Author: Prakash Chandra Chhipa

Problem (Assignment 1):

We have a set of 10,000 documents and we want to identify, Author, Method & Citation and want to find the relations among them. Describe the exact steps to do that. e.g. how pre-process, what do you need from manual annotation or data generally, what methods you use, what kind of ML task (s) this is and what are challenge and potential risks. Please detailed on what you would do.

Solution:

1. **Analysis & Observations:**

   By performing bird view analysis on given problem with respect to underlying requirements, below are the some ***observations*** which can be helpful to design the solution:

   1) Identifying tags e.g. author, method & citations in documents is very close to parsing problem by assuming the prior knowledge of document structure.
      a. A rule based parsing could only attain document for which structures are defined
   2) In order to support new structures or partially unstructured documents, a more comprehensive method is required to obtains the tags ex. Name-Entity Recognitions
      a. Manual tagging is also required for the documents where tags are not even present
   3) Document's actual content analysis is not mandatory to understand the relationship among documents if identified tags & their association can be stored efficiently.
      a. Relationship among documents are ultimately represented by relationship among authors for given corpus
      b. Citations are again a very rational tag which represents closeness of one document to other by citing one in another. Achieving some sort of correlation metric.
      c. By only analyzing tags relationship, we can have low cost solution in terms of infrastructure, development efforts and maintenance.
   4) In case, relationship among documents is more important and focused part then document similarity needs to accomplished using learning techniques e.g. topic modeling, cosine similarity on document term matrix, etc.

2. Solution Pipeline (work flow):

Below mentioned work flow showcases the task break down which typically includes textual data pre processing, document parsing and tag identification steps and exhaustive approaches for achieving relationship among documents.

Work flow diagram also depicts few decisions making in the pipeline with respect to document structuring, tag availability and consideration of intensive document categorization possibilities.

Work flow also segregates the nature of tasks such as rule based parser development to support various document structures. Other side machine learning based task e.g. Name Entity Recognition as

alternative approach to parsers and Topic Modeling approach to strengthen document similarity at the cost of more infrastructure and efforts.

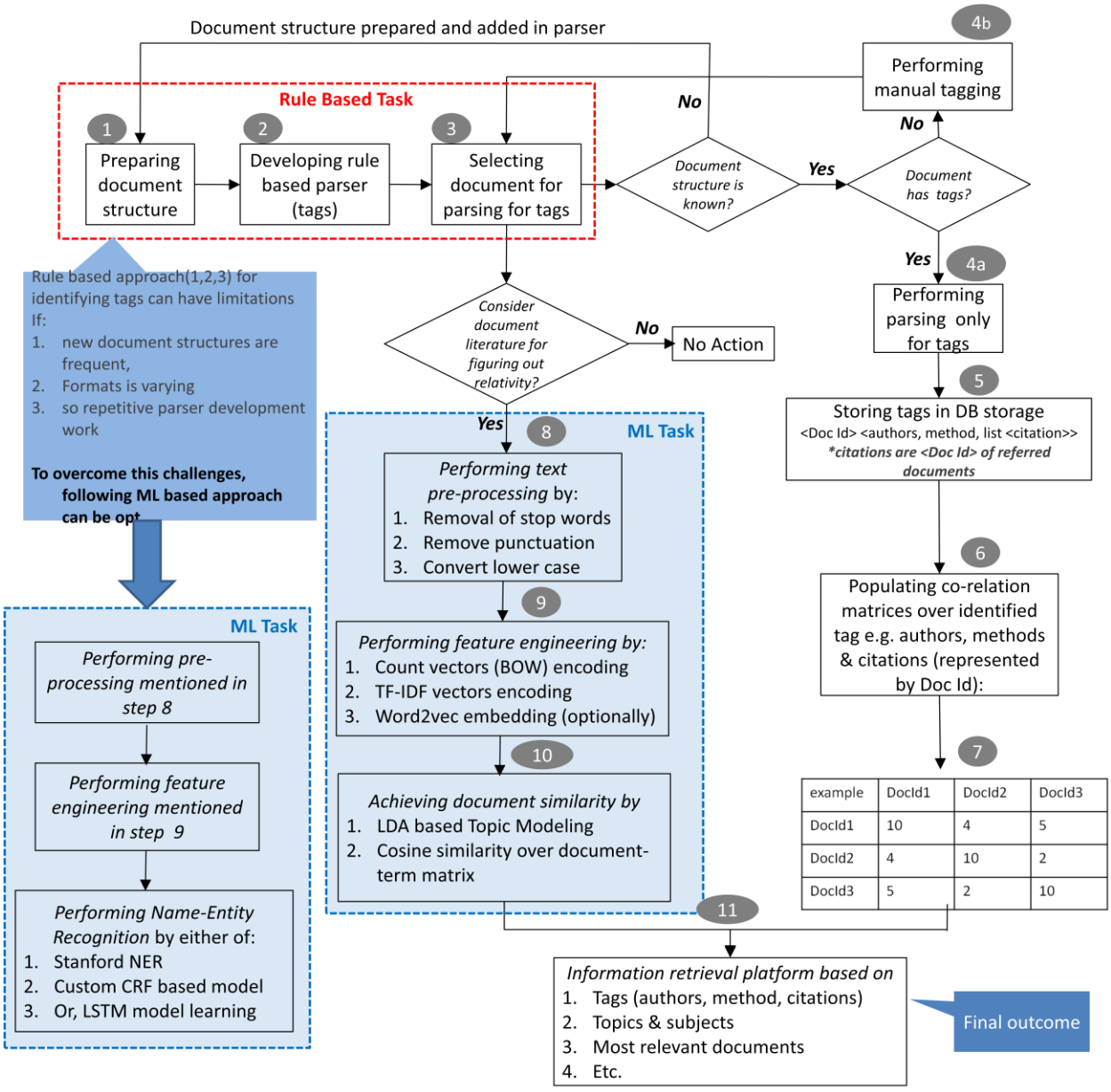

Figure 1: Work flow Diagram

2.1 Document Tagging:

Tags (authors, method & citations) identification part of the given problem is discussed with two possible approaches. Each approach has its own criteria to consider as main stream.

2.1.1 Rule Based Parsing for Tag Identification: This approach assumes the structure and format of the documents are fairly known in prior (depicted in step 1, 2 & 3 in work flow diagram). It also assumes that incorporations of new document structure are a rare event. By having this assumption in front, a tailor made textual data parsers are developed using libraries possible NLTKs (though it could be anything else also). Once parsers are developed then based on their customized rules, documents would be parsed to identify required tags (author, method & citations). In case overall platform decides to support a new format of document or new format of information tag then either customization is required to parsers or a new parser might be required to add up in family of parsers. These types of scenarios can be evident for examples, if previously citations were APA & MLA and now Chicago format has introduced.

2.1.2 Alternate Approach for Tag Identification (ML): Assumptions made in rule based parser approach are somewhat not so suitable in many cases where document base system is still evolving. In other words, if knowing document structure in well advanced becomes superficial and addition of new formats and structures in terms of citations or author details are very frequent. Then it is clearly a big overhead to maintain and develop rule based parsers. This kind of situations motivates to move forward towards learning based approach where unstructured textual data can be efficient processed to identify required tags. Here some documents posses' generic tag demands then already pre-trained Name Entity Recognition solution can be utilized effectively. For example, after performing textual data pre-processing (step 8) and typical feature engineering (step 9 in work flow diagram), it is advised to use available NER platform (https://nlp.stanford.edu/ner/). As state of art has proven performance records, CRF based NER are out-performing for name entity. In case of specific tag identification requirement, which are not so much used in open source prebuilt models, then custom made CRF model training is also possible at the cost of efforts, time and resources. Other alternative could be Bi-directional LSTMs for NER problem also possible and quite event in state of art (https://www.aclweb.org/anthology/Q16-1026)

2.1.3 Manual Tagging: We need to understand the one of bottleneck of current problem statement is unconditional requirement of manual tagging of document. In case, some of the documents are untagged, then we might not have choice but ot tag the required entities in documents manually by reading out and understanding it. It may take considerable time as there is no effect automation could be done.

2.2 Understanding Document Relationships:

2.2.1 Utilizing Tag information: At this stage when tags are identified by above mentioned methods or their combinations then the next challenge is organizing the tag details in effective ways. This phase of task is purely dedicated to store and arrange the tag details in either tabular or more structured way that it preserves the relations and relationship can be studied easily. One of the very straight forward method is to store the tags details (authors, methods and

citations as document id) in key value format in either in-memory databases or virtual memory based solutions(step 4a, 5 & 6 in work flow diagram). Here key represents document Id from which tags are identified and tags serves are value attributes. More specifically citation is nothing but reference of another document id in the provided corpus. Alternatively, such types of data records can be stored in structured database system e.g. MySQL where each records represent one document and its associated citations as relevant documents. Similar techniques can be implemented to retain author and method wise associatively also. This stored records are easily servers any programming platforms and libraries to build co relation matrix among documents by considering all the tag in all the possible combinations. One of the examples of such possibilities is shown in work flow diagram (step 7) where document similarity is counted by their citations to each other.

2.2.2 Utilizing Document Content (Additional ML approach): This particular step is not mandatory to the information retrieval platform to provide document similarities matrices. But in case, solution needs to be very exhaustive to analyzing document relationships then it is worth including. This step is data and computation intensive as it requires whole textual data to process. More specifically, a common pipelining task of textual data cleaning, pre-processing and features engineering has to be completed as mentioned in step 8 and 9. After achieving encoded and embedded features, here we learn the topic distribution among documents of the provided corpus. This task is being executed by training LDA model over the corpus. Learnt LDA model on training dataset of documents will be capable to provide topic distribution for any new documents. Finally based on similarity over topic distribution for documents, it becomes a very added advantage to figure out the document similarity in more comprehensive way. Keeping in mind, that this dynamic comprehensiveness is achieved at the extra cost of efforts and infrastructure.

2.3 Information Retrieval Platform

This platform is web based, GUI based interfaces which utilizes the exploited document relationships achieved by both the pipelining 2.2.1 Utilizing Tag information and 2.2.2 Utilizing Document Content. In order to support faster throughput, cache based solution could be used to store calculated matrices of document relevancy and learned topics patterns.

3. Challenges:

Designing & developing of such huge scale information retrieval solution can have multiple challenges with respect to different aspects. Some of the major challenges are:

1. Large corpus size – Performing operations on large corpus itself a big challenge statement it terms of resource scaling, electing distributed environments and result summarizations etc.

2. Manual Tagging: If data required to be tagged manually then it consumes linear time with many human resources efforts.
3. Supporting the change management is evidently tough as it demands lots of revision in data pre-processing and parsing operations.
4. Unsupervised Nature: Topic modeling could be very trickery job as it required extensive experimentation and involvement of intuitions to reach right number of topics and their interpretations.
5. Scalability while retaining the performance & throughput is sensitive trade-off to understand and judge.
6. Benchmarking of such system is itself very challenging and subjective matter.

4. Risk:

Risk associated with current systems could be broadly categorized in following manner:

1. Wrong Annotation of tags: In case documents needs to be tagged manually, it has potential risk of wrongly tagged documents as human error
2. Syntactic Failure of Parser(s): Mistakenly if parser parses tag data wrongly in specific circumstances which are barely possible to test then it has potential risk of wrong tag identifications. This type of risk are very crucial as it is seldom and periodic in nature. So have to be careful during the design and development of such components with multiple exhaustive reviews.
3. Ambiguity in Data: Ambiguity in names of author (examples authors are different but share same name, other way is same author but have two different name conventions) and years of publications in citations can mislead the analytics of document similarity. Again catching up and solving such errors are toughest job ever.
4. Hyper parameters of LDA in topic Modeling: Contextually inappropriate hyper parameters values in LDA can lead to diluted topics distribution where all documents are comprising average all topics presence or vice-versa where topics are not shared among documents.

5. Conclusion:

Though i have attempted to visualize the problem statement and hypnotized it with multiple possible scenarios but still it has many improvement scope as it goes to real ground of design and development with specific dataset.

Thanks & Sincere Regards,
Prakash Chandra Chhipa