

Annotation Tools and Beyond

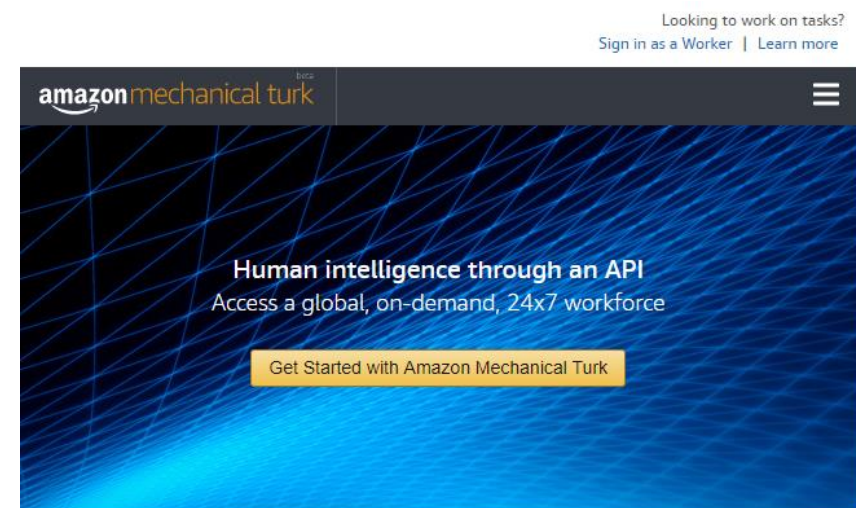
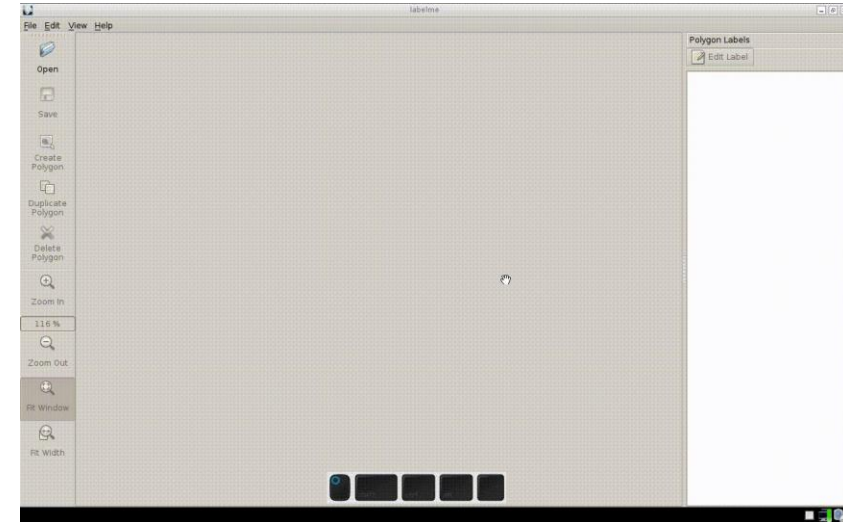
Guan Wang
2018/06/06

Table of Contents / Agenda

- Why Annotation Tools
- Annotator for Chinese Text Corpus
 - Intelligent Labelling with Active Learning
 - Data Pipeline and Modular Design
 - More than an Annotation Tool
- One more thing

Why Annotation Tools?

- Most ML Applications are **Supervised** Learning
- If lucky: automatically generated
 - Click Google ad on a webpage
 - Add stuff in shopping cart on Amazon
 - Listen to music on Spotify
- Not so lucky, but there are smart stricks
 - Crawl from Internet
 - Distant Supervision
- Not lucky, no tricks (usually)
 - Label by hands



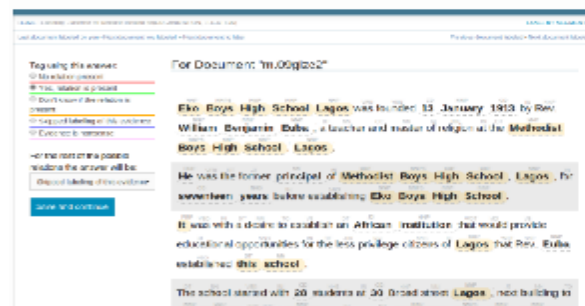
Amazon Mechanical Turk (MTurk) operates a marketplace for work that requires human intelligence. The MTurk web service enables companies to programmatically access this marketplace and a diverse, on-demand workforce. Developers can leverage this service to build human intelligence directly into their applications.

Annotator for Chinese

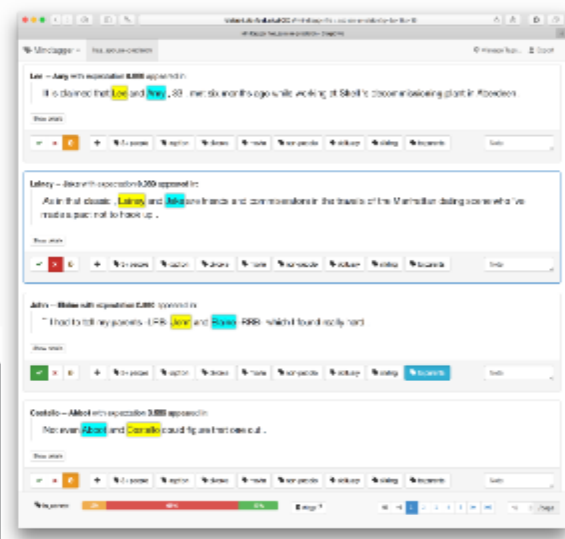
- Many tasks in NLP are supervised learning problem
 - Sequence Labelling (Tokenization, NER)
 - Classification (relation extraction, sentiment analysis, intent recognition)
- Not so much open-source annotated corpus for Chinese
- Vertical Applications require **Domain Knowledge** (insurance, finance, health, legal, public security etc.)
- Existing annotation tools are either
 - clumsy and complex to use
 - only supports English
 - not open-source and cannot use without public cloud
 - use out-dated software stacks

Annotator for Chinese

IEPY



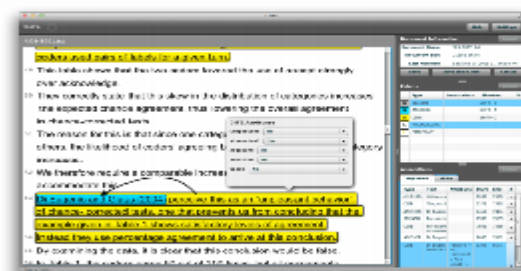
DeepDive (Mindtagger)



BRAT



Slate



Snorkel



Figure 3: Labelling functions which express pattern-matching, distant supervision, and weak classifier heuristics, respectively, in Snorkel's Jupyter notebook interface.

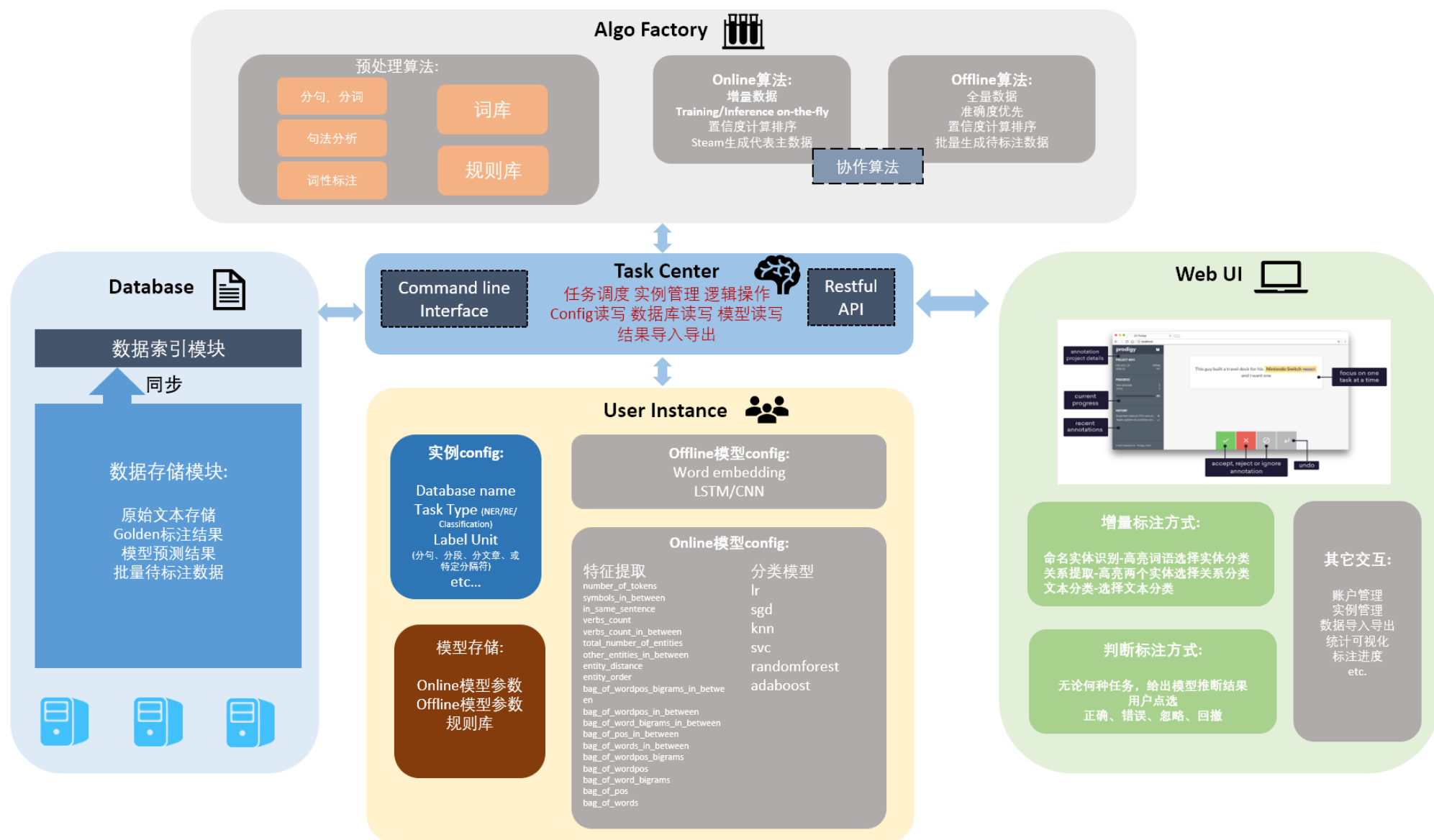


Figure 4: The Viewer utility in Snorkel, showing candidate company-employee relation mentions, comprised of candidate person and company mention pairs.

Intelligent Labelling with Active Learning

1. User make labels
2. Backend active learning algorithm will consist of "Online" part and "Offline" part.
 - "Online" part will do the online learning and update online model in real time, using fast traditional algorithms like Linear Classifiers or SVM
 - When label data accumulated to a certain amount, "Offline" part will update the offline model, using probably highly accurate deep learning models.
3. After model is updated, predict as much as possible within reasonable time, rank the confidence, and choose the lowest certain number of samples as datasets waiting to be labeled. Repeat step 1.

Intelligent Labelling with Active Learning



Intelligent Labelling with Active Learning

The screenshot illustrates the Anno data annotation platform interface, which is used for intelligent labeling with active learning. The interface is divided into several sections:

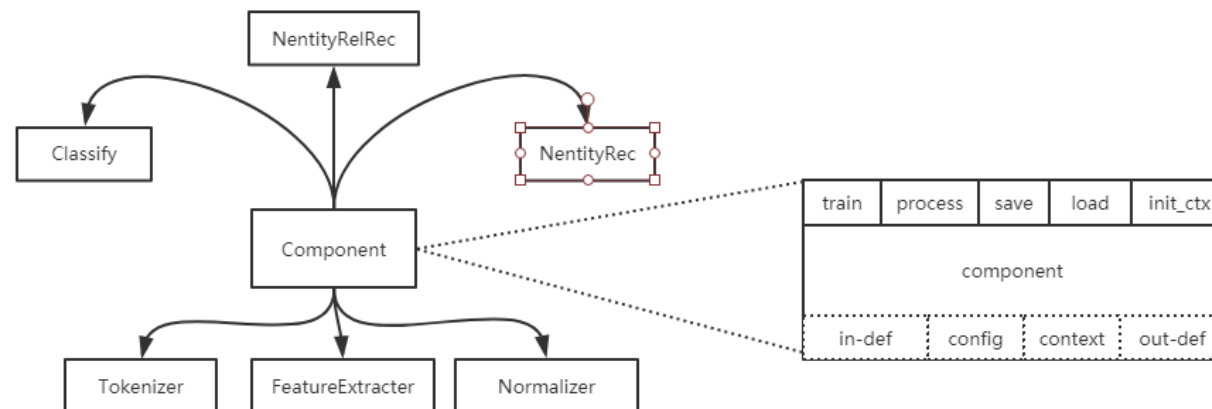
- Top Left: Anno 数据标注平台**
 - 最近使用的标注项目 (Recently used annotation projects)
 - * 公司名实体识别1 (Company name entity recognition 1)
 - * 快递地址文本实体识别 (Express address text entity recognition)
- Top Center: 选择您需要创建或导入的的标注项目类型** (Select the type of annotation project you need to create or import)
 - NER** (命名实体识别 - Named Entity Recognition)
 - TC** (文本分类 - Text Classification)
 - Face** (人脸识别 - Face Recognition)
 - 导入已有项目** (Import existing project)
- Top Right: 创建新标注项目** (Create new annotation project)
 - 项目名称 (英文或中文) (Project name (English or Chinese))
 - 项目类型 (Project type): 文本分类 (Text Classification)
 - 主动学习算法 (Active learning algorithm): 无 (None)
 - 待标注数据集 (待标注数据集 文件选择(File...)) (Dataset to be annotated (File selection))
 - 分类 (Classification): 劳动纠纷, 刑事纠纷, 消费者权益保护, ... (Labor dispute, Criminal dispute, Consumer rights protection, ...)
 - Buttons: 创建 (Create), 取消 (Cancel)
- Bottom Left: Anno 数据标注平台**
 - 类型 (Type): 文本分类 (Text Classification)
 - 数据集 (Dataset): iccpol-train
 - 已标注 (Annotated): 139/299
 - 历史 (History): 刑事纠纷的引起是因为... (Criminal dispute is caused by...), 五个老大爷们围堵在... (Five old men are surrounded by...), 五个老大爷们围堵在... (Five old men are surrounded by...)
- Bottom Center: 新增标注类型指南** (New annotation type guide)
 - 首先, 本案系劳动争议和工伤保险待遇纠纷, 并不是身体权、健康权纠纷, 一审法院混淆了二者的关系, 且将赔偿项目进行了选择性叠加, 造成错误的判决。 (First, this case is a labor dispute and a dispute over social security benefits, not a dispute over physical or health rights. The first instance court confused the relationship between the two, and selectively stacked the compensation items, resulting in an incorrect judgment.)
 - Buttons: 查看上一个(左键) (View previous (Left button)), 忽略(X键) (Ignore (X button)), 下一个(右键) (Next (Right button))
- Bottom Right: 创建新标注项目** (Create new annotation project)
 - 项目名称 (英文或中文) (Project name (English or Chinese))
 - 项目类型 (Project type): 命名实体识别 (Named Entity Recognition)
 - 主动学习算法 (Active learning algorithm): 无 (None)
 - 待标注数据集 (待标注数据集 文件选择(File...)) (Dataset to be annotated (File selection))
 - 实体类型 (Entity type): ORG, LOC, _ (Organization, Location, ...)
 - Buttons: 创建 (Create), 取消 (Cancel)

Legend:

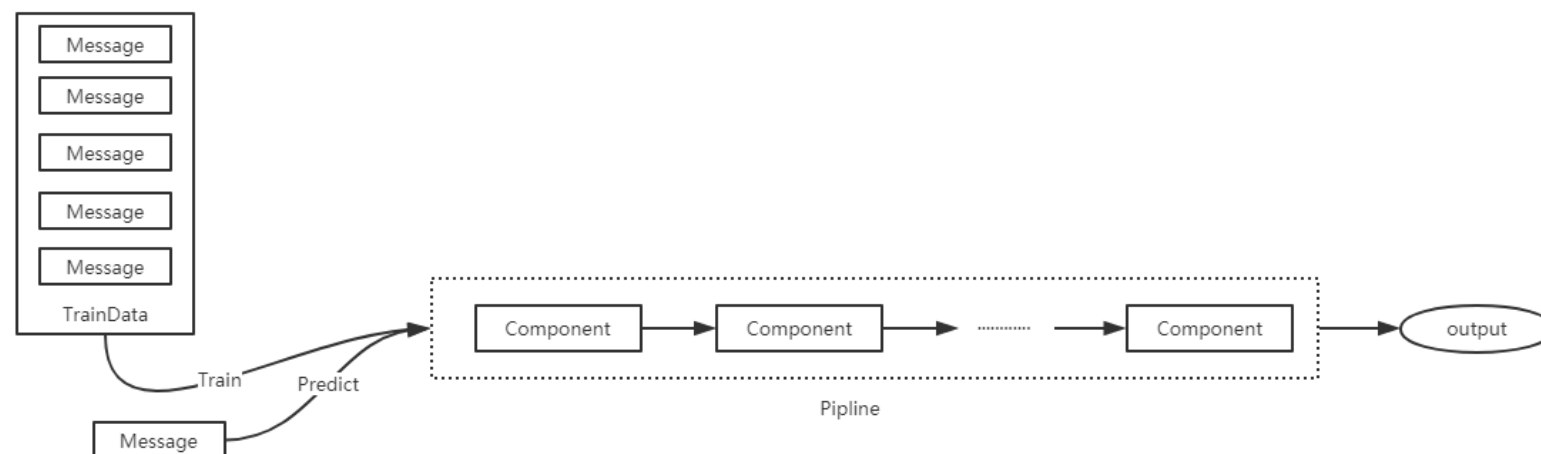
- 蓝色代表当前模型的预测结果 (Blue represents the prediction result of the current model)
- 绿色是在用户点击之后出现的颜色 (Green is the color that appears after the user clicks)

Data Pipeline and Modular Design

Modular Design



Data Pipeline Design



Data Pipeline and Modular Design

Configurable Programming

```
{
  "ip" : "localhost",
  "port" : "8000",
  "database_type" : "mongodb",
  "type" : "classification"
  "name" : "email_spam_classification",
  "model_type" : "classification",
  "pipeline": ["nlp_word2vec",
               "linesplit_preprocess",
               "feature_extractor",
               "online_svm_classifier_sklearn",
               "offline_svm_classifier_sklearn"],
  "language": "zh",
  "wordvec_file": "./tests/data/test_embedding/vec.txt",
  "path" : "./tests/models",
  "org_data" : "./tests/data/test_email_classify/email_classify_chi.txt",
  "database_name" : "spam_emails_chi",
  "labels": ["spam","notspam"],
  "batch_num" : "10",
  "inference_num" : "20",
  "low_conf_num" : "10",
  "confidence_threshold" : "0.95",
  "log_level": "INFO",
  "log_file": null
}
```

More than an Annotation Tool

Thanks to **Modular** and **API** design:

1. Human User Interface for Machine Learning Projects

2. Data Manager

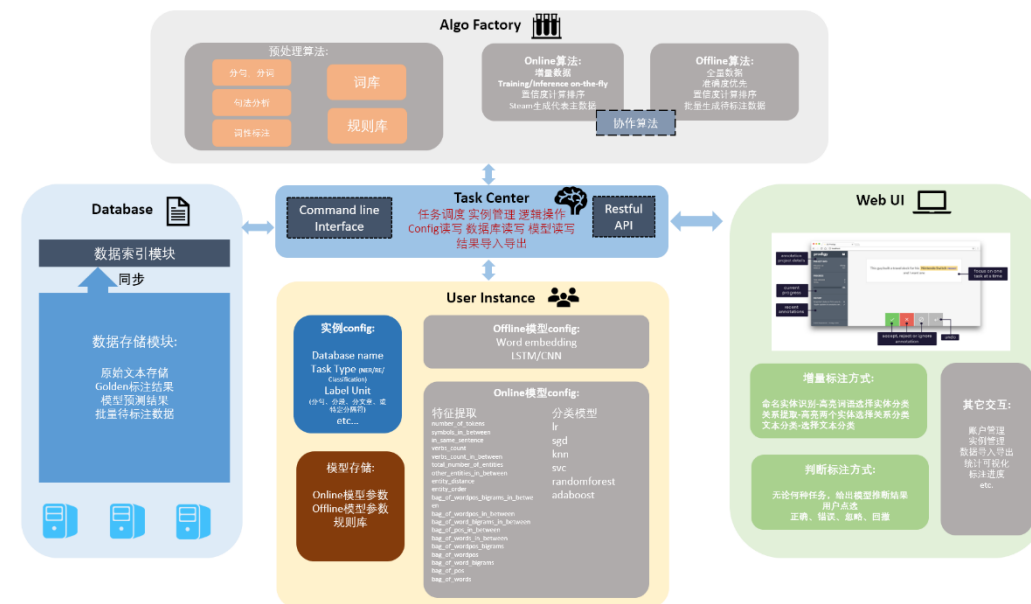
- raw data, pre-processed data, feature engineered data, labelled data, predicted data etc.
- Upstream module like Crawlers
- Downstream modules like Visualizations

3. Model Manager

- pre-trained models, configuration of online and offline models, persisted models

4. Prediction Service

Full Pipeline Machine Learning Tool



One More Thing

- 10 People
- Full Stack hackers, algorithm experts, professional software engineers
- Scattered in Chengdu, Shanghai, Nanjing, Beijing, Guangzhou, Shenzhen and Hong Kong (luckily same time zone)
- Process
 - Vague idea
 - White Board design
 - Software architecture diagram
 - Choice of stack back/front end
 - Configurable and Pluggable Algorithm Design
 - Unit Testing Cases
- As of Feb.22nd 2018:
 - 30k lines of codes
 - 153 commits
 - 248 stars
 - 76 forks
 - 14 issues
 - 41 pull requests.

One More Thing

Tools we use:

- Github
 - Code review
 - issue tracker
 - code discussion
 - wiki documents
 - Travis-CI (continuous integration)
- Asana
 - project management
 - allocate and claim tasks
 - set deadlines
 - synchronize progress
 - share resources
 - discuss specific tasks
- Wechat (core team discussion)
- Gitter chatroom (public discussion)

The image shows a composite view of development tools. The top part is the GitHub repository page for 'crownpku / Chinese-Annotator', showing repository statistics (35 watchers, 128 stars, 44 forks) and navigation tabs (Code, Issues, Pull requests, Projects, Wiki, Insights, Settings). The bottom part is the Asana task management interface, displaying a chat-like view for the 'Chinese-Annotator/Lobby' project. The chat messages include:

- Liu Fan @feanlau: hello
- Guan Wang @crownpku: Hi @/all, 我们的项目现在还在开发流程框架、算法模块、前端webui和后端数据库, 已经渐渐成型了。各位如果有兴趣的话可以在这里讨论想法和实现。谢谢!
- Yunong Pang @ppn029012: 现在是不是应该分出一个dev分支? 然后写一个简单部署配置的说明文档?
- Guan Wang @crownpku: @ppn029012 我是想我们先开发出第一个可用的spam email classification demo, 出一个alpha的release版本完善说明文档, 然后再分dev分支吧。现在整个系统还没跑通。
- Guan Wang @crownpku: webui + database
前端界面和数据库都还在开发中, 重点是要写好API和相应文档以便未来后端算法模块的接入, 下一步希望能尽快出一个能看到界面的demo~
task_center + user_instance
当前已经搭建好一个最简单的offline training的pipeline。
下一步既然我们已经有spam email的数据和label, 在前端完成之前, 需要模拟用户“拿到confidence最低的一批数据”->“续标数据(其实就是从数据里把该部分的label拿出来)”->“重新训练给出confidence ranking”这样一个过程, 完成一个模拟的online training & inference pipeline的test case。未来就可以方便接入前端与数据库的部分。
另外一个, 就是要注意把具体任务(如spam email classification)的所有配置文件(.config), 文本数据(.sqlite/mongodb), 模型数据(tensorflow/sklearn/jieba词库)乃至状态数据全都实例化单独放在同一个

The Asana interface also shows a sidebar with 'My Favorites', 'Reports', and a search bar. The right sidebar of the Asana view shows 'PEOPLE' and 'ACTIVITY' sections.

One More Thing

- Working Style: **Distributed Asynchronous Collaboration**
- No compulsory meetings, every communication is online and persisted in an asynchronous way
 - Pull requests at mid-night
 - Merge and review code on the subway
- Every member chooses when, where and how he works.
 - beds, sofas, cafes, libraries, swimming pool or anywhere with good quality Wifi. No commute.
 - Save rent; Help environment protection
- Flat and Transparent, Spontaneous Collaboration, Full Trust and Support for Human Nature



Q&A



Legal notice

©2018 Swiss Re. All rights reserved. You are not permitted to create any modifications or derivative works of this presentation or to use it for commercial or other public purposes without the prior written permission of Swiss Re.

The information and opinions contained in the presentation are provided as at the date of the presentation and are subject to change without notice. Although the information used was taken from reliable sources, Swiss Re does not accept any responsibility for the accuracy or comprehensiveness of the details given. All liability for the accuracy and completeness thereof or for any damage or loss resulting from the use of the information contained in this presentation is expressly excluded. Under no circumstances shall Swiss Re or its Group companies be liable for any financial or consequential loss relating to this presentation.