Patent Citation Generation

Arvin Sahni, Sarah Kelley

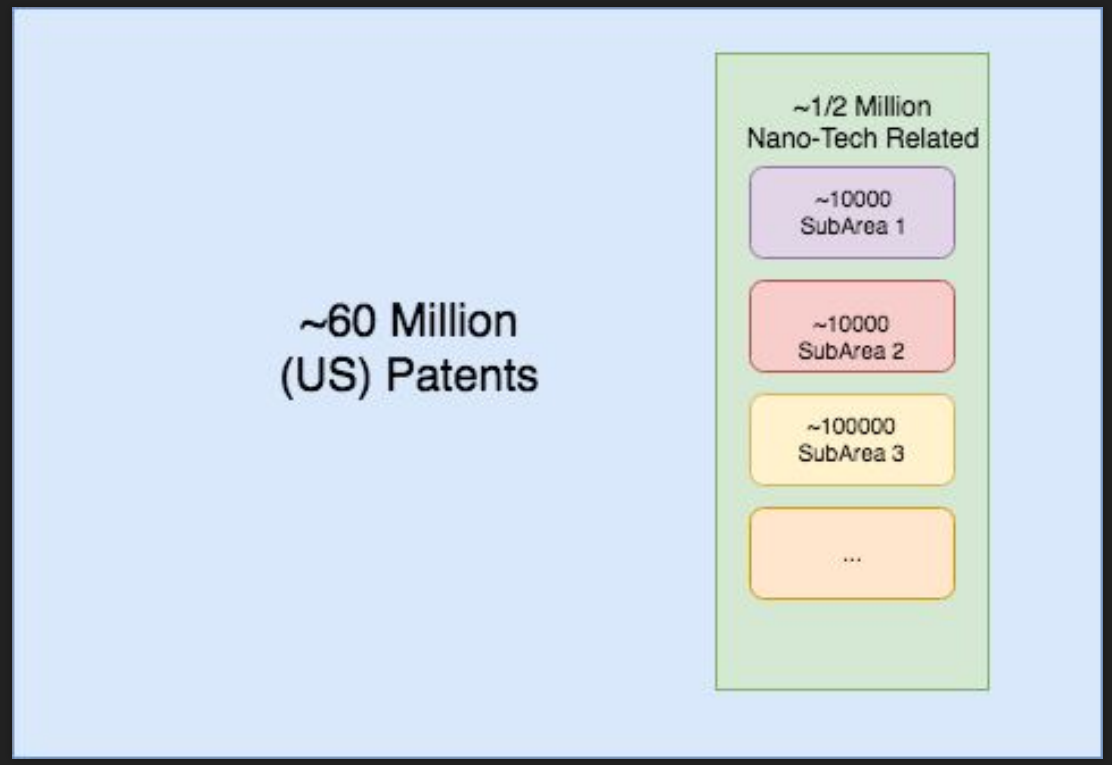
What are we trying to do?

- Patents have citations much like academic articles
- Importance/Practical applications

Text Data Available

- Patents have various text fields, we focus on title and abstract
- Possible extension to descriptions and claims

Scope of the Problem



- Difficult efficiently sort through 60 million records, limit natural language processing technique we could use
- Therefore focus on patents in one particular field (we chose nanotech)
- And use topic modelling to break it down into even smaller areas

Sampling NanoTech Patents

- Key word search strategy (Arora et al, 2013)
  - Series of words to match
  - Exclusion of records that have only nano-size or nano-organism words
  - Results (on journal articles) verified by nanotechnology experts
- Implementation:
  - Python: search over all 60 million records to generate list of related patents
  - BigQuery: select abstract, title, and citation information for the given patents
- Gives us 450 thousand patents to work with

Developing Sub-Area Groups through Topic Modelling

- Problem: Subdivide the nanotechnology patents into smaller fields which will provide pools of candidate patents
- Solution: parallelized implementation of topic modeling (to work on all half million records)
- Data Prep: snowball stemming
- Topic Modelling

Calculating Similarities

- TF-IDF Cosine Similarity
- Jaccard Distance/Similarity
- Second Degree Cosine Similarity*

$$sim2(d_i, d_j) = \frac{\sum_{m=1}^{N} sim1(d_m, d_i) \times sim1(d_m, d_j)}{\sqrt{\sum_{m=1}^{N} (sim1(d_m, d_i))^2} \times \sqrt{\sum_{m=1}^{N} (sim1(d_m, d_j))^2}}$$

where ,

sim1 (d_x, d_y) refers to the tf-idf cosine similarity measure between document d_x and d_y

N refers to total number of documents in the topic

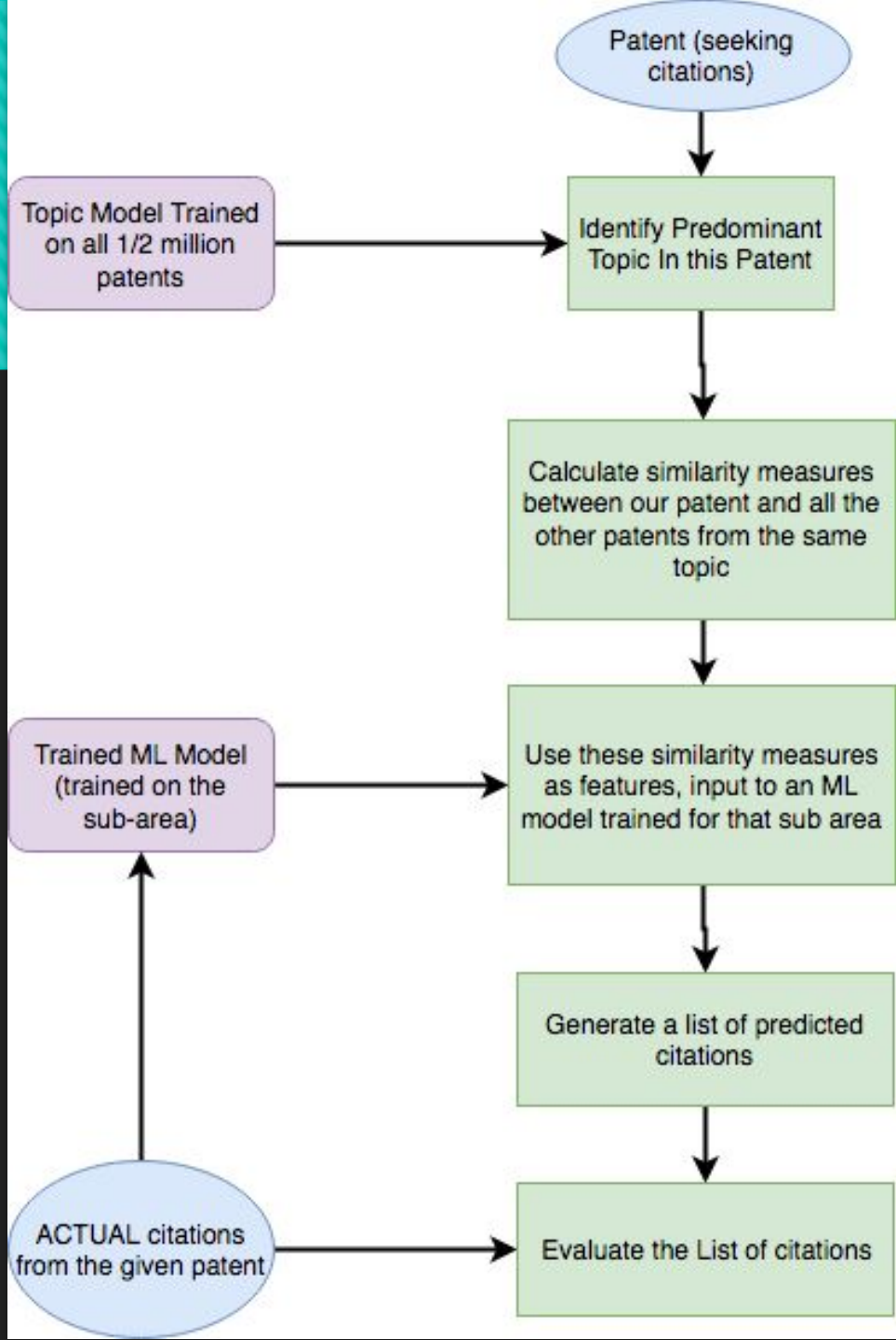
*Document-document similarity approaches and science mapping: Experimental comparison of five approaches - Per Ahlgren, Christian Colliander-Journal of Infometrics

Machine Learning on Similarity Features

- Logistic Regression Model
  - Classification : 0/1
  - Features : Similarity Measures

Prediction Pipeline

Topic Modeling → Similarity Measures → Logistic Regression → Predicted Citations



Model Evaluation

- Test sample had classification accuracy around 98.6% (improvement from 55% in initial iterations)

Further Extension

- Integrate structured information about the patents
  - Inventor/Company information
  - Structured patent data
- Refine topic modeling and/or similarity measure calculations using secondary topics
- Add other possible document similarity measures