ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Topic Modeling using TF-IDF and Linked Data

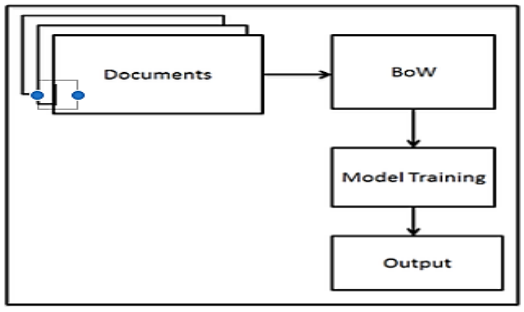
[1] Jali Sumalatha, [2] Dr.H.Girish
[1] PG Student, Dept of CSE,RYMEC College
[2] Professor, Dept of CSE,RYMEC College

Abstract: **- An information processing unit takes an input of raw data, this data is passed through various steps to generate useful information. Topic modeling is a process in which raw text corpus is passed onto a series of steps so that the document can be categorized into a set of topics.There are various methods for achieving topic modelling like Tf-Idf, LSI,LDA. These topics are then looked up onto global repository of linked information such as dbpedia. This paper explores ways to map local topics with global counterparts and retrieve useful information so as to develop intelligent systems capable of understanding semantics.**

***Index Terms*— Dbpedia, Linked Data, Topic Modeling, Tf-Idf.**

I. INTRODUCTION

In recent years, we are witnessing the rapid growth of digital data every minute in the internet world. All the data can be transformed into information when it can be managed efficiently. This task of extracting information and relations from huge corpus of data is a biggest challenge for the data scientists. Topic modeling is one of the compelling techniques for finding structure in the accumulation of records. Topic modeling is considered as an effective mechanism as it acts more than grouping approach. It can model objects as latent topics that can reflect meaning of the collection of document. The Topic is a collection of words that are likely to appear in the same context. It is a hidden structure that helps to determine what words are likely to appear in documents. By discovering hidden structure patterns of word use and connecting documents that exhibit similar patterns, a topic model has emerged as a powerful technique for finding useful structure. Topic modeling refers to an algorithm or method that identifies short and informative descriptions of a document in a large collection that can further be used for various text mining task such as summarization, document classification. The main significance of topic modeling is to find the structure of word use and how to link documents that share the same structure. So, the notion of topic modeling is that term which is dealing with documents and these documents are a combination of topics, where we can say that topic is a probability distribution over words. A topic model is a generative model for a collection of documents which de- scribes a simple probabilistic procedure. By using prob-abilistic procedure documents can be
produced and a new document generated by choosing a distribution over topics. Then, each and every word in that document choose a topic randomly depends on the distribution. After that, take a word from that topics[1] In the field of Information retrieval significant develop-ments have been made to achieve the goal of topic model-ing. The objective is to discover short depictions of the indi-viduals from a gathering that empower productive preparing of substantial accumulations while saving the fundamental measurable relationships. There are several approaches to achieve the goal.



1.Basic method
2. Tf-Idf method
3.LSI method
4.LDA method

1. Basic method

The essential system proposed by IR scientists for content corpora—a technique effectively conveyed in current Inter-net web crawlers diminishes each report in the corpus to a vector of genuine numbers, every one of which represents ratios of counts.

2. Tf-Idf method

In this method, a fundamental vocabulary of "words" or "terms" is picked, and, for each report in the corpus, a check is framed of the quantity of events of each word.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

After ap-propriate standardization, this term recurrence count is con-trasted with a reverse archive recurrence count, which measures the quantity of events of a word in the whole cor-pus. The final product is a term-by-archive grid X whose segments contain the tf-idf esteems for every one of the reports in the corpus.Thus the tf-idf conspire diminishes rec-ords of self-assertive length to settled length arrangements of numbers.

3.LSI method

To address these shortcomings, IR researchers have pro-posed several other dimensionality reduction techniques, most notably latent semantic indexing (LSI). LSI uses a singular value decomposition of the X matrix to identify a linear subspace in the space of tf-idf features that captures most of the variance in the collection. This approach can achieve significant compression in large collections. Fur-thermore, the derived features of LSI, which are linear com-binations of the original tf-idf features, can capture some aspects of basic linguistic notions such as synonymy and polysemy [2].

4. LDA

Latent Dirichlet Allocation (LDA) is the most effective topic modelling, which has been applied to information re-trieval and other application domains and achieved good performance. It is reasonable to expect that applying LDA to
IF could make a breakthrough for current IF models due to two advantages of LDA: first, the topic based representation generated by using LDA conquers the problem of semantic confusion compared with the traditional term based document representation. Second, LDA can describe documents at a detailed level with multiple topics instead of a single topic in traditional IF. However, directly applying LDA to IF using topic distributions to represent documents cannot produce satisfactory results due to limited dimensions in the topic representation, meanwhile, word based topic representation lacks distinguished semantic meaning [3].

II. TF-IDF METHOD

In this paper ,Tf-Idf method is used as a topic model-ling method to extract the topics from the document. Tf-Idf stands for term frequency-inverse document frequency.Term Frequency–Inverse Document Frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

The Tf-Idf method has two weights first, the Term Fre-quency, measures how frequently a term occurs in a docu-ment. Term Frequency (t) = Number of times term t appears in a document.Inverse Document Frequency, measures how important a term is. While computing TF, all terms are con-sidered equally important.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).

Inverse Document Frequency (t) = where, N is the total number of documents and Nt is the number of documents with term t in it.

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

For eg. Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency (i.e.TF) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (i.e. IDF) is calculated as log(10,000,000 / 1,000) = 4. Thus, the TF-IDF weight is the product of these quantities: 0.03 * 4 = 0.12. Term Frequency, which measures how frequently a term occurs in a document. Term Frequency (t) = Number of times term t appears in a document [4].

Let us consider term t and document d. then, t exists in n of N documents in D. The Tf-Idf function can take the form of

$$TFIDF(t,d,n,N) = TF(t,d) \times IDF(n,N)$$

TF(Term Frequency) can take the form of

$$TF(t,d) = \begin{cases} 1 & \text{if } t \in d \\ 0 & \text{else} \end{cases}$$

$$TF(t,d) = \sum_{word \in d} \begin{matrix} 1 & \text{if } word = t \\ 0 & \text{else} \end{matrix}$$

Inverse Term Frequency can take the form of

$$TFIDF(t,d,n,N) = \left(\sum_{word \in d} \begin{matrix} 1 & \text{if } word = t \\ 0 & \text{else} \end{matrix} \right) \times \log\left(\frac{N-n}{n}\right)$$

The combined weights of TF-IDF can take the form of

$$IDF(n,N) = \log\left(\frac{N}{n}\right)$$

$$IDF(n,N) = \log\left(\frac{N-n}{n}\right)$$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

III. EXPERIMENTAL SETTINGS

Setup

A corpus of dataset of around 1000000 questions published by quora in a kaggle competition is used. The data is the following format.

"test_id","question1","question2"
0,"How does the Surface Pro him self 4 com pare with iPad Pro?","Why did Microsoft choose core m 3 and not core i3 hom e Surface Pro 4?"
1,"Should I have a hair transplant at age 24? How m uch would it cost?","How m uch cost does hair transplant require?"
2,"What but is the best way to send m oney from China to the US?","What you send m oney to China?"
3,"Which food not em ulsifiers?","What foods fibre?"

There are some preprocessing steps to be done before the data can be used:-

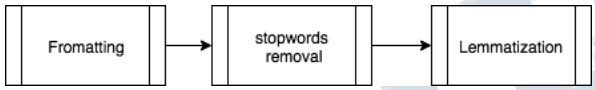


Figure 2 : Pre-Processing Steps

1.Formatting

We have taken only question1 in the above field. So the preprocessing step involves taking the question1 and formatting it in an array which can further be processed.

2.Stop-words removal

This is one of the basic steps. In this step a set of common words which are used in the language is removed. This words makes semantic sense in the grammar but for the algorithm it is not required. The formatted array is passed on to this step so as to remove this words.

i
me
my
myself
we
our
ours
ourselves
you

3. Lemmatization:-

A word can be used in multiple ways in sentence for eg:- car, cars, car's, cars' car All the above words on the left hand side point to the word car. So removing them is important so as to reduce the di-mensionality. Note that we are not using stemming here since it is possible that the actual meaning of the word will be removed by this process.

Methodology

The cleaned data is sent to the tf-idf algorithm as an input. The output of the algorithm is stored into a file. This file contains keywords which are in between of the max repeated as well as min repeated ranges. It takes into account an n-gram factor of 4. The output is a text file as below. The size of this file was around 304677 n-gram to-kens.

enough
enough live
enough live comforta-
bly
euros
euros good
euros good enough
every
every month
feet

Linked Data category list

A copy of dbpedia category for the English lan-guage is used to do a lookup of global categories. This can be downloaded from the official dbpedia website in various format. The format used for the purpose of this paper is the RDF turtle format. Let us go through some common terms in related with linked data.

Linked Data

This is also called the sematic web of data. This is a collection of structured data having various attributes which are globally collected an inferred. It also has various attributes related with the keyword. So any keyword can have a lookup in the Linked Data repository and find if they match with a global keyword.

RDF (Resource Description Framework)

It is an open model for data interchange on the web. It has features to join multiple schemas. If there are different attributes between different categories still this format will be able to merge them and provide a format capable of integrating those.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Turtle(Terse RDF Triple Language) :- This is a format which is used to express RDF. Its made up of triples. This contains three parts which are

1. Subject
2. Predicate
3. Object

A dot(.) at the end of the sentence indicates a delimiter. Example:-

<http://dbpedia.org/resource/Category:Futurama> <http://www.w3.org/2000/01/rdf-schema#label> "Futurama"@en .
<http://dbpedia.org/resource/Category:World_War_II> <http://www.w3.org/2000/01/rdf-schema#label> "World War II"@en .

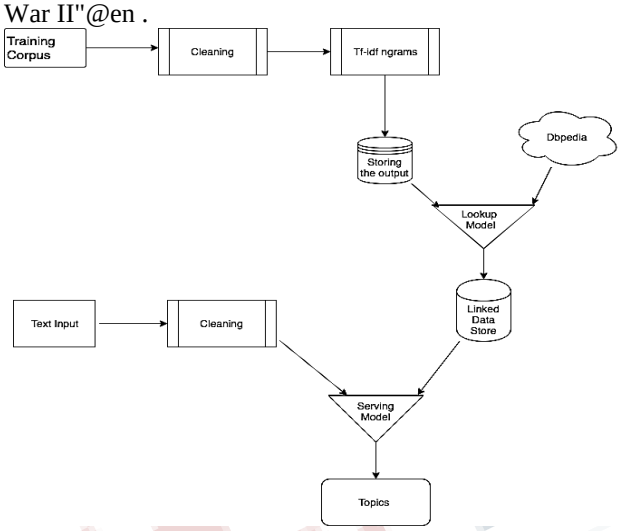


Figure 3 : Topic modelling using tf-idf ngrams and dbpedia

Once the corpus is cleaned and tf-idf with n-grams is taken into consideration, the next step is to store this output into a file or a database. It is then used as a lookup against Dbpei-da categories and the n-grams which are found are then stored into mongodb database. This becomes our linked data store. If required further information regarding the same can be found by using the respective Dbpedia uri's. The serving is a process in which the existing model is used for new data. Here say if a line of text is given as an input, the output will be the global categories which is represented by the model. Also for an actual use case, we can give a description of the topics.

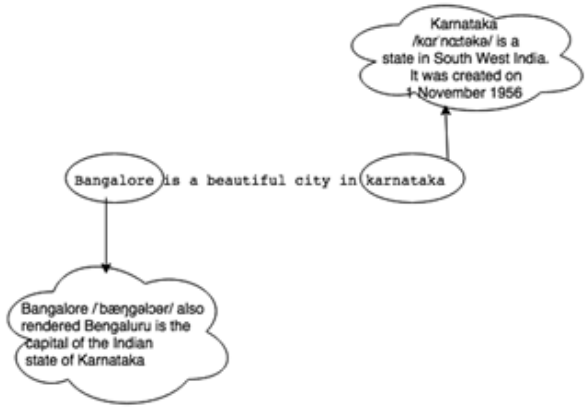


Figure 4 : Dbpedia Topics extracted

As shown in the above figure, a simple text sentence is bro-ken down into multiple n-gram tokens. Each token is checked with its corresponding keyword/description inside the no-sql reference. If found it is sent back as an output. A common way to implement this can be using a python serv-er/flask. The input of this server will be a text and the output will be the json array of topics.

Eg:-

Input

```
{"text" : "bangalore is beautiful city in Karnataka"}
```

Output

```
[
  {
    "topic": "bangalore",
    "description": "Bangalore /ˈbæŋgəlɔːr/ also rendered Bengaluru is the capital of the Indian state of Karnataka."
  },
  {
    "topic": "karnataka",
    "description": "Karnataka /kɑrˈnɑːtəkə/ is a state in South West India."
  }
]
```

IV. CONCLUSION

This paper describes a process of connecting local corpus data with the global linked data. This enables local system data to have more meta-data information structuring based on global data.The current implementations deals with se-mantic topic modelling but it can be extended by using neu-ral networks and

algorithm's like vord2vec or fasttext which can understand contextual meaning of the input sentence.

REFERENCES

1. B. V. Barde and A. M. Bainwad, "An overview of topic mod-eling methods and tools," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017

2. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research

3. Pattern-based topics for document modelling in information filtering,Gao Y, Xu Y, Li Y, IEEE Transactions on Knowledge and Data Engineering, vol. 27, issue 6 (2015) pp. 1629-1642 Published by IEEE Computer Society

4. Mishra, Apra, and Santosh Vishwakarma. "Analysis of TF-IDF Model and Its Variant for Document Retrieval." 2015 In-ternational Conference on Computational Intelligence and Communication Networks (CICN), 2015, doi:10.1109/cicn.2015.157.