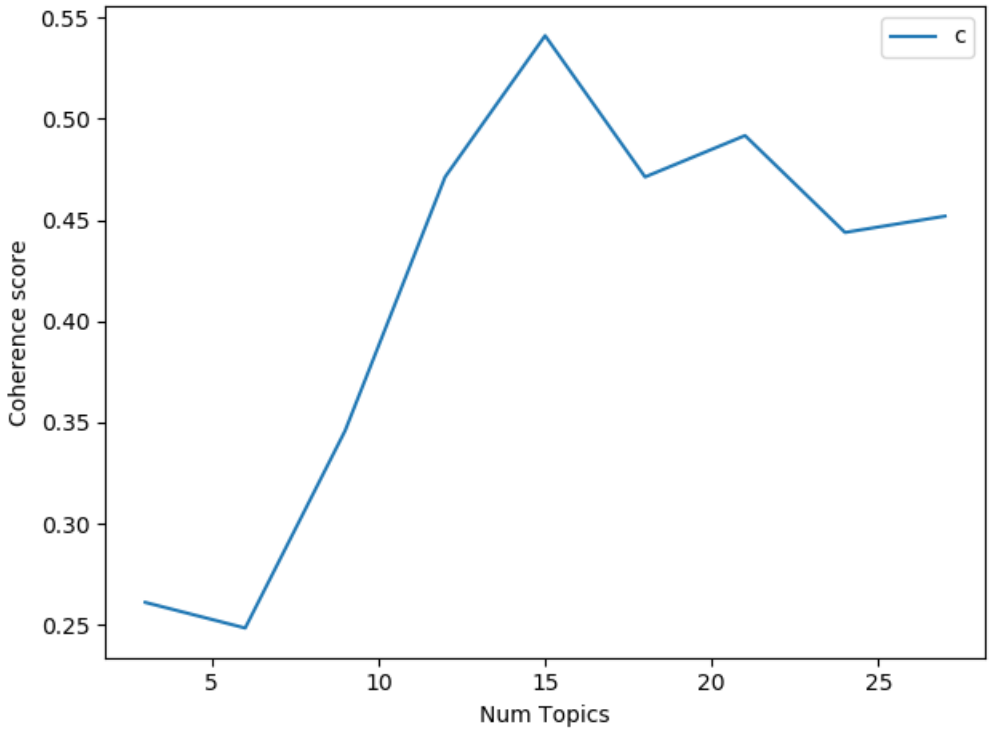
- Project Name
  - AI 领域专利结构分析
- Team
  - 范潇雄、李煜泽、马啸阳、尤予阳
- Problem statement
  - 分析提供的 AI 领域专利数据集，找出专利间的引用与依赖关系。
  - 引用关系可直接由专利局官网获取，而依赖关系较为隐蔽，需要分析专利语义，提取主题，获取相似性。
  - 以数据可视化方法展示上述关系。
- Data sources
  - 原始数据为专利 pdf 格式，在 patent-data 文件夹中，分为多个类别，文件名为专利号。
  - 美国国家知识产权局数据库 Patent Full-Text Database (PatFT)
- Design/methodology
  - 整个方案分为四个部分：数据获取、预处理、建模及训练、可视化
  - 数据获取：通过爬虫取得美国国家知识产权局数据库 Patent Full-Text Database (PatFT)中数据，存于 data 文件夹中，每个类别的专利存入一个 json 文件，每个 json 中包含一个由该类别的专利组成的数组，每个专利有 id、patent_code、patent_name、year、inventor_and_country_data、description、application_number、abstract、citations、related（即 prior arts）属性。
  - 预处理：使用正则表达式去除文本中的标点符号并将文本切分成单词，使用 gensim 库去除停止词并生成二元分词（bigram）和三元分词（trigram）模型，并将处理过的文本转换为 TF-IDF 形式的语料库模型。
  - 建模及训练：使用 LDA 与 HDP 文本主题模型对专利文本进行建模，详见 Algorithm/model 部分
  - 可视化：

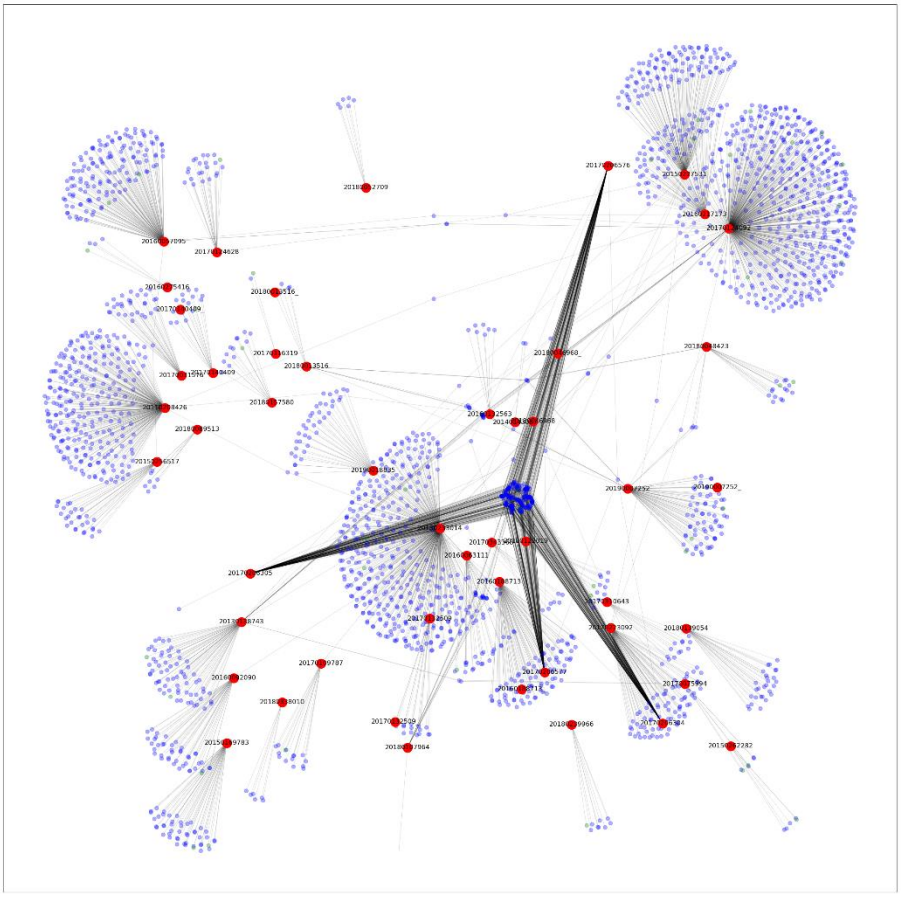
- 通过分析专利中的引用与关联信息，绘制引用图。
- 相似度图：模型训练完成后，计算其文本相似度矩阵，对于一对专利文本，当二者在两个模型下的相似度均超过阈值（LDA 模型取阈值为 0.7，HDP 模型取阈值为 0.8）时，即认为二者高度关联，由此获得一联合相似度矩阵。将相似度矩阵作为邻接矩阵画图
- 文本主题可视化

- Algorithm/model
  - 文本主题模型构建：使用预处理后的 TF-IDF 语料库分别训练 LDA 与 HDP 模型，其中 LDA 模型主题数定为 15，这个数值是通过主题相干性分数（coherence model score）确定的，一般认为相干性分数越高，模型约优。LDA 模型中主题数与 coherence value 的关系见下图，HDP 模型的相干性分数约为 0.71。使用 gensim 库进行训练，设置 LDA passes=2，这一数值为测试出能覆盖全部文本的最小 passes 值。

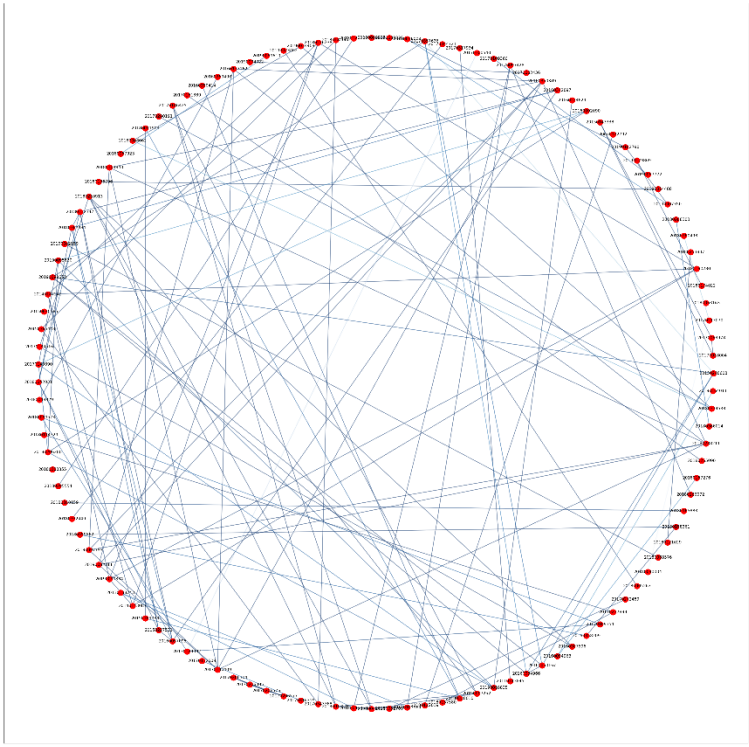


- Result
  - 引用图：利用爬取到的数据绘制了专利的引用关系图，图中红色点代表数据集中的专利并标注了专利号，蓝色的点为 citations 中出现的专利，绿色点为 related（即 prior arts）中出现的专利。如果被引用的专利是在数据集中的专利，它也会被标为红色，并在标签后加“_”加以区分。我们分别绘制了不同类别以及全部类别的专利引用图，以及一张数据集中专利间的直接引用关系图，从这几张图中可以明显的看出一些专利在第一层引用关系上就已经有一些重合。

    以下为引用图实例，详见 visualization/citation graph



– 相似度图：相似度高低以边的颜色深浅进行区分，相似度越高，颜色越深，见 visualization/joint_similarity.png



– 主题可视化：使用 pyLDAvis 对训练得到的 LDA 模型进行可视化，见 visualization/LDA_topics.html