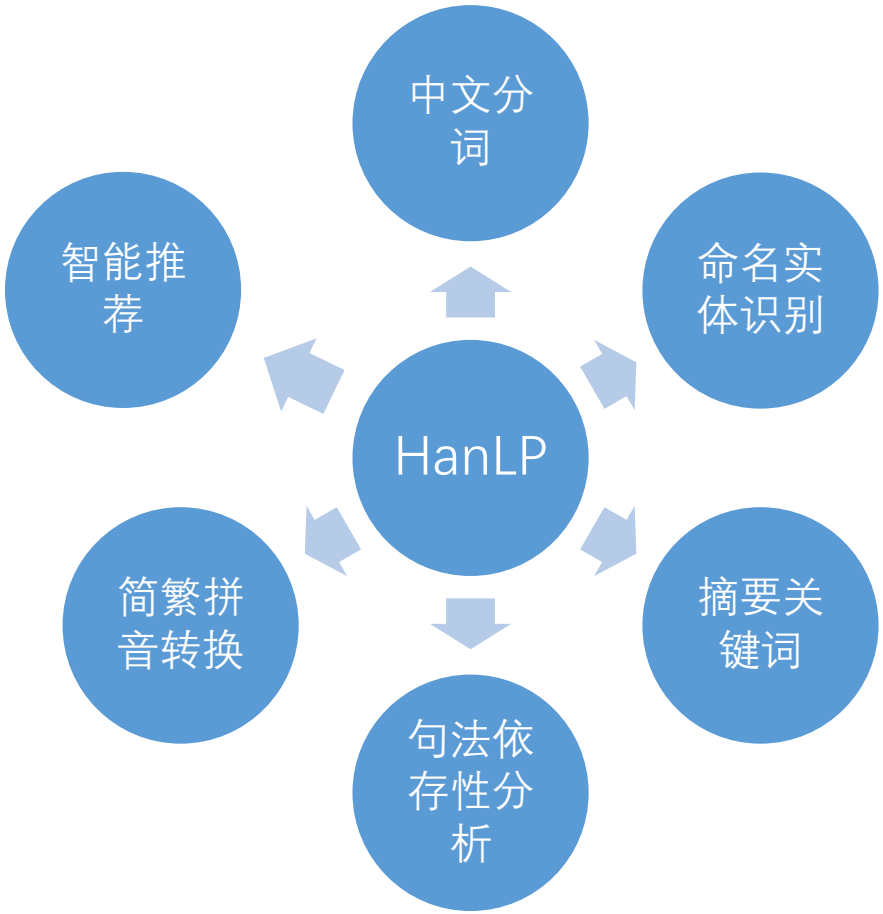
自然语言处理

HanLP自然语言处理包

自然语言处理

- 中文分词 （ chinese-word-segmentation ）
- 词性标注 （Part-of-Speech tagging）
- 命名实体识别 （Named Entity Recognizer）
- 依存句法分析 （Dependency parsing ）
- 关键词提取 （Keyword extraction）
- 新词发现 （New Word Discover）
- 短语提取 （Phrase extraction）
- 自动摘要 （Text summarization）
- 文本分类（Text classification ）

HanLP自然语言处理包——框架



https://github.com/hankcs/HanLP

中文分词功能（静态分词器）

- 标准分词（最常用）
- NLP分词（命名实体识别和词性标注）
- 索引分词（面向搜索引擎）
- 繁体分词
- 极速词典分词（速度快，精度一般）
- N-最短路径分词（精度高，速度慢）
- CRF分词（新词识别，无法利用自定义词典）

中文分词功能（静态分词器）

- 例子：

  标准分词：

  ```
  List<Term> termList = HanLP.segment("商品和服务");
  System.out.println(termList);
  ```

  索引分词：

  ```
  List<Term> termList = IndexTokenizer.segment("主副食品");
  ```

- HanLP, IndexTokenize都是类名

中文分词功能（动态分词器）

- Viterbi算法
- Dijkstra算法
- 基于CRF的分词器
- 例子：

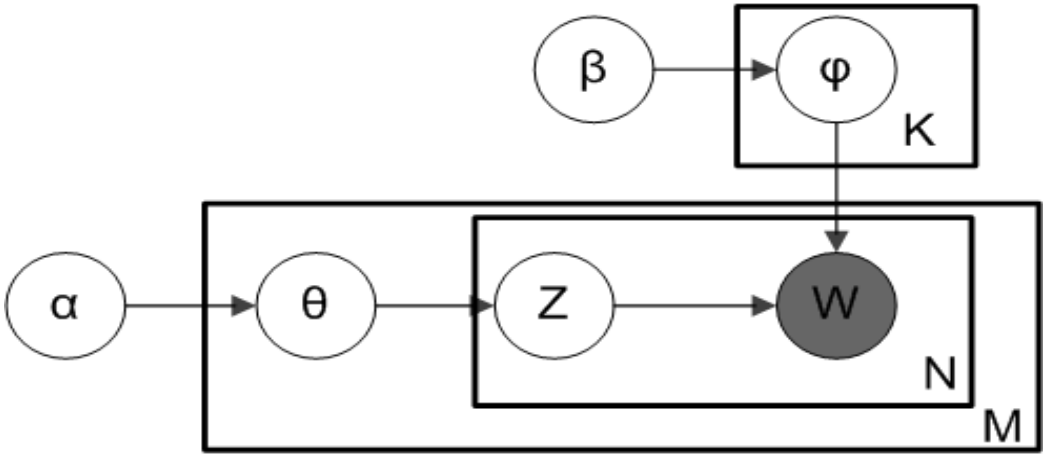
```
Segment segment = new DijkstraSegment();
List<Term> termList = segment.seg(“商品和服务”)
```

主题模型（topic model）

LDA算法

算法原理

- 主题模型（Topic Model）
- 三层贝叶斯概率分布（词语、文档、主题）
- 大量语料训练，对新语料作主题分类



算法支持——GibbsLDA++

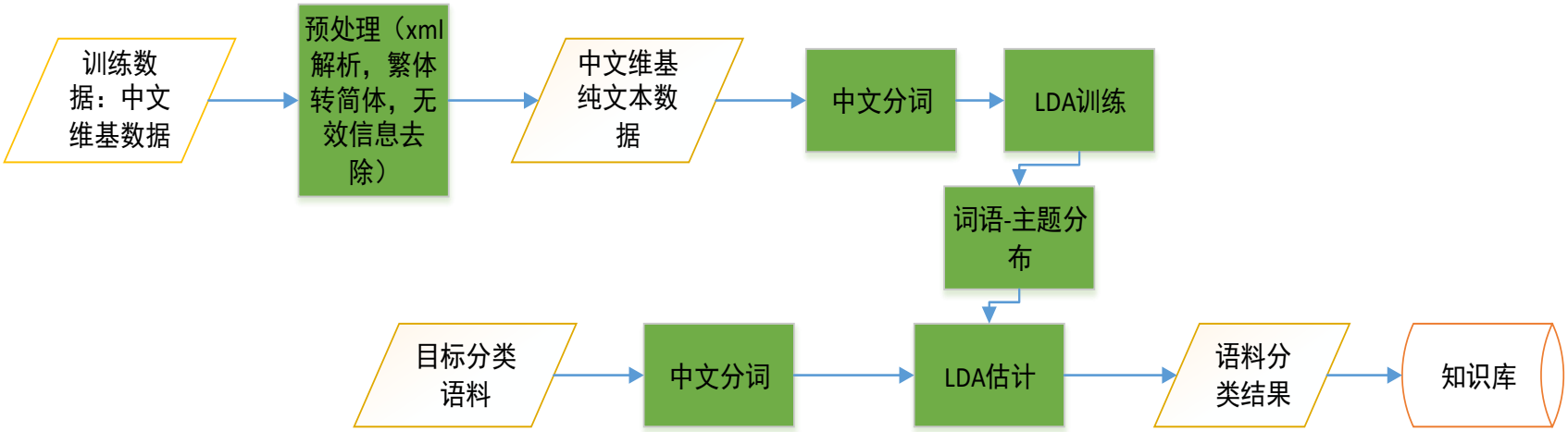
- 官网： http://gibbslda.sourceforge.net/
- 使用：
  - 从原始数据中进行参数估计
  - 利用已有模型进行参数估计
  - 利用已有模型进行预测

算法支持——GibbsLDA++

- 生成文件：
  - <model_name>.others：
    包含LDA参数信息
  - <model_name>.phi：
    包含词语-主题分布的信息（$p(word_w|topic_t)$）
  - <model_name>.theta：
    包含主题-文档分布的信息（$p(topic_t|document_m)$）
  - <model_name>.tassign：
    包含每个主题类的代表性的词语
  - <model_name>.twords：
    包含每个词语的最有可能的主题

算法示例

- 训练语料：中文wiki数据集，大约105万个文档
- 中文分词工具：HanLP中文分词包
- 预计分类数目：500
- 测试语料：清华校内的景点介绍



算法示例

Topic 496th:

三角形 0.004316381521306001
镶嵌 0.0033541007201961914
多面体 0.002584276079308344
顶点 0.002499090565767476
三角 0.002013217636682523
多边形 0.001962737332362008
立方体 0.0018333815525406897
截角 0.001435849156016637 2
正方形 0.0014106090038563796
六边形 0.001353818661495801
底面 0.0010004565312521987
八面体 9.941464932121343E-4
堆砌 9.878364551720701E-4
几何学 9.089609796712661E-4
边形 9.026509416312017E-4
斜方 7.985353139701403E-4
四面体 7.606750857297543E-4
面体 7.385899525895292E-4
截半 6.849546292489824E-4
菱形 6.660245151287894E-4
半正 6.628694961087573E-4
四角 6.376293439485E-4
对偶 6.344743249284678E-4
五角 6.186992298283 07E-4
边长 6.155442108082749E-4
星形 5.934590776680497E-4
正四面体 5.839940206079533E-
施莱夫利 5.429787733475352E-
十二面体 5.240486592273421E-
密铺 4.893434500069883E-4
五边形 4.830334119669 24E-4
算经 4.293980886263773E-4
梯形 3.915378603859913E-4
二十面 3.8207280332589486E-4
詹森 3.789177843058627E-4
四边形 3.789177843058627E-4
帐塔 3.694527272457662E-4
均匀 3.599876701856697E-4
锥柱 3.5683265116563754E-4
扭棱 3.536776321456054E-4

Topic 280th:

芬兰 0.011597692273984356
亚美尼亚 0.008347910627194 2
土耳其 0.00798591652415984
阿富汗 0.007645696502510959
伊朗 0.0051416771431752
波斯 0.004839561763950994
阿塞拜疆 0.004678977913732 7
格鲁吉亚 0.004434019498145 5
王朝 0.0036174914461882157
安哥拉 0.002817293955270048
高加索 0.002430804010676921
赫尔辛基 0.001796633890323 4
共和国 0.001674154682529810
亚美尼亚人 0.00148907499075
安息 0.0013529869820932679
波斯语 0.00125500361585839
土库曼 0.001167907290316276
乔治亚 0.001151576729277130
萨珊 0.0011406896885843665
巴库 0.001053593363042253
喀布尔 9.773840781929037E-4
亚塞 9.25670634902274E-4
拜然 9.147835942095098E-4
阿布哈兹 8.821224721312172E
里海 8.494613500529247E-4
中亚 8.440178297065426E-4
芬兰语 7.977479067622949E-4
黑海 7.786955855499576E-4
巴基斯坦 7.40590943125283E-
南奥塞梯 7.269821422593277E
塔利班 6.507728574099785E-4
塞尔柱 6.453293370635964E-4
第比利斯 6.371640565440232E
莫卧儿 6.317205361976412E-4
帝国 6.289987760244501E-4
地毯 6.126682149853039E-4
普什图 5.772853327338203E-4
阿拉伯 5.60954771694674E-4
土耳其语 5.55511251348292E-
俄罗斯 5.473459708287188E-4
奥斯曼 5.391806903091457E-4

Topic 9th:

出版 0.02811097960922827
文学 0.012784609724522673
小说 0.011474578325331003
杂志 0.011245156582832969
作家 0.010151247114980026
作品 0.0073865488702392285
出版社 0.007142164840186976
编辑 0.005911932307951142
发表 0.005594399316522704
作者 0.005002557719933573
读者 0.0046667374881610895
翻译 0.004580288715625599
文章 0.004560338998886639
发行 0.004294342775700513
书籍 0.00428769287012086
报纸 0.00354955335077936
写作 0.0034830542949828287
漫画 0.003473079436613349
主编 0.0034464798142947364
创刊 0.0034365049559252564
内容 0.0033616935181541587
台湾 0.0033018443679372804
文化 0.0032320203593509222
中文 0.0029444119430309234
著作 0.0028862252692089586
文艺 0.0028546382177056058
创作 0.0028213886898073403
周刊 0.002670103337870231
月刊 0.0026501536211312716
中国 0.0025354427498822545
刊物 0.0024174069258434115
期刊 0.0023725200631807527
图书 0.0023309581533079203
散文 0.0022727714794859555
书店 0.0022727714794859555
评论 0.0022046099472945104
诗人 0.0021464232734725456
文学奖 0.0021198236511539327
日报 0.0020200750674591356
阅读 0.0019718632520066503
笔名 0.001935288771318558

算法示例

| <http://localhost/清华学堂> | <http://localhost/要点> | <历史发展，建筑风格>
| <http://localhost/清华学堂> | <http://localhost/建筑风格> | <清华学堂在翻修过程中拆除了一楼腐朽的木地板，填平了积水的地下室；为了抗震还用钢材加固了二层楼板和整体结构，外观也被修葺
| <http://localhost/清华学堂> | <http://localhost/历史发展> | <1909年清政府成立了游美学务处，负责直接选派学生游美，同时着手筹设游美肄业馆。清华学堂于1911年4月29日在清华园开学，这就
| <http://localhost/闻亭> | <http://localhost/要点> | <建筑风格>
| <http://localhost/闻亭> | <http://localhost/建筑风格> | <1986年，为纪念闻一多牺牲40周年，清华大学于大礼堂西侧土山坡下建立闻一多纪念像。此像为一尊红色花岗岩石雕坐像，高2.80米，
| <http://localhost/二校门> | <http://localhost/要点> | <建筑风格>
| <http://localhost/二校门> | <http://localhost/建筑风格> | <二校门位于清华主干道之一清华路，是清华园内最具代表性的标志性建筑之一，被认为是清华大学的象征。二校门为一座古典优雅的青砖
| <http://localhost/西区体育馆> | <http://localhost/要点> | <建筑风格，研究，历史故事，新时期发展，历史发展，目标，精神传承，运动项目>
| <http://localhost/西区体育馆> | <http://localhost/建筑风格> | <在清华校园的西校区有一座西洋古典式样的体育馆——清华大学西区体育馆。它是清华第一个体育馆，位于清华第一个运动场——西大
| <http://localhost/西区体育馆> | <http://localhost/历史发展> | <“为祖国健康工作五十年”：清华素来重视体育锻炼，1919年西区体育馆建馆之前，就已经开始了正规的体育课，直到今天老体育馆
| <http://localhost/西区体育馆> | <http://localhost/研究> | <见证反帝反侵略的历史：西区体育馆有着丰富的历史和文化价值，一些重大历史事件与之紧密相连。1919年北京爆发了反帝爱国的“五
| <http://localhost/西区体育馆> | <http://localhost/历史故事> | <“一二·九”运动期间，体育馆曾做过保护进步学生的“掩体”。1935年12月，为反对日本策划的华北五省“自治运动”，北平爆发了
| <http://localhost/西区体育馆> | <http://localhost/新时期发展> | <20世纪50年代初期，清华体育馆曾一度成为毛泽东主席冬季游泳的场所。当时中南海只有一个露天游泳池，秋冬季节不适合主席游泳
| <http://localhost/西区体育馆> | <http://localhost/目标> | <体育锻炼在清华蔚然成风，并有所建树，这与马约翰等人的努力分不开，“为祖国健康工作五十年”的口号的提出，也与马老的工作与贡
| <http://localhost/西区体育馆> | <http://localhost/精神传承> | <“为祖国健康工作五十年”这一口号50多年来在清华园一次又一次唱响，它是清华优良体育传统的集中体现，一直激励着清华园一代又
| <http://localhost/西区体育馆> | <http://localhost/运动项目> | <清华大学在以前每年的10月份举行新生运动会，目的主要是为了选拔体育生，后来为了达到群众普及的目的，从1998年起，改为赤足
| <http://localhost/图书馆> | <http://localhost/要点> | <历史发展，新时期发展，研究，作用与影响，使用情况>
| <http://localhost/图书馆> | <http://localhost/历史发展> | <1911年建立清华学堂。1912年清华学堂改建为清华学校，正式建立了小规模的图书室，称清华学校图书室。1919年3月图书室独立馆舍
| <http://localhost/图书馆> | <http://localhost/研究> | <　1952年国家教育体制改革，清华大学由一所综合性大学调整为工科大学，所有文、理科及一部分工程技术院、系被调整到其他院校，
| <http://localhost/图书馆> | <http://localhost/新时期发展> | <　抗日战争胜利后，清华大学迁回北京清华园。1946年复校时，图书馆已面目全非。抗战期间，日军以图书馆为外科病房，书库为手
| <http://localhost/图书馆> | <http://localhost/作用与影响> | <　1978年党的十一届三中全会以后，教育战线拨乱反正，迅速发展。清华大学自70年代末期开始进行学科调整，逐渐恢复理科、文科
| <http://localhost/图书馆> | <http://localhost/使用情况> | <　目前，校图书馆有宽敞明亮的阅览室9个，设置阅览座位1535个，每周开放时间105小时，实行开放式借阅一体化服务，为师生提供
| <http://localhost/工字厅> | <http://localhost/要点> | <历史发展，建筑风格>
| <http://localhost/工字厅> | <http://localhost/建筑风格> | <　工字厅原称“工字殿”，始建于1762年，以它为主体的一组清代皇室园林即是最早的“清华园”，距今已有近250年的历史。工字厅
| <http://localhost/工字厅> | <http://localhost/历史发展> | <1909年清政府在京设游美学务处，1911年2月，游美学务处与肄业馆(后改称清华学堂)迁入清华园，办公场所由史家胡同迁到了清华园

谢谢指正！

zhenchen@tsinghua.edu.cn