Lec 2 爬虫

常见反爬虫策略：User Agent：利用 Request Headers 信息中的 User-Agent 信息，判断用户使用何种方式浏览。爬虫往往 U-A 部分为空。应对：Request 库自定义请求头信息的手段，将 U-A 的值改为浏览器的请求头标识。IP/账号访问次数/频率：限制特定 IP 地址/账号访问频率和次数。其本质在于判断浏览行为是否是人类行为。应对：IP 代理池、爬取后间隔一段时间、注册多个账号。验证码：伤害用户体验。应对：字符识别：基于机器学习与模式识别相关技术；逻辑推理：人工。动态网页：从网页的 url 加载网页的源代码之后，会执行 JS 程序加载网页内容。多用于音视频文件。应对：模拟调用请求，使用审查元素分析 ajax 请求，如此循环直到获得包含数据信息的 json 文件。蜜罐技术：留下一些人类看不到或者绝对不会点击的链接，爬虫会从源代码中获取内容，所以爬虫可能会访问这样的链接。发现有 IP 访问这个链接，立刻永久封禁访问者。应对：干涉爬虫路径。通过工具库判断页面上的隐含元素，使爬虫避开这些元素。用户权限限制：不同类型/级别的用户给予不同的内容权限。应对：氪金。不同的网页结构：相同类型的页面的源代码格式不同。多模态的呈现方式：文字转为图像或视频。应对：OCR、语音识别、图像/视频标签技术。

Lec 3 网页文字处理

基于匹配的分词方法：机械分词方法，按照一定的策略将待分析的汉字串与一个充分大的机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。正向最大匹配 FMM、反向最大匹配 RMM、双向最大匹配 BM（综合比较前两者的切分效果，直接合并结果或选择词数最少的结果）、最少切分分词（最短路径）(使句子中切出的词数目最少，等价于在有向图中搜索最短路径的问题，将每个字视作节点，前后自然连接边，词形成一条边，即寻找有向图最短路径。边权=1 或根据词频）(拓展：N-最短路径法，保留 N 条最短路径)。优点：效率高、直观性好。缺点：对词典的依赖性。维护词典开支、难应对新生词、词频与重要性无影响。

基于统计的分词方法：字与字相邻共现的频率或概率能够较好的反映成词的可信度。计算不同分词方案的概率，选出概率最大的分词结果。N-gram：文本中第 N 个词出现的概率仅仅依赖于它前面的 N-1 个词。一元：$P(w)=P(w1)\,P(w2)\cdots P(wn)$。二元：$P(w)=P(w1)\,P(w2|w1)\cdots P(wn|wn-1)$。$P(wn|wn-1)=C(wn-1\,wn)\,/\,C(wn-1)$,优点：减轻了对于词典的依赖性（并非完全消除），结合机械分词则效率提升但依赖词典；减少对词典的依赖则解空间巨大。缺点：依赖已有数据中词频的统计，对于新生词汇或专业词汇不友好。冷门领域的稀有词汇往往难以准确划分；易受数据集先验偏差的影响。

隐马尔可夫模型：两个集合：观测值集合（字符集合）、隐藏状态值集合（BEMS）。三个矩阵：初始状态概率矩阵：第一个字属于某种隐含状态（BMES）的概率；隐含状态转移概率矩阵：各种隐含状态（各种标签）之间的转移概率；观测状态概率矩阵：从隐含状态（标签）到观测值（字符）的转移概率。当观测到句子$w1,w2\cdots wn$，希望找到相应的标签序列$s1,s2\cdots sn$，使得$P(s1,s2\cdots sn|w1,w2\cdots wn)$概率最大。齐次假设：当前隐藏状态只与上一个状态有关系。观测独立性假设：观测值之间互相独立的，只与生成它的状态有关系。目标：使得$P(si|si-1)P(wi|si)$ 乘积最大。

初始化：在 t=1 时，对每个状态 si，计算状态为 si 观测 w1 为指定状态的概率。

递归：对每个状态 si，计算前一状态为 sj（且观测 wt-1 为指定状态）且观测 wt 为指定状态的路径最大概率。同时，记录该最大路径的前一状态作为前驱。

终止：在最后一个字，取得到的最大概率，并得到最后一个字的状态。

回溯：从最后一个字的最优状态向前不断找前驱，还原全部标签。

停用词：文档中频繁出现或对实际语义影响不大的词语。The、of、的、是、数字、副词。为什么：重复率高，索引中的倒排表很长，影响查询性能。对排序没贡献，反而可能干扰。停用词识别方法：文本频率、词频统计、熵计算、统计与句法或内容分析。隐患：特定场景下有意义、停用词的组合有意义。现代搜索引擎的趋势是逐渐减少对停用词的使用，更关注利用语言的统计特性来处理常见词问题。压缩降低停用词表的存储开支。词项权重，将高频词的影响降至最低。索引去除技术，低于权重的词项将被排除。

归一化/词根化：指还原词语的特殊形式的过程。词干提取（缩减）：去除单词前后缀，获得词根，不一定是完整单词。词形还原（转变）：将单词的复杂形态转变成最基础的形态，根据词典将单词进行转换，更依赖于词典，转变得到完整单词。两者不是互斥关系，其结果有部分交叉。词形还原需考虑词缀转化、词性识别等，更为复杂。

Lec 4 网页索引

倒排表的构建：1.检索每篇文档，获得<词项，文档 ID>对，并写入临时索引；2.对临时索引中的词项进行排序；3.遍历临时索引，对于相同词项的文档 ID 进行合并。

倒排表的查询：本质上是倒排记录表的合并过程。

倒排表的优化：改变处理顺序：按照文档频率的顺序进行处理。首先，获得所有词项的文档频率；其次，保守地估计出每个 OR 操作后的结果大小（考虑最坏情况）；最后，按照结果从小到大的顺序执行 AND。

跳表指针：较多的指针+较短的步长=更多的跳跃机会+耗费更多的存储空间。较少的指针+较长的步长=更少的指针比较次数+存储空间消耗更少+跳跃机会更少。以根号表长的间隔均匀放置跳表指针（没有考虑查询词项的分布，未必导致结果优化）。加入新属性：词项频率、词项类型。二元词索引：将文档中每个连续词对看成一个短语。更长的短语查询可以分为多个短查询处理（不能确定其中是否真正包含最原始的长短语）。位置信息索引：在记录词项的同时，记录它们在文档中出现的位置，位置信息索引能够用于邻近搜索（例如，间隔 k 个词）。可变长度编码：先存储 G，并分配 1bit 作为延续位。如果 G<128，则采用第一位延续位为 1 加 7 位有效二进制编码的格式。如果 G>=128，则先对低阶的 7 位编码，然后采取相同算法对高阶位进行编码。最后一个字节（低位）的延续位为 1，其他字节延续位为 0。压缩词项列表：将词典视作单一字符串，词项之间用指针分割，指向下一个词项的指针同时也标识着当前词项的结束。按块存储：单一字符串在词项指针上需要占用较多额外空间。通过为每 k 个词项存储一个指针，来减少指针的总数量，需要额外 1 个字节用于表示词项长度。K 越大，存储空间越小，查找越慢（块外二分，块内线性）。前端编码：使用特殊字符表示公共前缀。

Lec 5 查询和评估

相关性反馈类型：显式反馈：用户点击记录（只有正样本），拓展：不感兴趣、用户评论。隐式反馈：鼠标键盘动作（揭示用户身份）、用户眼球动作（揭示用户关注的内容）。优点：不需要用户显式参与，减轻用户负担，提升用户体验。用户行为某种程度上可以反应其兴趣，因此具有可行性。缺点：对行为分析有着较高的要求。准确度难以保证。某些情况下需要增加额外设备。伪反馈：无需用户参与反馈过程，而直接根据检索结果自动反馈。对于用户查询返回的有序结果，假定前 K 篇文档是相关的，在此基础之上，进行相关性反馈。可以提升检索的效果，但难以保证其准确性，甚至出现查询漂移。

查询扩展：用户针对词项的合适程度给出反馈，这些反馈将被用来构建更为完整的查询条件。用户选择和确认的查询扩展能够更好表达其查询意图。拼写检查：基于编辑距离。同义词拓展：对于某个查询词汇，使用辞典中的同义词或相关词进行扩展，维护词典需要代价。人工编纂。全局分析：分析文档集中的词项分布，来自动生成词典。计算相似度：与相似词共同出现、与特定语义关系共同出现。搜索日志。潜在问题：词项关联的质量。

准确率（Precision）：指检索出的文档中，相关文档所占的比例，也称查准率。计算公式为 TP/(TP+FP)。召回率（Recall）：指所有相关文档中，被检索出来的部分的比例，也称查全率。计算公式为 TP/(TP+FN)。无法列举所有相关文档，故召回率无法准确计算。解决方案：缓冲池。针对某一检索问题，各个算法分别给出检索结果中的 Top N 个文档，将这些结果汇集起来并进行人工标注，从而得到一个相关的文档池。假设大多数相关文档都在这个文档池中。

Accuracy：(TP+TN)/(TP+TN+FP+FN)，在信息检索的相关任务中并不常见。不返回搜索结果，能保证 TN 最大。

F 值：准确率与召回率的加权调和平均数。当两者同等重要时，F=2PR/(P+R)。算数平均和几何平均在处理极端情况下的效果并不够合理。

P-R 曲线：以准确率和召回率分别作为两条轴线，通过选定不同的阈值得到不同的 P-R 点并连接成线。可通过平衡点（P=R）比较算法好坏。

ROC 曲线：以真正率 TP/(TP+FN)和假正率 FP/(FP+TN)作为两条轴线，通过选定不同的阈值得到不同的真正率假正率点并连接成线。对角线表示区分能力为 0，即随机猜测。在对角线上端越远，效果越好。低于对角线的结果无意义。P-R 曲线和 ROC 曲线可以通过比较曲线下的面积（AUC）判断算法好坏。

P-R 曲线和 ROC 曲线的选择：ROC 曲线兼顾正负样例，更为全面，而 P-R 曲线则只考虑正例。用户往往更关心正样本，如果面向特定应用场景（如检索），P-R 曲线是个好选择。单份数据正负样本比例失调时，P-R 曲线更合适。当负样本比重过高时，负例的数目众多致使 FPR 的增长不明显，导致 ROC 曲线呈现一个过分乐观的效果估计，从而难以体现出性能的差异性。P-R 曲线受分布影响大，多份数据且正负比例不一时 ROC 曲线更合适。

P/R@N：前 N 个检索结果文档的准确率/召回率。大多数用户只关注第一页或前几页，因此对于大规模搜索引擎来说是很合适的评价指标。如果相关文档数小于 N，P@N 的理论上限必定小于 1。由于返回结果有限，R@N 值，甚至其理论上限往往都远小于 1。新的不相关文档被检索时，Recall 不变，Precision 下降，故 P/R@N 曲线呈现锯齿状。

R-Precision：在所有相关文档总数位置上的正确率，即 P@N，N=相关文档总数。

AP：平均准确率，对不同召回率点上的正确率进行平均。未插值 AP：查询 Q 共有 6 个相关结果，排序返回了 5 篇相关文档，位置分别是第 1，第 2，第 5，第 10，第 20 位，则 AP=(1/1+2/2+3/5+4/10+5/20+0)/6。

插值 AP：事先选定插值点数并进行插值。例如，当计算 11 点平均时，计算在召回率分别为 0（第一条），0.1，0.2，…1 的十一个点上的正确率求平均。简化 AP：只对返回的相关文档进行计算，没有考虑召回率和补零的情况（不补 0）。

CG：累计增益。用于衡量位于位置 1 到 p 的检索结果的相关度之和。未考虑文档位置。CG=sum{rel_i}

DCG：折损累计增益。若搜索算法把相关度高的文档排在后面，则应该给予惩罚。DCG=rel_1+sum{rel_i/log i}或 DCG=sum{((2^rel_i)-1)/lb(i+1)}。后者采用指数，更突出相关性。

NDCG：归一化折损累计增益。将 DCG 除以完美结果下得到的理想结果 iDCG。

MAP：多查询评价，对所有查询的 AP 求算数平均。

MRR：第一个相关的文档的位置的导数 RR。对多个查询的 RR 做算数平均。

多样性评估：为什么：用户的单次搜索可能体现出多方面的需求。用户搜索可能存在歧义，需要展示多方面内容加以确认。要避免信息茧房的产生。核心思想：降低用户无法获得所需信息的风险，尽可能确保排序靠前的结果中至少有一个结果满足用户的需求。两种方式：隐式模型：只计算文档之间的差异性。显式模型：更加具体地考量文档所对应的用户意图。

Lec 6 排序（上）

信息检索模型的形式化表述：[D,Q,F,R(Di,q)]。D：文档表达；Q：查询表达；F：查询与文档间的匹配框架。R：查询与文档间的相关性度量函数。

布尔模型：D：词项的组合；Q：布尔表达式；F：完全匹配；R：满足布尔表达式。

Jaccard 系数：JACCARD(A,B)=|A∩B|/|A∪B|。将文档视作词项的集合。不足：查询词本身未做重要性区分、没有仔细考虑文档的长度因素、不考虑词项频率、未考虑罕见词比高频词的信息量更大。

词项频率 TF：查询词在文档中出现得越多，该文档越相关。TF 指词项 t 在文档 d 中出现的次数。不足：相关性与频率并不线性相关。改进：引入对数词频 wf(t,d)=1+lg tf(t,d) if tf(t,d)>0 else 0，数量级的差异性所造成的影响变得更为缓和。将文档与词项的匹配得分定义为所有同时出现在查询与文档中的词项其对数词频之和。

文档频率 DF：有些词在单个文档中出现的多，是因为这个词本身就很常用（eg 停用词）。罕见词的信息量更为丰富，而频繁词的信息量相对较少。df(t)指出现词项 t 的文档数量。idf=lg(N/df(t))。

TF-IDF：W(t,d)=(1+log tf(t,d))*log(N/df(t))直接相乘。在少数文档内多次出现的词更适合衡量文档相关性。

向量空间模型：每个文档和查询视作一个词项权重构成的向量，查询时通过比较向量之间相似性来进行匹配。步骤：首先，将文档与查询表示成词项的 tf-idf 权重向量。其次，计算两个向量之间的某种相似度。最后，按相似度大小进行排序，将 Top-K 的文档返回给用户。优点：简洁直观，可以支持多种不同度量或权重方式，实用效果不错。缺点：缺乏语义层面的理解和匹配，同时依赖 tf-idf 值也可能造成干扰（用户无法描述词项之间的关系，词项之间的独立性假设实际上不成立）。

查询与文档的匹配：欧氏距离：对于向量长度（文档长度）非常敏感。余弦相似度。

基于迭代的查询意图更新：用户的查询意图可能无法一蹴而就，而需要通过相关性反馈实现逐步更新。在本质上，这一过程是使查询意图的表达逐步逼近用户目标文档的过程。罗基奥算法使得查询尽可能离与之相关的文档更近，离与之不相关的文档更远。q_m=αq+β/|D_r|*sum{D_r}-γ/|D_nr|*sum{D_nr}，即原查询向量、相关文档质心、不相关文档质心的加权。通过设置 β>γ 来给予正反馈更大的权重，很多系统甚至只允许正反馈，即 γ=0。

Lec 7 排序（下）

Pointwise：将排序退化为分类或回归问题。基本假设：训练样本中的任何一个查询-文档对，都可以映射到一个分值或一个有序的类别（如优良中差）。往往更为注重文档的相关度得分，而并不注重文档之间的相关性排序。

Pairwise：将排序问题转化为分类问题（二分类或三分类）。每次比较一个查询与两个文档，衡量两个文档的偏序，分类器判断哪个文档应该排在前面（分类对应的标签为{1,-1}或{1,0,-1}）。优点：实现了从绝对相关性（分值）到相对偏序的改进。缺点：两两成对导致样本数大为提升，计算资源开支增加，受样本不平衡问题的影响，无法体现全局排序的合理性。

Listwise：直接面向整体排序结果进行优化，将排序的完整队列作为学习的对象。采用某种 IR 指标对排序进行优化，或设计面向完整排序的损失函数。

PageRank：将网页视作点，网页间的超链接视作有向边，从而形成一个巨大的有向图。网页入度越多，网页被引用和被推荐就越多，重要性就越大。公式：PR(pi)=(1-d)/N+d*sum{PR(pj)/L(pj)} for pj in M(pi)。L(pj)为 pj 发出的链接数量，d 为阻尼系数，N 为网页总数，M(pi)为链入 pi 的页面集合。计算过程：首先给每个网页赋予一个初值，例如 1/N，然后利用之前的公式进行迭代有限次计算，得到近似结果。Restart 机制：(1-d)/N 相当于以一定概率重新选择起点，跳出陷阱（仅指向自己）与黑洞（无出边），d 一般 0.85。收敛性：马尔科夫过程，A=dM+[(1-d)/N]E，则 P(n+1)=AP(n)。收敛条件：A 矩阵所有元素都大于等于 0，并且每一列的元素和都为 1；转移矩阵 A 为不可约的，当图是强连通时，A 为不可约，而 Restart 保障了这一条件；转移矩阵 A 为非周期的。

HITS：权威（Authority）网页与枢纽（Hub）网页的区分。权威网页：指某个领域或某个话题相关的高质量网页。中心网页：类似中介，指向了很多高质量的权威网页。HITS 的目的即在海量网页中找到并区分这些与用户查询主题相关的高质量“Authority”与“Hub”网页，尤其是“Authority”。计算过程：邻接矩阵为 M，Authority 向量为 a，Hub 向量为 h。迭代式：a_{k+1}=M^T*h_k，h_{k+1}=M*a_{k+1}。或 a_{k+1}=(M^T*M)^k*M^T*a_0，h_{k+1}=(M*M^T)^{k+1}*h_0。两初始向量可均设为全 1 向量。优点：能够区分网页功能的排序算法，更好地描述互联网组合特点；主题相关，因此可以单独用于网页排序。缺点：需要在线计算，时间代价较大；可能受到链接作弊的影响；初始页面的选择对查询结果有影响。

Lec 8 个性化检索（上）

基于内容的推荐：基本想法：用户的偏好一般相对稳定，因此给用户推荐他以前喜欢的项目。为实现基于内容的推荐，对于每个备选项目，需要给出相应的画像（往往以向量的格式存在）。用户的画像可由他曾经评分过的项目画像所估计。一般采用加权平均的方式得到用户画像向量（基于评分进行加权）。基于用户与项目画像，可采用相似性度量进行评分。优点：每个人的推荐过程相互独立，不需要其他用户的数据；可以为具有独特偏好的用户进行有效推荐，不受大众倾向性和热度的影响；可以推荐新项目或非热门项目；推荐结果有着较好的可解释性，可列举内容特征作为推荐的依据。缺点：找到合适的特征困难，特征的提取可能存在误导性；过度特化，信息茧房。

基于协同过滤的推荐：基于用户推荐：找到相似用户并基于历史行为推荐。Sim(a,b)=ab 向量的相关系数（两向量协方差除以两向量标准差）。Pred(a,p)=a 的平均分+对(a 的 n 最近邻给 p 的打分去中心化后)使用 sim 进行加权平均。无评分的项均跳过，相关性小于 0 的邻居不考虑。基于项目推荐：同一个人给两个项目打出相似分数，说明他认为两个项目相似。计算方法与基于用户推荐相似，但最终计算评分时不需要平均分修正（因为只针对该用户自己的评分算加权和）。基于项目推荐效果更好：项目的属性相对单一，而用户的偏好则更为丰富多样；某样项目受欢迎的理由相对固定，而用户可能在不同情境下体现出不同的偏好。优点：可适用于任意种类项目，不受多模态、非结构化信息表征与特征选取的困扰。缺点：冷启动、稀疏性（用户评分记录严重稀疏，很难找到评价过同一项目的用户）、热度偏差（更倾向于推荐热门项目，对具有独特偏好的用户推荐效果差，小众偏好很容易被热门偏好所淹没）

矩阵分解：评分矩阵 R 被近似视作项目属性矩阵 Q 与用户偏好矩阵 P 的乘积。估计两个潜在因子矩阵目标：最小化估值与实际值的差的平方和（误差平方和）min sum (r_xi-q_i*p_x)^2。问题：潜在因子的维度 K 过大，数据稀疏的情况下会导致过拟合。解决方法：引入更多训练数据；收缩参数，引入正则项 min sum (r_xi-q_i*p_x)^2+(λ_1*sum||p_x||^2+λ_2*sum||q_i||^2)。问题：潜在因子矩阵元素不允许为负。解决方法：引入噪声矩阵 E=R-PQ，最小化噪声矩阵。考虑优化噪声所对应的 2-范数，通过梯度下降法，迭代式：p_ik←p_ik+α_1[(R-PQ)Q^T]_ik，q_kj←q_kj+α_2[(P^T(R-PQ))]_kj，其中中括号部分为对矩阵元素的偏导。取 α_1=p_ik/[PQQ^T]_ik，α_2=q_kj/[P^TPQ]_kj，则由于 R 矩阵非负，在 P、Q 初始值非负的情况下，迭代一直非负。

概率矩阵分解：在数据稀疏且有噪声的情况下，引入某些规律，实现对于参数更好的描述。比如参数符合高斯分布。

Lec 9 个性化检索（下）

维度归约：删除不具有区分度的特征，同时可能降低噪声。避免维度灾难的同时，模型更容易理解，也更易于可视化。维度灾难：计算量呈指数级增长，难以处理；数据稀疏，没有足够数据可建模。方法：主成分分析、特征子集选择、为特征赋予不同权值。维度归约可能造成信息损失，甚至产生误导效果。

主成分分析：通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。通过这种方式，可以采用较少的综合指标综合先前存在于各个属性（且相关）中的各类信息，而综合指标之间彼此不相关。

最大特征值对应的特征向量可以最大化投影方差。要求得数据样本的最大的 K 个特征值，其特征向量所对应的线性组合就可以形成 K 个新的综合指标。K 个特征值的比重反应了主成分的信息量，一般应大于 0.85。

计算过程：1.标准化，计算每个特征的均值，原始特征减去均值；2.计算各个特征间的协方差矩阵 C；3.计算协方差矩阵的特征值与特征向量，并单位化特征向量；4.计算投影矩阵（特征向量按列拼接），投影。

主成分分析依赖于原始变量，也只能反映原始变量的信息，因此，原始变量的选择很重要。主成分分析的内在假设之一是原始变量直接存在一定的关联性。PCA 的结果未必清晰可解释，与选取的原始变量及数据质量等都有关。

情境感知的查询理解：动机：用户查询存在歧义、用语精简，缺乏精确性，借助情境信息协助判断用户意图。情境信息：所有与人机交互相关，用于区分标定当前特殊场景的信息，例如搜索的上下文。查询概念：一组有着相同语义的查询词。

数据采样：简单随机采用：对于所有对象，采用简单的等概率方式进行采样，分为有放回、无放回。分层采样：对数据进行分组，从预先指定的组里进行采样。样本容量：影响采样效果的重要因素之一；较大的样本容量更能完整代表数据，但降低了采样的收益；较小的样本容量在采样收益上更高，但可能造成信息的损失。启发式采样：分组采样，组内的数据高度相似，而不同组的对象差异性较大。渐进式采样：从一个小样本开始，然后逐步增大采样规模；在模型的准确率趋于稳定的时候停止采样，从而确定采样规模。优点：不需要在一开始就确定采样的容量。缺点：计算开销大（需要多次迭代）。

数据离散化：将连续属性变换为分类属性。非监督离散化：等宽、等频率、K 均值。有监督离散化：更注重问题导向，其目的在于取得更好的结果。基于熵的方法是最重要的有监督离散化方法之一。

Lec 11 知识图谱导论

信息抽取：从语料中抽取指定的事件、事实等信息，形成结构化的数据。从文本中获取用户感兴趣的事实信息，借助于自然语言处理技术，通常领域相关（借助领域知识辅助抽取）。“抽取实体，确定关系”。

实体：即命名实体，指文本中的基本构成块，如人、机构等。属性：实体的特征，如人的年龄、机构的类型等。关系：实体之间存在的联系，也称事实，如公司和地址之间的位置关系、公司与人之间的雇佣关系。事件：实体的行为或实体参与的活动。基本任务：命名实体 NE（实体抽取）：命名实体抽取是信息抽取最重要的任务。模板元素 TE（属性抽取）：模板元素又称为实体的属性，目的在于更加清楚、完整地描述命名实体。通过槽描述了命名实体的基本信息。共指关系 CR：如果不同的命名实体表达了相同的含义，即为共指关系，也称为等价概念。模板关系 TR（关系抽取）：实体之间的各种关系，又称为事实，通过关系抽取，将实体关联起来，并为推理奠定基础。场景模板 ST（事件抽取）：又称事件，是指实体发生的事件。常见的新闻事件描述模板 5W1H。

信息检索：从文档集合中找文档子集，通常利用统计与关键词等技术，通常领域无关。

知识图谱的表示：由结点和结点之间的边组成，结点表示概念（或实体），边表示关系（或属性）。关系：侧重实体（Entity）之间的关联，例如“高于”。属性：用于描述实体的特征，例如尺寸。点和边组成知识图谱的基本单位：三元组（实体-关系-实体）。

知识图谱的优点：知识图谱至少可从以下三个层面提升搜索的效果：1.找到最想要的信息：将信息直接呈现，无需用户劳动；2.提供最全面的摘要：对搜索对象进行总结，提供更完整的信息和关联；3.让搜索更有深度和广度：构建完整知识体系，使用户获得意想不到的新发现。

事理图谱：描述逻辑社会，研究对象是谓词性事件及其内外联系（区别），借助事理逻辑链接形成对于事件的推理（应用）。不足：关系的边界比较难以界定；对于事件关系的研究大量集中于因果关系等，对于其他关系的研究较少；事件的定义不够明确，从而引起抽取事件比较困难。
多模态图谱：实体和属性可能是多模态的（区别），表示与整合多模态知识（应用）。作用：模态知识互补：不同模态协同服务于实体的理解；模态信息搜索：以不同模态作为信息入口；模态语义增强：利用知识图谱增强多模态任务。属性多模态：优点：易于从现有图谱中进行扩展，即通用图谱可以轻松扩展，概念和关系不用改变；可以推理视觉知识。缺点：实体仍仅限于文本概念描述；1 个概念可以对应 N 张图片，但 1 个图片无法对应 N 张概念。实体多模态：优点：视觉语义信息丰富，场景多源化；关系丰富。缺点：图谱庞大，符号复杂。

Lec 12 知识抽取与表达（上）

命名实体识别是信息抽取中的核心任务，它往往包含两个子任务：判别实体边界；判别实体类型。难点：与分词的难点非常相似：不断有新的命名实体涌现；命名实体存在严重歧义；命名实体构成结构复杂；命名实体类型多样（如共指关系复杂）。
命名实体识别方法：基于词典：预先构建一个命名实体词典，词典中的词汇即识别为命名实体。优点：简单快速，与具体语境无关，容易部署和更新。缺点：难枚举所有的命名实体名；维护代价，实体歧义。
基于规则：手工构造规则模板，对符合规则的实体进行识别。优点：当提取的规则能较精确地反映语言现象时，性能较好。缺点：不同表达对应不同规则，规则库大；规则往往依赖于具体语言、领域和文本风格；代价太大，系统建设周期长、移植性差。基于统计：抽象为序列标注问题。四类标注：B（词的开始）、M（词的中间）、E（词的结束）、S（单字词）。分支一：基于分类的命名实体识别方法，将 NER 视作一个多分类问题，通过设计特征训练分类器的方法加以解决。分支二：基于序列模型的命名实体识别方法，与分词中的序列标注方法思路类似，区别在于标注的不同。
实体对齐：指对于异构数据源知识库中的各个实体，找出属于现实世界中的同一实体（如跨社交网络用户匹配）。基于表征的知识图谱实体对齐：利用相似性合并使关系表征拥有统一的向量空间。
实体链接：将文本中的提及链接到知识库中的实体上。方法：神经网络、预训练语言模型。

Lec 13 知识抽取与表达（下）

关系抽取：从文本中识别出两个实体（或多个实体）之间存在的事实上的关系。意义：搜索引擎发现和关联知识的重要渠道，知识库构建与知识关联的基础性手段，是支持问答系统、推荐系统等应用的有力工具。方式：1.基于规则；2.基于模式；3.基于机器学习。
基于规则：优缺点：1.通常针对特定领域的特定关系抽取任务，可以根据想抽取的关系的特点设计针对性的规则，但部分任务可能很难制定规则；2.基于手工规则的方法需要领域专家构筑大规模的知识库，这不但需要有专业技能的专家，也需要付出大量劳动，因此这种方法的代价很大；3.知识库构建完成后，对于特定领域的抽取具有较好的准确率，但移植到其他领域十分困难，效果往往较差。
DIPRE：基本元素：元组：表示关系实例，如<Foundation, Isaac Asimov>—<Title, Author>。模式：包含常量和变量，例如 ?x , by ?y 的形式（可表示“title”by“author”）。基本假设：元组往往广泛存在于各个网页源中；元组的各个部分往往在位置上是接近的；在表示这些元组时，存在着某种重复的“模式”。流程：首先，输入一组种子元组实例 R，如若干<title, author>的实体对；其次，基于种子实例集合 R，找到这些元组在网页中出现的内容 O（Occurrence），注意寻找的时候保留上下文信息（Surrounding Context）；进而，基于找到的元组实例 O，生成模式 P；最后基于生成的模式，找到更多的元组实例 R，此时可选择停止，或返回第二步继续基于新实例生成新模式（此时生成的新模式可能与之前的模式有所差异）。
Occurrence：元组在网页中的呈现形式，一般而言，只有元组的元素在网页中非常接近。模式：将同一关系的不同实例在网页上所呈现的不同 Occurrence 中，相同内容保留下来，不同内容采用通配符取代，即可得到近似的模式。将 URL 的前缀（Prefix）引入模式中，用于描述模式的限定范围。生成模式的基本步骤：首先，将 Occurrence 归纳为 Order（元素的顺序）和 Middle（中间部分）；其次，定义模式如下：模式的 Order 和 Middle，即为 Occurrence 集合的 Order 和 Middle；模式的 URLPrefix、Prefix、Suffix，分别为 Occurrence 集合中最长的公共（Shared）URL 前缀与前、后缀，其他部分采用通配符填充。
远程监督：思想：如果某个实体对之间具有某种关系，那么，所有包含这个实体对的句子都是用于描述这种关系。局限性：语义漂移，不是所有包含该实体对的句子都表达该关系，错误模板会导致关系判断错误，并通过不断迭代放大错误。优化方案：动态特征矩阵：引入一个动态转移矩阵，描述各个类之间相互标错的概率。在利用算法得到的关系分布的基础上乘以这一转移矩阵，即可得到相对更为准确的关系分类结果。规则学习：模拟远程监督的启发式标签过程，设计相应的否定模式列表 NegPat(r)，专门用于去除错误的标签，即某些关系的判断是否为错误。注意力机制：即使是被打入同一个包里的句子，不同句子对于训练关系到别模型的贡献度也不相同，这一贡献度可以采用注意力模型加以衡量。采用深度学习技术，获取对于整个句子的表示。进而，通过注意力机制，将最能表达这种关系的句子们挑选出来。
事件抽取：事件是信息的一种表现形式，其定义为特定的人、物，在特定时间和特定地点相互作用所产生的客观事实。基本要素：事件类型：与触发词相对应，往往可以通过触发词分类加以识别；事件元素：事件的参与者，主要由实体、时间等组成；事件元素角色：事件元素在事件中充当的角色。

Lec 14 知识图谱与图计算

图表示学习算法：将图数据进行向量化表征，映射到一个低维的向量空间，在这个低维向量空间中，图的结构特征和语义特征得到最大限度的保留。邻接图矩阵：每一行表示一个节点，1/0 分别表示与对应节点是/否连接，这一行可以视作该节点的一个表示向量；该思路可用于最基础的图聚类问题；局限性：未能充分融入节点结构信息，节点属性信息无法加入。基于随机游走的图表示学习：基于随机游走的邻居节点序列，挖掘图结构信息。基于图神经网络的图表示学习：利用神经网络来学习图结构数据，提取和挖掘图结构中的特征和模式。图神经网络技术：路径游走、面向表征整合的消息传递。

Lec 16 社会网络

独立级联模型：“独立”体现在，每次激活都是一次独立事件，相互不产生影响。同时，每个已激活节点，只有一次机会尝试激活他/她的未激活邻居节点。
如果某个节点在第 t 轮被激活，那么，他仅有一次机会，即仅能在 t+1 轮，尝试激活他所有未被激活的邻居节点。t=1 时，仅有种子节点可以尝试激活其他节点。整个传播过程直到所有节点都被激活，或没有新节点可以被激活为止。
对于节点 v 而言，他激活邻居节点 w 的概率采用 P_vw 表示。基本传播模型里，为简化考虑，一般将 P_vw 设为 1/N，N 为 w 节点的入边的数量。
线性阈值模型：将信息传递过程视作多人影响的叠加过程。一个用户会被某个信息激活，如果来自他已激活邻居的影响超过某个阈值。阈值预先设定，往往为从[0,1]均匀分布中随机抽取的一个数值（或根据用户对信息的兴趣等决定）。
对比：线性阈值模型与独立级联模型的区别：随机性。对于独立级联模型来说，其随机性在于抛硬币的过程。独立级联模型是完全随机过程，每一次的结果可能都不相同，一般需要重复多次以确定个体节点被激活的可能性。对于线性阈值模型来说，其随机性在于边权重/阈值的确定。如果采用启发式方法确定边权/阈值，则该方法结果完全由方法设计决定，一旦确定边权/阈值（无论何种方式），其结果具有唯一性。

信息传播最大化方法：PageRank 及其衍生模型、核心性（Centrality）度量、计算单个节点所能够激活的邻居数量，再进行排序。存在问题：在寻找“最具影响力的节点”时可行，在确定影响力节点集合时不可行（可能存在影响范围重叠）。
目标：找到一个节点集合 S，使得 f(S)的期望最大，并且|S|=k 只选择 k 个节点作为初始节点。
f(S)特性：子模特性。1.f(S)非负；2.f(S)单调非减，f(S+v) >=f(S)，新增加一个节点，至多不增加新激活，不至于减少；3.f(S)具有子模特性：对于任何集合对 S,T 且满足 S⊆T 时，给定节点 v，有 f(S+v)-f(S)>=f(T+v)-f(T)。
在 ICM / LTM 等模型定义下传播最大化问题可以归约为集合覆盖和节点覆盖问题，是 NP 难问题。
由于 f(S)函数具有子模特性，我们可以采用贪心算法近似求解：1.以空集合为起点，即初始 S=∅；2. 经过 k 次迭代，每次选择最大化 f(S+v)-f(S)的节点 v。贪心算法可以实现至少 (1-1/e) 的近似效果。
信息传播元素：发送者，也称作信息源或“种子节点”。指在信息传递开始时拥有信息的那一小部分用户集合。接收者，指作为潜在传播目标的广大用户集合。接收者集合的规模要远大于前者，且不同发送者的目标集合存在重叠。媒介，指传播过程发生的平台。

Proof. Consider an instance of the NP-complete *Set Cover* problem, defined by a collection of subsets $S_1, S_2, \ldots, S_m$ of a ground set $U = \{u_1, u_2, \ldots, u_n\}$; we wish to know whether there exist k of the subsets whose union is equal to U. (We can assume that $k < n < m$.) We show that this can be viewed as a special case of the influence maximization problem.

Given an arbitrary instance of the Set Cover problem, we define a corresponding directed bipartite graph with $n + m$ nodes: there is a node i corresponding to each set S_i, a node j corresponding to each element u_j, and a directed edge (i, j) with activation probability $p_{i,j} = 1$ whenever $u_j \in S_i$. The Set Cover problem is equivalent to deciding if there is a set A of k nodes in this graph with $\sigma(A) \geq n + k$. Note that for the instance we have defined, activation is a deterministic process, as all probabilities are 0 or 1. Initially activating the k nodes corresponding to sets in a Set Cover solution results in activating all n nodes corresponding to the ground set U, and if any set A of k nodes has $\sigma(A) \geq n + k$, then the Set Cover problem must be solvable. ■

1.4 假设有三个城市，编号分别为 1、2、3。现在有一个商人在三个城市之间来回穿梭，已知三个城市作为起点的概率分别为（0.2，0.4，0.4）。同时，这个商人在城市之间旅行或同城停留的跳转概率如下表所示：

初始状态概率向量

城市编号	->1	->2	->3
1	0.5	0.2	0.3
2	0.3	0.5	0.2
3	0.2	0.3	0.5

状态转移概率矩阵

同时，还知道三座城市各自晴天/雨天的概率如下表所示：

城市编号	1	2	3
晴天概率	0.5	0.4	0.7
雨天概率	0.5	0.6	0.3

观测状态概率矩阵

在某一次旅行中，商人连续三天观测到的天气状态是（晴天、雨天、晴天），请问，这三天内该名商人最有可能的旅行轨迹是什么？请给出计算过程。

三个城市

记 S 为所有可能状态的集合 $S = \{s_1, s_2, s_3\} = \{1, 2, 3\}$

$I = (s_{i_1}, s_{i_2}, s_{i_3})$ 是长度为 3 的状态序列，其对应的观测序列为 $W = (w_1, w_2, w_3) =$（晴天，雨天，晴天）

出行路径　天气序列

A, B, π 分别为隐含状态转移概率矩阵，观测状态概率矩阵和初始状态概率向量

(1) 初始化，在 $t = 1$ 时，对每个状态 s_i，计算状态为 s_i，观测 w_1 为晴天的概率，记此概率为 $\delta_1(s_i)$，则

$$\delta_1(s_i) = \pi_i b_i(w_1), i = 1, 2, 3$$

代入实际数据得，

$$\delta_1(1) = 0.2 \times 0.5 = 0.1$$
$$\delta_1(2) = 0.4 \times 0.4 = 0.16$$
$$\delta_1(3) = 0.4 \times 0.7 = 0.28$$

(2) 在 $t = 2$ 时，对每个状态 s_i，求在 $t = 1$ 时状态为 s_j 观测为晴天，并在 $t = 2$ 时状态为 s_i 观测 w_2 为雨天的路径的最大概率，记此概率为 $\delta_2(s_i)$，同时记录概率最大路径的前一个状态 s_j 所对应的下标 $\psi_2(s_i)$

$$\delta_2(1) = \max_j \{0.1 \times 0.5, 0.16 \times 0.3, 0.28 \times 0.2\} \times 0.5 = 0.028$$
$$\psi_2(1) = \operatorname{argmax}_j \{0.1 \times 0.5, 0.16 \times 0.3, 0.28 \times 0.2\} = 3$$

$$\delta_2(2) = \max_j \{0.1 \times 0.2, 0.16 \times 0.5, 0.28 \times 0.3\} \times 0.6 = 0.0504$$
$$\psi_2(2) = \operatorname{argmax}_j \{0.1 \times 0.2, 0.16 \times 0.5, 0.28 \times 0.3\} = 3$$

$$\delta_2(3) = \max_j \{0.1 \times 0.3, 0.16 \times 0.3, 0.28 \times 0.5\} \times 0.3 = 0.042$$
$$\psi_2(3) = \operatorname{argmax}_j \{0.1 \times 0.3, 0.16 \times 0.3, 0.28 \times 0.5\} = 3$$

$x_1 = (4, 1)$，$x_2 = (2, 3)$，$x_3 = (5, 4)$，$x_4 = (1, 0)$

$$X = \begin{pmatrix} 4-3 & 1-2 \\ 2-3 & 3-2 \\ 5-3 & 4-2 \\ 1-3 & 0-2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}$$

$$C = X^T X = \begin{pmatrix} 1 & -1 & 2 & -2 \\ -1 & 1 & 2 & -2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix} = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

特征值

$$\begin{vmatrix} 10-\lambda & 6 \\ 6 & 10-\lambda \end{vmatrix} = 0 \Rightarrow \lambda = 4, 16$$

归一化

特征向量：$\begin{pmatrix} 10-16 & 6 \\ 6 & 10-16 \end{pmatrix} w_i = 0 \Rightarrow w_i = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow W^* = (w_i)$　$X' = X \times W^*$

1.1 给定以下词项的 idf 值，以及在三篇文档中的 tf，已知总文档数为 811,400，请完成如下计算任务：

	df	tf@Doc1	tf@Doc2	tf@Doc3
Car	18,871	34	8	32
Auto	3,597	3	24	0
Insurance	19,167	0	51	6
Best	40,014	18	0	13

1) 计算所有词项的 tf-idf 值。
2) 试采用欧式归一化方法（即向量各元素平方和为 1），得到处理后的各文档向量化表示，其中每个向量为 4 维，每一维对应 1 个词项。
3) 基于 2)中得到的向量化表示，对于查询“car insurance”，计算 3 篇文档的得分并进行排序。其中，查询中出现的词项权重为 1，否则为 0。

1)

	tf-idf@doc1	Tf-idf@doc2	tf-idf@doc3
Car	4.135	3.109	4.092
Auto	3.476	5.601	0
Insurance	0	4.404	2.892
Best	2.947	0	2.763

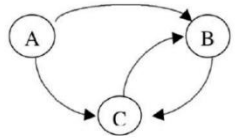
2)

$$\text{doc1} = \frac{\vec{W}_{t,d}}{|W_{t,d}|} = \left(\frac{4.135}{\sqrt{37.86561}}, \frac{3.476}{\sqrt{37.86561}}, 0, \frac{2.947}{\sqrt{37.86561}}\right) = (0.672, 0.565, 0, 0.478)$$
$$\text{doc2} = (0.400, 0.720, 0.567, 0)$$
$$\text{doc3} = (0.715, 0, 0.505, 0.483)$$

注意归一化方法，欧式归一

$$\vec{Q} = (1, 0, 1, 0)$$
$$Cosine(\vec{d_1}, \vec{Q}) = 0.475$$
$$Cosine(\vec{d_2}, \vec{Q}) = 0.683$$
$$Cosine(\vec{d_3}, \vec{Q}) = 0.863$$
$$\therefore Doc3 > Doc2 > Doc1$$

1.2 考虑右图的网络结构图



1) 当 Restart 部分的随机跳转概率为 0.15 时，写出 PageRank 的（随机）转移概率矩阵。
2) 计算各个节点所对应的 PageRank 值、Hub 值和 Authority 值。

(1) 归一化计算跳转矩阵R

$$R = \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 1 \\ 0.5 & 1 & 0 \end{bmatrix}$$

(2) 计算转移矩阵A

$$A = dR + \left[\frac{(1-d)}{N}\right] ee^T = 0.85 * \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 1 \\ 0.5 & 1 & 0 \end{bmatrix} + 0.05 * \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0.05 & 0.05 & 0.05 \\ 0.475 & 0.05 & 0.9 \\ 0.475 & 0.05 & 0.05 \end{bmatrix}$$

(4) 迭代计算Pagerank值

令每个节点的初始PageRank值为1/3

由 P_{n+1}=A P_n迭代计算可 P = $\begin{pmatrix} 0.05 \\ 0.475 \\ 0.475 \end{pmatrix}$

(1) 写出跳转矩阵

$$M = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

(2) 迭代计算并每步进行归一化

$$a_{k+1} = Mf h_k, \quad h_{k+1} = Ma_{k+1}$$

$$hub = \begin{bmatrix} \sqrt{6}/3 \\ \sqrt{6}/6 \\ \sqrt{6}/6 \end{bmatrix}, authority = \begin{bmatrix} 0 \\ \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$$