网络舆情分析

- 网络舆情概述
- 网络谣言
- 网络水军
- 话题检测与跟踪
  - 单边聚类算法
  - 基于平均分组的层次聚类算法GAC
- 社交网络事件检测

网络舆情概述

1. 网络舆情：指在互联网背景之下，众多网民关于社会（现实社会、虚拟社会）各种现象、问题所表达的信念、态度、意见和情绪表现的总和，即网络舆论和民情，简言之为网络舆情。
2. 网络舆情的特点：（1）直接性：**没有中间环节、随意性强**；（2）突发性；**无法预测**；（3）偏差性：**所表达观点与实际不符。**
3. 网络舆情传播：（1）信源：事件；（2）传播者：网民。
   - 传播特征：（1）一对一交流；（2）一对多交流；（3）多对多网络式交流。

网络谣言

4. 网络谣言：**在互联网中传播的、没有事实根据或凭空捏造的虚假信息。**
5. 特点：**传播速度快、周期短、波及范围广、表现形式多样、隐蔽性强、社会危害大、治理难度大。**
6. 类型：（1）政治谣言；（2）经济谣言；（3）军事谣言；（4）社会民生谣言；（5）自然现象谣言。
7. 辨别与检测方法：
   - 从**发布主体层面**分析：信息被“转手”次数越多，越易失真。**一手信源往往更有助于辨析真伪。**
   - 从**信息内容层面**分析：
     - **发生时间、地点、人物是否清晰具体且具备可回溯性**；
     - **信源是否多元、均衡**；
     - **物证是否可核查**；
     - **内容是否具备逻辑性、有无前后矛盾。**

网络水军

8. 网络水军：以**获取收益为主要诉求，受雇于公关公司或营销公司**，在短时间内通过**大量发帖、转帖、回帖**等方式满足雇佣者**建构舆论、制造荣誉**或**恶意抹黑**的特定需求。
9. 机器人水军：**扩散速度更高、传播量更大、覆盖面更广、具有病毒性传播特质。**
10. 运作模式：**需求方、中介方和服务提供方。**
11. 危害：（1）助推谣言滋生；（2）制造大量网络噪声；（3）导致社会经济受损；（4）诱发敲诈勒索等犯罪行为。
12. 检测方法：（1）**文本内容特征**、（2）**账号信息特征**；（3）**关系用户特征。**

- 文本内容特征：（1）具有**强烈感情倾向**；（2）群体活动以**评论、转发、点赞**为主；（3）**包含大量商业广告或垃圾信息。**
- 账号信息特征：（1）**创建时间较短**；（2）**名称较随机**；（3）**活动时间较集中。**
- 用户关系特征：（1）**大范围关注正常账号**；（2）“**回关**”率低。

话题检测与跟踪

13. 话题检测与追踪（Topic Detection and Tracking, TDT）：旨在没有人工关于的情况下，由机器自动判断新闻数据流的主题。
14. TDT任务：（1）报道切分；（2）话题检测；（3）首次报道检测；（4）话题跟踪；（5）关联检测。
15. 关联检测：
    - 基于**向量空间模型**：将报道表示成一个向量，然后使用**向量余弦距离**计算方法计算两个报道向量之间的相似度。最后将相似度与设定的**阈值**进行比较。
    - 基于**语言模型**：由**贝叶斯原理**计算报道S与话题T_j之间的相似度，也可使用词项分布间的相似度度量指标**K-L距离**来计算报道与话题之间相似度，也可计算话题与话题之间的相似度。
16. 话题检测：使用**聚类算法**来实现，可分为**在线检测**或新事件检测（New Event Detection, NED）和**回溯检测**（Retrospective Detection）。
    - 在线检测：输入**实时的报道数据流**，能够**在线判断**新报道是否属于一个新的事件。
    - 回溯检测：输入**所有时刻的完整数据集**，要求**离线地判断**数据集中报道所属的事件。

单遍聚类算法

17. **单遍聚类算法**：增量式的在线聚类算法。
    - 算法流程：
        1. 按顺序处理输入的报道，计算新报道与所有已知话题之间的相似度；
        2. 报道与话题的相似度为**话题中心向量或平均向量**之间的相似度。
        3. 反复执行，直到所有的报道都处理完，整个过程只读取数据一遍。
    - 改进：新闻报道的**时间特征**有助于提高在线话题检测的性能，数据流中**时间接近的报道更有可能讨论相同的话题。**
    - 优点：（1）**原理简单**；（2）**计算复杂度低**；（3）**支持在线计算。**

基于平均分组的层次聚类算法（GAC）

18. GAC：针对回溯检测的一种较好算法。
    - 特点：是一种自底向上的贪心算法，采用**分而治之**的策略。输出为**层次式话题类簇结构。**
    - 算法流程：
        1. 初始将文档集合中的**每篇文档当作一个单独的话题类簇**；
        2. 将所有话题类簇按顺序连续且不重叠地划分为大小为m的桶中；
        3. 对每一个桶进行聚类；
            - 重复合并桶内最相似的底层话题类簇，直到类簇数量比例减少到预设的p，或者任意两个类簇间的相似度低于某一预定义的阈值s为止。
        4. 保持事件顺序，去除桶边界，此时对文档集合的划分即为当前类簇集合；
        5. 重复2,3,4步，直到顶层话题类簇数目达到某预定的数值为止；
        6. 定期得将每个顶层类簇中的所有新闻文档按照前五步**重新聚类。**

社交网络事件检测

1. 突发特征检测方法：
    - **基于假设检验的方法**：

    > 假设在一个给定的窗口内，特征词w_k的生成概率服从**正态分布**：
    >
    > $$w_k \sim w_k^t = N(\mu_k^t, (\delta_k^t)^2)$$
    >
    > 其中$\mu_k^t = \frac{1}{L}\sum_{i=1}^{L} w_k^i$, $\delta_k^t = \sqrt{\frac{1}{L}\sum(w_k^i - \mu_k^t)^2}$。特征词$w_k$的频率大于阈值是**小概率事件**，因此若发生该情况，则说明该特征词是该窗口内的突发特征词。

    - **基于能量值**：考虑了频率和发帖者的权威度。

    > 根据过去几个时间窗口内的特征的权重值计算当前窗口内的能量值，增速越大能量值越大。

    - **Kleinberg方法（基于隐马尔可夫模型）。**
2. 事件检测：**根据突发词之间的关系构建关联图。**