Help collect dataset(s)

See marked TODOs

1. 【done】Basic learning about quantitative literary research => doc already
   1. 【TODO】(search for big data literature)
2. 【done but dataset requirements vague!!】Read their websites and log requirements
   1. Intelligent archive似乎比较image focus啊，比如古物文字识别，我不是特别感兴趣
   2. [TODO] What is **HASS RDC**?
   3. **AI Thematic Workbench/ Codifynd似乎还可以**
      1. Tag, 可视化等等，和gale那个viz也有点像，和我那个工具research rabbit也有点像，很多都这样
         1. 但他想主要做语义理解queries，直接show viz结果，不只是文献viz而已（‘who said what, where, in what context, and with what relationships’ ）（见下）
      2. Nodes & edges & **entities**
      3. **3D vizting挺有意思，我觉得可以和VR结合，形成学习机制**
      4. **主要features和use cases：**
         1. **和国家研究数据中心连接，可以查询研究尖端做到啥程度了做到哪里了，有利于researchers定位研究方向和尖端项目**
         2. **项目metadata。。**
         3. **跨学科。。。**
   4. [TODO]（我以前FB有什么entity insights可以给他？）
   5. [TODO get demo] Ask about a demo / someone to connect to so I can get API etc
      1. Able to auto-recognize dataset format and intent?
   6. What about dataset discovery? Want to support?
3. Choose among the following approaches:
   1. ~~The search engine approach: e.g. google dataset, gale dataset etc~~
   2. ~~The article writing approach: see notes~~
   3. The question/research approach: what am I interested in?
      1. Find a dataset
      2. [TODO] See photo for common research dataset engines / databases
   4. The helping approach: what are they interested in using, especially testing format?
   5. a simpler / faster ready dataset: <http://www2.informatik.uni-freiburg.de/~cziegler/BX/>
      1. good for building the entire system fast, but no use for me to get information about big book data
4. Collect dataset(s), each will:
   1. Input format
   2. [TODO] (cleaning or state that a notebook/pipeline is needed for the dataset)
   3. Output format (storage required?)