小诗机

孙玉霞 2017年12月09日

自我介绍

孙玉霞

南京师范大学硕士，曾就职于1号店，三星等公司，从事自然语言相关工作，在数据挖掘，语义表征等方面有一定的研究。

现为携程基础业务部算法工程师，参与小诗机，智能客服，产品强化等相关项目。

大纲

小诗机项目

Input:

景点写诗

城市写诗

天气、季节

心情

照片写诗

格律：

宝塔诗

律诗

绝句

景物识别&
成诗：
红楼；
梅花；

照片地点
知识库：
西湖；
西湖的知
识库

小诗机项目

Input:

景点写诗

城市写诗

天气、季节

心情

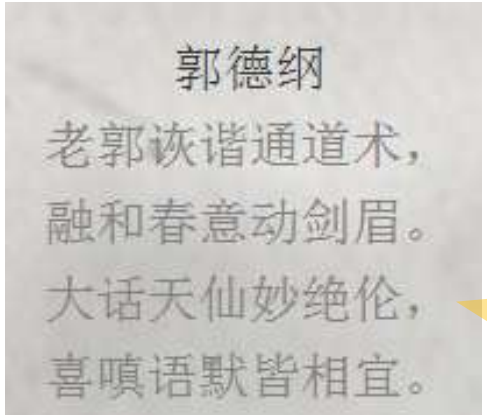
照片写诗

格律：

宝塔诗

律诗

绝句



人物画像：职业；作品；身材；风格

小诗机项目

人机盲测 第3名(11场)

	冠军	亚军	季军
赵玉玮	5	0	1
邹竞夫	2	6	2
小诗机	**2**	**3**	**2**
卿云子	2	2	2
沈以昕	0	0	2
张一双	0	0	0

与上海地区诗人比，略胜

	思念	孤独	夏之愉悦	秋之悲伤	北京的雾	杭州的雨	黄山的雪	青海湖的晴	沙-正宝塔	沙-倒宝塔	景点 愉悦
赵玉玮	7.45	7.155	6.72	6.47		6.52	6.73			6.68	6.72
邹竞夫	8.37	7.575	6.705	6.49	8.22	6.41	6.685	6.32	6.425	6.24	6.375
AI LEE	7.885	8.14	6.195	6.535	7.61	6.225	6.07	6.4	6.15	5.91	6.4
卿云子		8.395	6.315		8.655		6.6			6.365	6.515
沈以昕	7.725	6.725	5.85	6.32	7.565	6.395			5.61		6.31
张一双	7.115			6.155			6.325	5.87	5.175		6.115

宝塔寺、情感、季节、天气、人物、绝句、律诗、城市、美食

西湖

凉风乱入杭城路，
秋水孤山似旧年。
遍地桂花停过客，
侵帘绿影满湖边。

西湖

绿树余杭花草畔，
通都月色谁自怜。
喜听潇雨鸣枫叶，
安得凉风起桂筵。
久靠林荫堤岸伫，
长依湿地碧波边。
西湖桥下潺湲水，
凭有孤山却不前。

西湖. 夏

重云深处引溪涧，
湖水孤山路远长。
遥爱断桥残夏早，
风清杨柳映通方。
莲花倚槛雨生雾，
荷叶绕船橹走旁。
林缝落光波滟滟，
寺边湿地色苍茫。

西湖. 雪

城巷夜空却白昼，
西湖山色故依然。
敢危湿地凌冬约，
不负君梅岁竹缘。
万里重云寒水边，
满城娑树雪中天。
萧疏枫叶扁舟岸，
灵隐冷风谁顾怜。

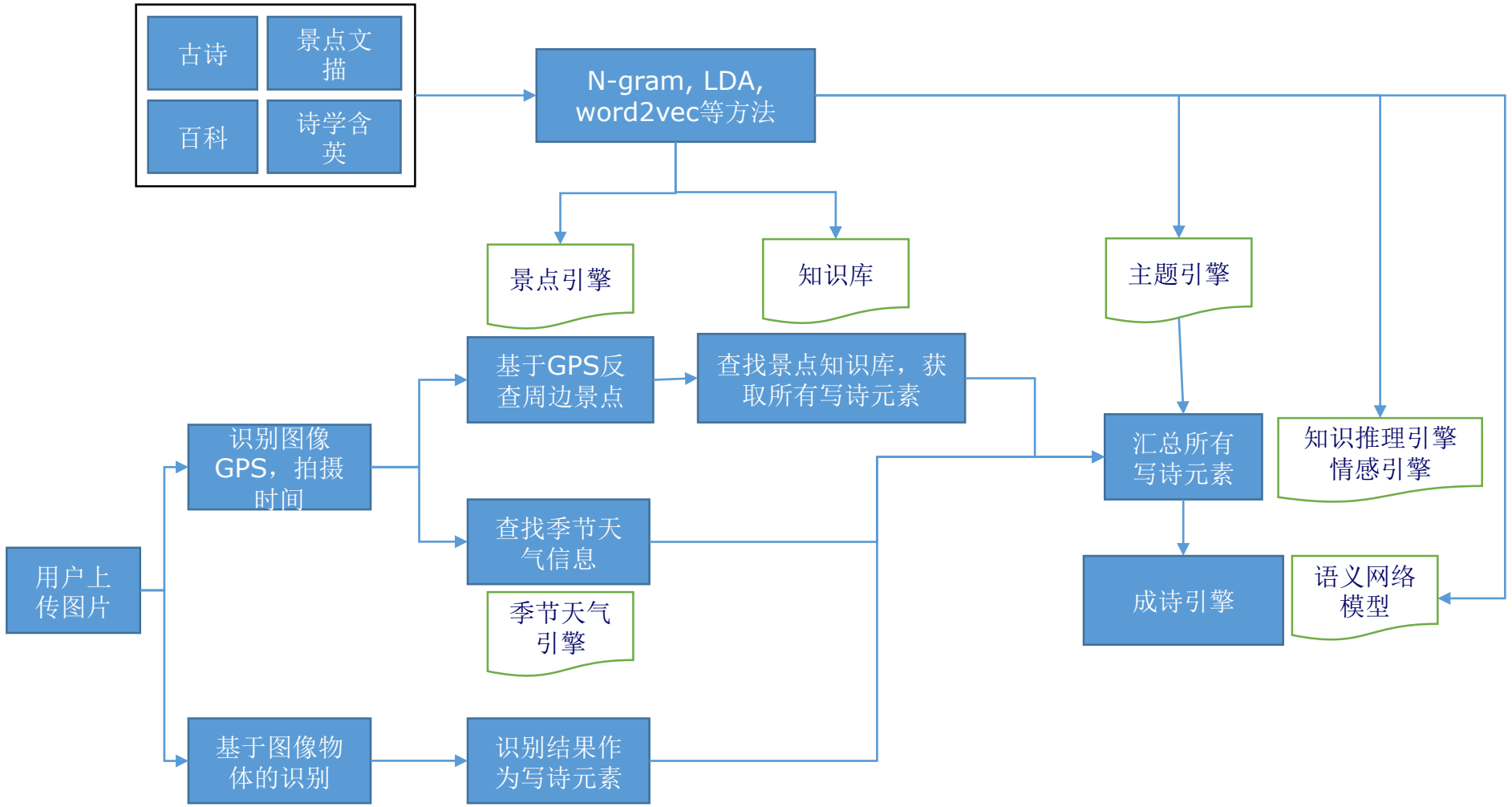
澜

惊湍
水流漫
鲸波海观
江涌万卷蟠
涛濑掀翻浪寒
激荡澎湃渺云端

香葱燕雀凉醉颊

玉眉梅蕊翠靥
偷香团扇妾
芳情细贴
花重叠
春叶
蝶

小诗机



小诗机


应用层
七言绝句
七言律诗
宝塔诗
藏头诗
技术
服务层
成诗引擎
图搜索算法
遗传算法
RNNLM
情感引擎
知识推理引擎
季节天气引擎
知识层
知识库
语义网络模型
基础词
模块
三层语义体系
情感词
景点知识库
同义词林
笠翁对韵
诗学含英
20w+古诗
百度百科
携程景点文描

小诗机

图像识别

知识库

成诗引擎

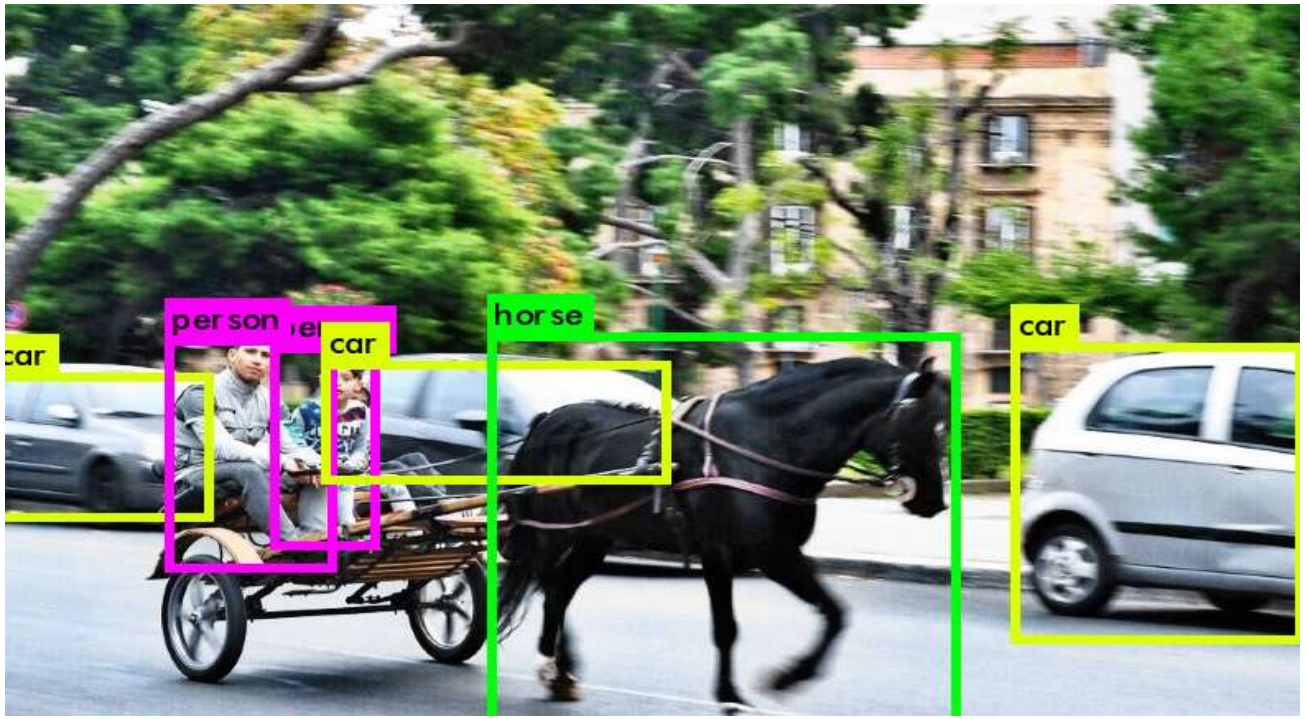
小诗机

图像识别

Image detection - SSD

存在的问题

- 语料有限
- 标注较为复杂



小诗机

图像识别

Multi-label – inception V3 + transfer learning

人: 0.959
冲浪板: 0.663
船: 0.372
水: 0.396
海: 0.251

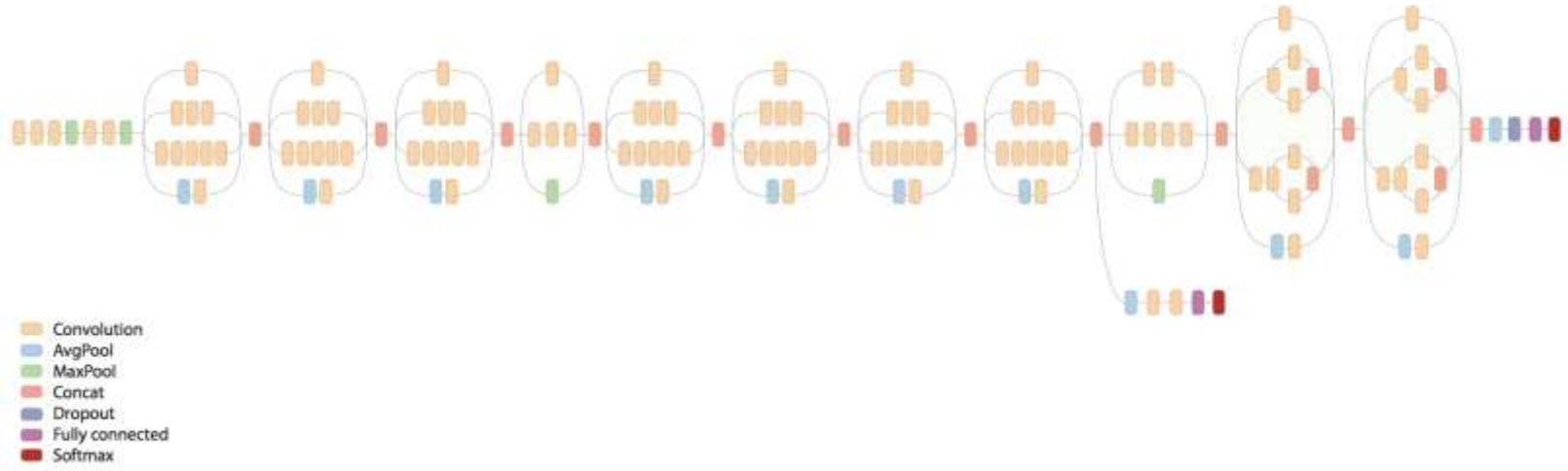
图像识别

Multi-label: Transfer Learning

condition	action
New dataset is small and similar to original dataset	train a linear classifier on the CNN codes
New dataset is small and similar to original dataset	fine-tune through the full network
New dataset is small but very different from the original dataset	train the SVM classifier from activations somewhere earlier in the network.
New dataset is large and very different from the original dataset.	it is very often still beneficial to initialize with weights from a pretrained mode

图像识别

Multi-label: Inception V3



https://github.com/tensorflow/models/tree/master/research/inception

小诗机

图像识别

知识库

成诗引擎

小诗机

知识库

- 常识能力
  - I. 特征知识库
  - II. 公共知识库
  - III. 特征知识引擎
- 诗词能力
  - I. 诗词知识库
    - a. 特征三层语义知识库
    - b. 特征-候选写诗词知识库
    - c. 相关主题知识库
  - II. 诗词引擎

next

小诗机

知识库

景点-景点特征知识库

- 包含49266个poi
- 包含8722类基础特征，如湖水

景点	特征	相关分值
西湖	西湖	1.05
西湖	苏公堤	0.55
西湖	小瀛洲	0.52207
西湖	杭城	0.1
西湖	断桥	0.099524
西湖	湖水	0.081093
西湖	西泠桥	0.071546

知识库

特征三层语义知识库

将特征/季节/节日/天气/情感等相关信息映射到三层语义体系

- 包含80左右的一级语义
- 包含400左右的二级语义
- 包含3,000左右的三级语义
- 共包含6万条三层语义Item数

候选写诗词	三级语义	二级语义	一级语义
正午;日午;午;中午;晌午;当午;昼	晌	早中晚	时间
薄雾;烟雾;云雾;氤氲	雾	天气	自然天气现象
村落;古村;村庄	村	村镇	建筑

小诗机

数据来源

- 结构化数据：百度，用户画像，景点数据
- 非结构化数据：用户评论，游记，景点文描，百度百科
- 字典：同义词词林，知网，诗学含英

数据挖掘

- TF-IDF：有效评估当前特征的的重要性

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad \mathrm{idf}_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

- Mutual Information：有效量化随机变量之间的相关性

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \qquad I(X;Y) = H(X) - H(X|Y)$$

- 关联规则

数据挖掘

- 词袋模型-One-hot Representation，这种方法把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。

For Example

“话筒”表示为 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 ...]

“麦克”表示为 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ...]

问题:

1、向量的维度会取决于词语的量

2、任意两个词之间都是孤立的，无法进行语义层面的表征

小诗机

数据挖掘

- LSA
  1. SVD分解
  2. 没有概率含义
- PLSA
  1. 添加概率信息
  2. 优化似然函数
- LDA
  1. 生成式概率模型

小诗机

数据挖掘

- Word2vec
  - 浅层神经网络

- 余弦相似度
  1. 各个维度指标必须一致
  2. 适合比较稠密的矩阵

$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k}\, x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^{2}}\,\sqrt{\sum_{k=1}^{n} x_{2k}^{2}}}$$

小诗机

图像识别

知识库

成诗引擎

小诗机

机器写诗

- 传统方法
  1. 基于模板和模式的方法
  2. 基于遗传算法的方法
  3. 基于统计机器翻译的方法 PMT

- 基于 encoder-decoder 框架
  1. Attention机制: Q Wang , T Luo , D Wang , C Xing. Chinese song iambics generation with neural attention-based model. International Joint Conference on Artificial Intelligence , 2016 :2943-2949
  2. 规划模型+Attention机制: Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, Enhong Chen. Chinese Poetry Generation with Planning based Neural Network. COLING 2016.
  3. hierarchical 的 RNN: i,Poet: Automatic Poetry Composition through Recurrent Neural Networks with Iterative Polishing Schema

- 基于GAN
  1. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient

语言模型

➢给定一个句子，计算这个句子出现的概率？假设T是由词序列W1,W2,W3,…Wn组成的，则这个话存在的概率为： *P(T)=P(W1W2W3Wn)=P(W1)P(W2|W1)P(W3|W1W2)…P(Wn|W1W2…Wn-1)*

存在问题：参数空间过大，不可能实用化；数据稀疏严重

➢马尔科夫假设：一个词的出现仅仅依赖于它前面出现的有限的一个或者几个词

$$p(q_t = w_j \mid q_{t-1} = w_i, q_{t-2} = w_k, \ldots) = p(q_t = w_j \mid q_{t-1} = w_i)$$

$$p(\mathrm{T}) = p(w_1 w_2 w_3 \ldots w_n) = p(w_1)\, p(w_2 \mid w_1)\, p(w_3 \mid w_2) \ldots p(w_n \mid w_{n-1})$$

➢ 对于古诗而言，短句表达，语言模型的稀疏性更严重。

语义层级语言模型

- 写景诗歌的成诗体系进行分析和挖掘，提取常用语义构建三层语义体系结构，基于语义体系获取语言模型

1. 从多个层次上来把握诗歌的语义搭配粒度
2. 基于提取出的语义结构，挖掘常用诗歌组合模型，将诗歌组合模型和语义层级相结合
3. 表达更加的流畅

$$p(s_1 \mid s_2) = \frac{freq(s_1, s_2)}{freq(s_2)} = \sum_{w_i \in S_1} \sum_{w_j \in S_2} freq(w_i, w_j) / \sum_{w_j \in S_2} freq(w_j)$$

小诗机

计算模型

- 构建语义模式库
  1. 利用语义级的语言模型进行语义模式挖掘
  2. 利用候选写诗词进行浅层语义模式挖掘

小诗机

计算模型

- 检索引擎 – 粗排
  1. S(query，patterns，position)
  2. 主题相关性
- 二次排序 - 细排
  1. 差异性排序(主题表达/写诗词)
  2. 贪婪算法/局部最优
- 基于二次排序的BFS搜索 – 成诗
  1. BFS (pattern, rhythm, state) > threshold
  2. 贪婪算法/局部最优
  3. 可回退

计算模型

- 遗传算法：解决最优化的一种搜索启发式算法

1. 求解问题编码
2. 初始化种群
3. 适应度函数设计
   - 综合语言模型因子，韵律因子，主题因子，对仗因子
4. 遗传操作
   - 借助于自然遗传学的遗传算子（genetic operators）进行组合交叉（crossover）和变异（mutation），产生出代表新的解集的种群

Thank you