Quality Assessment of Biomedical Metadata Using Topic Modeling

Stuti Nayak[1], Amrapali Zaveri[1], and Michel Dumontier[1]

Institute of Data Science, Maastricht University, Maastricht, The Netherlands,
firstname.lastname@maastrichtuniversity.nl

Abstract. There is an abundance of biomedical data present on the Web. However, this data is not re-usable because it is insufficiently described using rich metadata. The recently published FAIR principles specify desirable criteria that metadata and their corresponding datasets need to be Findable, Accessible, Interoperable, and Reusable. However, currently the biomedical metadata quality is poor which makes data re-use extremely difficult. To tackle this problem, we propose the use of topic modeling, specifically non-negative matrix factorization (NMF), as a first step towards dimensionality reduction when dealing with large amounts of data. In this position paper, as a use case, we apply NMF to the BioSamples metadata and present preliminary results.

Keywords: Metadata, Quality, Biomedical, NMF, Topic Modeling

1 Introduction

There is an abundance of biomedical data present on the Web [5]. This biomedical data is instrumental in enabling several medical use cases which should be shared and re-used by other investigators. In order to understand the structure of the data, there is an urgent need for accurate, structured and complete description of the data – defined as *metadata* . The recently published FAIR principles specify desirable criteria that metadata and their corresponding datasets should meet to be Findable, Accessible, Interoperable, and Reusable (FAIR) [14]. For data to be FAIR, metadata needs to be accurate and uniform (e.g., relying on controlled terms where possible), However, currently there is a large amount of biomedical metadata, which is of poor quality i.e. extremely heterogeneous and which makes data re-use extremely difficult [4]. Thus, we need to perform quality assessment of metadata to identify and ultimately improve the metadata quality. Currently, the challenges with metadata quality assessment are: (i) size of the data and (ii) heterogeneity of data.

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents [9]. In particular, topic modeling techniques allow examining a large set of documents and discovering, based on the occurrence frequency of the words, what the topics might be. The metadata elements are then associated to one or none of the topics, thus allowing one to easily detect erroneous ones.

Thus, the research question we aim to address is: *Can we use topic modeling to identify meaningful topics in biomedical metadata?*. By meaningful, we refer to being representative of the set of input metadata.

In this position paper, we aim to tackle the size and heterogeneity issues by using topic modeling techniques to discover the (metadata) 'topics' that occur in a large collection of (metadata) documents. Identifying topics will help reduce the large amount of heterogeneous metadata elements into smaller manageable clusters, which can be further analyzed for quality issues. The 'topics' identified as a result of applying topic modeling, will be considered as clusters of similar words. Thus, topic modeling is considered a first step towards data quality assessment.

2 Related Work

In [6] and [16], gene expression metadata quality assessment was performed on the Gene Omnibus Expression (GEO) database. The assessment was performed using (a) clustering methods and (b) crowdsourcing (i.e. non-expert human workers). In the former, the metadata (keys) were clustered based on (a) lexical similarity, (b) core concepts and (c) value similarities. In the latter, empirical analysis was performed on the same set of keys using crowdsourcing by submitting microtasks on the CrowdFlower[1] platform. While both methods were able to classify keys that contained the category term (e.g. 'disease state' in the category 'disease'), the clustering algorithm misclassified certain keys (e.g. stage in the category 'age') and there was low consensus amongst workers for key that could belong to more than one category (e.g. 'disease specific survival years' that was categorized either into the 'disease' or 'time' category).

In [4], a survey was carried for assessing the quality of Biosamples metadata. The analysis established that the quality of metadata in BioSamples is poor because of a lack of structured format and vocabularies to describe it. However, no further analysis was performed. In [15], topic modeling is used to understand gene expressions and build local gene networks. Supervised learning of gene expression data was performed and it was concluded that this method is a useful tool to computationally understand biological meaning from intricate and noisy gene expression data. Further related work includes the use of clustering and concept matching methods[2],[3] , [10], [12] for semantic matching of data. Our approach is novel as it proposes the use of topic modeling for tackling large amounts of heterogeneous biomedical metadata, as a firs step towards data quality assessment.

3 Topic Modeling

Topic modeling is a statistical technique which is used to discover abstract topics from a (large) collection of documents. Topic models help categorize heteroge-

[1] `crowdflower.com`

neous data and offer insights into large collections of unstructured text documents. The popular topic modeling techniques are: (i) Latent Semantic Indexing (LSI), (ii) Probabilistic Latent Semantic Analysis (PLSA), (iii) Latent Dirichlet allocation (LDA), (iv) Non-Negative Matrix Factorization (NMF) [9]. All the mentioned methods, except for NMF, use probability distributions to determine the topics whereas, NMF uses TF-IDF (Term frequency-Inverse Document frequency). The TF-IDF method uses term frequency to see how important a word is in one document whereas the inverse document frequency checks how many documents contain a term. Both these combined help in determining the relevant topics in a given input. The probability distributions tend to get complex and difficult to estimate therefore NMF is easier to use.

Our method uses the NMF algorithm to predict the topics. It is a topic modeling technique which outputs relevant topics from a large text. NMF was initially used for *learning parts of objects*. They use parts of a face as an input for NMF and then NMF learns the whole face from these parts [7]. NMF involves linear approximation for the data representation. It converts the input data into a document-term matrix ($\mathbf{X}$). Then it transforms it into 2 lower rank matrices topic-document matrix ($\mathbf{W}$) and topic-term matrix ($\mathbf{H}$). The algorithm then finds a lower rank approximation of $\mathbf{X}$ using the mentioned equation 1:

$$\mathbf{X} \approx \mathbf{WH} \tag{1}$$

. The two matrices W and H are the decomposition of original matrix X. The factorization is calculated by an optimization problem that minimizes a objective function given by a frobenius norm of the difference W and H.

$$(\mathbf{W}, \mathbf{H}) = \mathbf{argmin_{W,H}} \|\mathbf{X} - \mathbf{WH}\|_{\mathbf{F}}^{\mathbf{2}} \tag{2}$$

$\|.\|_F$ represents the Frobenius norm. This optimization problem was solved by [8]. NMF was chosen because it can represent the data using frequency of terms. In this work, the NMF algorithm was used from the python library scikit-learn [11]. NMF was applied to each of the three groups of the BioSamples metadata.

4 Methodology

4.1 Use Case

As a use case, we chose the BioSamples dataset owned by The European Bioinformatics Institute (EBI) [1]. It aggregates information of reference samples for e.g. Coriel Cell lines and samples for which data exists in one of the EBI's databases such as ArrayExpress. The dataset has a total of $1,332,354,592$ statements and is approximately 77GB in size. The metadata is structured in a format of attributes in the form of `key:value` pairs, examples of which are: `Organism: Homo Sapiens`, `Sex: Female`, `Disease: Cancer`. An example of a BioSamples Sample along with the metadata (attributes) is available at `https://www.ebi.ac.uk/biosamples/samples/SAMEA491372`.

4.2 BioSamples Metadata Extraction

We first retrieved the RDF [13] version of the BioSamples data (`https://www.ebi.ac.uk/rdf/datasets/`) and imported it into a local GraphDB (`http://graphdb.ontotext.com/`) repository. Then, we queried the repository using the SPARQL query language [2] to extract the metadata.

We extracted the keys and values according to the following four groups:

- Lexically similar values: values which have similar names. For example: `cancer` and `pancreatic cancer`.
- Lexically similar keys: keys which have similar names. For example: `disease` and `disease state`.
- Non-Lexically similar keys: keys which have similar meanings. For example: `clinical status`, `tissue`, `source`, `cell type` and `disease and its variants`.
- Non-lexically similar values: values which have similar meanings. For example: `heart attack` and `myocardial infarction` which are synonyms.

We applied NMF on the first three groups, the fourth group will be dealt with in future work. The example SPARQL query for extracting lexically similar values, lexically similar keys and non-lexically similar keys is shown in Listing 1.1. This query can be adapted to retrieve lexically similar values and non-lexically similar keys. These SPARQL queries are available at `https://github.com/stutinayak/BioSamples-Metadata-Quality-Assessment`.

```
PREFIX biosd-terms: <http://rdf.ebi.ac.uk/terms/biosd/>
PREFIX pav: <http://purl.org/pav/2.0/>
SELECT DISTINCT ?value
WHERE
{?smp  a                                  biosd-terms:Sample;
       biosd-terms:has-sample-attribute   ?pv;
       pav:derivedFrom                    ?webRec.
?pv    rdfs:label    ?value;
       dc:type       ?key.
FILTER (LCASE (STR ((?key) = "disease")))
FILTER (LCASE (STR ((?value) = "cancer")))
}
```

Listing 1.1. Example SPARQL query for extracting lexically similar values.

4.3 BioSamples Metadata Preprocessing Step

First, the stop words such as 'a', 'the', 'and' etc. were removed. Then, the metadata was converted into a vector as required for the input of the NMF algorithm. In order to convert the data in a vector form, the function *TfidfVectorizer* was used [11]. This function helps in removing the stop words along with two parameters *max_df* and *min_df*. max_df is used for removing terms that appear too frequently, also known as "corpus-specific stop words". For example: max_df = 0.50 means "ignore terms that appear in more than 50% of the documents". The default max_df is 1.0, which means "ignore terms that appear in more than

[2] `https://www.w3.org/TR/sparql11-query/`

100% of the documents". Thus, the default setting does not ignore any terms. min_df is used for removing terms that appear too infrequently. For example: min_df = 0.01 means "ignore terms that appear in less than 1% of the documents". The default min_df is 1, which means "ignore terms that appear in less than 1 document". Thus, the default setting does not ignore any terms. We assigned the max_df value to be 0.95 and min_df to be 2, which were close to the default to ensure that we removed the most and the least frequently occurring metadata terms. Then, we applied NMF on the processed metadata to extract the topics followed by an analysis of the results.

5 Preliminary Results

The source code, input data, SPARQL queries and results are all available at: `https://github.com/stutinayak/BioSamples-Metadata-Quality-Assessment.git`.

5.1 Lexically Similar Values

For each of the following listed value types, 1000 variants for each value were extracted. These specific values were chosen because these were the top most frequently occurring in BioSamples and are also relevant in the biomedical domain.

- Strain
- Species
- Source Name
- Host
- Cell Type
- Age
- Sample Type
- Sex
- Diseases
- Organisms
- Genotype
- Environmental Biome
- Disease State

The topics extracted of the key 'Sample Type' were: `21 cell line p4 stem passage mesenchymal labeled ferritin cells r60 rna reference p2 24`. While topics such as 'cells', 'cell line' and 'stem cells' are relevant, the topics '24' and 'p4' are extracted since 25% of the values are, for example, 'Mesenchymal stem cells labeled with ferritin dosper, cell line 24, passage P4'[3].

5.2 Lexically Similar Keys

We applied NMF for both (relatively) small and large set of keys. The number of keys for each key type in the small set of keys was 10. The larger set of keys had 100 keys of each key types. An overview of the number of the keys extracted for each type is in Table 1. For the smaller dataset, the resultant extracted topics were: `type cell thaliana stage species sample plant organism mutant model genotype disease columbia arabidopsis 2001`.

[3] This value is in the BioSample record: `https://www.ebi.ac.uk/biosamples/samples/SAMEA1326915`.

We observe that the main topics (the ones listed in Table 1) are identified correctly except for the key 'age'. Instead, the number '2001' is retrieved as a topic due to the occurrence of 'Arabidopsis thaliana (col-0) - dev.stage (Boyes et al. Plant Cell 2001)'. This is a value for the key 'sample type'[4] but is retrieved as a variant of a 'age' key due to the keyword 'stage'.

For the larger datset, the topics extracted were: `genotype host cell type survival status state stage species sample organism model free disease age`. We observe that while the main topics are identified correctly including 'age', topics such as 'survival', 'host', 'free' also appear when we increase the number of topics to be extracted in the output (in this case 15).

Table 1. Statistics of lexically similar keys extracted for each type.

Lexically similar key types	**Number of Keys**
Sample	100
Genotype	100
Species	35
Organism	38
Model	35
Disease	165
Cell Type	69
Age	100

5.3 Non-Lexically Similar Keys

Table 2. Statistics of Non-Lexically similar keys extracted.

Non-lexically similar key types	**Number of Keys**
Cancer	181
Heart Disease	22
Homo Sapiens	10
Mus Musculus	12
RNA	635
DNA	372

An overview of the number of non-lexically similar keys extracted is shown in Table 2. The topics extracted for the key 'disease' were: `patient clinical history groups disease death cause`. While the topics extracted are representative of

[4] present in `https://www.ebi.ac.uk/biosamples/samples/SAMEA416754`.

the sample of key variants in the input dataset, further investigation is required to verify the relevance of the keys *and* their respective values for each key type.

Overall, by investigating the topics retrieved for the keys vs. values, we observe that NMF performs better in extracting topics for keys as opposed to for the values. Also, the larger the dataset, the better it is for identifying the most important topics. Moreover, extracting topics based on specific types (all variants of the key 'sample type') is more accurate than for the entire set of lexically similar keys. We observe that NMF provides more meaningful results for keys rather than values.

6 Conclusions, Limitations and Future Work

In this position paper, we have proposed the use of topic modeling, particularly NMF, on BioSamples metadata to extract meaningful representative topics for different groups of metadata elements: (i) Lexically similar values, (ii) Lexically similar keys and (iii) Non-lexically similar keys. While we only show the feasibility of applying NMF on a small part (10%) of the dataset, it produced promising results in identifying meaningful topics for the heterogeneous metadata available. We observe that the results improve with large amount of data as an input which was also the motivation to use NMF as a data representation technique. With these topics, we can then divide metadata elements into different groups, which resolves the scalability issues for data observation and cleaning. The elements in the same cluster with duplicates and errors can easily be found. Thus, topic modeling is seen as a first step towards data quality assessment.

There are, however, some limitations of the method. The algorithm does not output keywords that are less frequently occurring, since it uses TF-IDF to predict the topics. This leads to losing out on important information. We propose to tackle this by adding weights on the less frequently occurring keywords as input.

In the future, extracting and analyzing non-lexically similar values will be performed along with scaling the method to the entire dataset for other metadata elements which have the same quality issues. We will also compare our method with baselines and current state-of-the-art methods. Moreover, we propose to add humans in the loop via crowdsourcing mechanisms to verify the topics identified for each metadata element. This, in turn, can be used as input for training, for example, machine learning algorithms to predict the topic for each metadata element. The ultimate aim is to assess and improve the quality of biomedical metadata so as to make datasets maximally FAIR.

References

1. Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., et al.: BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic acids research 40(D1), D57–D63 (2011)

2. Freudenberg, J.M., Joshi, V.K., Hu, Z., et al.: CLEAN: CLustering Enrichment ANalysis. BMC Bioinformatics 10, 1–15 (2009)
3. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic matching: Algorithms and implementation. Journal on Data Semantics IX, 1–38 (2007)
4. Gonçalves, R.S., O'Connor, M.J., Martínez-Romero, M., Graybeal, J., Musen, M.A.: Metadata in the BioSample Online Repository are Impaired by Numerous Anomalies. arXiv preprint arXiv:1708.01286 (2017)
5. Hoffman, S., Podgurski, A.: The use and misuse of biomedical data: is bigger really better? American journal of law & medicine 39(4), 497–538 (2013)
6. Hu, W., Zaveri, A., Qiu, H., Dumontier, M.: Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. BMC Bioinformatics 18(1), 415 (Sep 2017), https://doi.org/10.1186/s12859-017-1832-4
7. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
8. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Neural computation 19(10), 2756–2779 (2007)
9. Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W.: An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5(1), 1608 (2016)
10. Loureiro, A., Torgo, L., Soares, C.: Outlier detection using clustering methods: a data cleaning application. In: Proceedings of KDNet Symposium on Knowledge-based systems for the Public Sector (2004)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
12. Ulrich, B., Andreas, K., Sepp, H.: APCluster: an R package for affinity propagation clustering. Bioinformatics 27, 2463–2464 (2011)
13. W3C: Resource Description Framework (RDF). http://www.w3.org/RDF/ (2004)
14. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., andCarole Goble and-Jeffrey S. Grethe, P.G., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (2016)
15. Wu, S., Joseph, A., Hammonds, A.S., Celniker, S.E., Yu, B., Frise, E.: Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proceedings of the National Academy of Sciences 113(16), 4290–4295 (2016)
16. Zaveri, A., Dumontier, M.: MetaCrowd: crowdsourcing gene expression metadata quality assessment. F1000Research 6 (2017)