EasyChair Preprint

№ 2050

Influencing Trends for Neural Information Processing Systems

Akshay Lunawat, Shreyas Sawai, Sairaj Nanaware, Jhanvi Bhatia and Swapnil Jawahire

November 29, 2019

Influencing Trends for Neural Information Processing Systems

Akshay M. Lunawat[1], Shreyas D. Sawai[2], Sairaj V. Nanaware[3], Jhanvi M. Bhatia[4]

Department of Computer Engineering, Vishwakarma Institute of Technology
Upper Indira Nagar, Bibwewadi, Pune

[1]akshay.lunawat18@vit.edu [2]shreyas.sawai18@vit.edu [3]sairaj.nanaware18@vit.edu
[4]jhanvi.bhatia18@vit.edu

Prof. Swapnil M. Jawahire

Department of Computer Engineering, Vishwakarma Institute of Technology
Upper indira nagar, Bibwewadi, Pune

swapnil.jawahire@vit.edu

Abstract*— Neural Information Processing Systems (NIPS) is leading Machine Learning and Computational Neuroscience conference in the world where innovative work is published. The NIPS include over million research papers describing various projects in the latest technologies. However, there are various problems to find out the trending topics in recent technologies for studying and research purpose due to its large amount of data. Consequently, Data Analytics may pinpoint the evolving trends within technologies like Machine Learning, Artificial Intelligence, etc. Natural Language Processing techniques allow us for relatively rapid and largely automated analysis of large collections of text data. The objectives of our project are to carry out a review of the papers accepted at NIPS in the 1990-2015 decade, analyze and characterize the topics and trends in technology and identify challenges for the evolution of the area in the near future. In our project, we will review the title, the abstract, and the keywords of the papers that are published in NIPS and identify the technological topics involved in these research works and will try to provide a classification of these papers in trending technological perspectives and obtain the timeline of these topics in order to determine interest growths and declines. The techniques used here to handle large amount of data, can also be applied to other text datasets as well.

***Keywords*— NIPS, Natural Language Processing, LDA, Bag of words.**

I. INTRODUCTION

Do you know? The only fix thing in the world is itself a "Change". From operating Nokia phones to gradually adopting MI phones, everything transformed with a rapid growth. In this transforming world, it is mandatory to be routinely updated. These changes with respect to technical world are represented by making use of trends. Trends describe the technologies which are currently in demand. In this paper we are addressing influencing trends in NIPS.

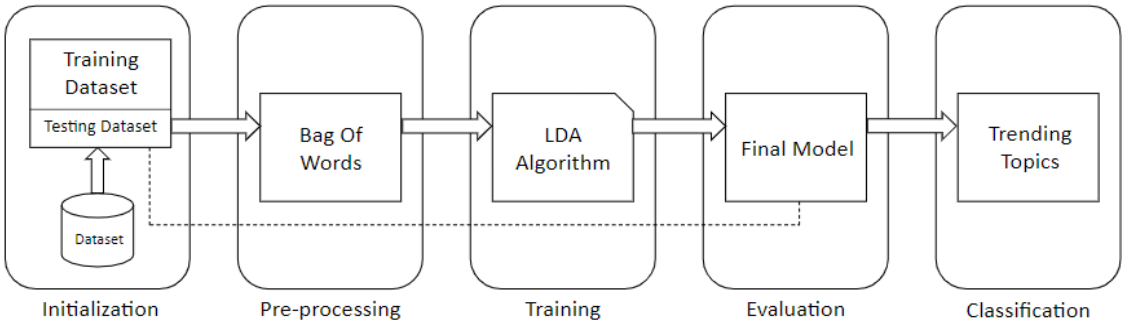
The system work on research papers collected from NIPS conference. NIPS conference is the biggest paper publishing conference of Machine Learning. The technologies we are implementing in this system are Natural Language Processing (NLP), Bag of Words, Latent Dirichlet Allocation (LDA) and many more. NLP is a sub-field in Artificial Intelligence and Machine Learning which works on interaction between computer and human languages. In particular, NLP is about how program computes large amount of natural language data. The techniques used in NLP to perform any task are syntax analysis and semantic analysis. In this system, the technique used is syntax analysis.

In NLP, the processing for output cannot be done directly on text. The processing is done on numbers. Hence, bag of words (BoW) is used to preprocess the text into bags which keeps the total count of particular word in the given text. To make bags, first all the given input is processed and unnecessary words and punctuations are removed and text data is converted into lower case. Then every word is tokenized and a dictionary is declared to hold BoW. Now for ever word encountered, it is checked that the word exists in the dictionary or not. If the word exists, the count is incremented by 1.

After BoW, LDA is used to extract trending topics for NIPS. In LDA, each document is considered as combination of topics and each topic as combination of words. LDA mathematically estimates both of these at the same time: finding the combination of words that is associated with each topic, while also determining the combination of topics that describes each document. The process of system is carried out in two phases i.e. Preparation Phase and Execution Phase. Starting for preparation to execution it has many steps which involves Collecting & Analyzing data, Project Planning, Creation of Project Architecture, Data pre-processing are the steps involved in Preparation phase. Execution phase includes Data Training, Data Testing, Accuracy checking and Deployment.

SR NO.	TITLE	YEAR	AUTHOR	DESCRIPTION
1.	Sensing trending topics in Twitter.	2013	Luca Maria Aiello et.al	Novel techniques are best for heterogeneous and NLP are best for homogenous data.
2.	Popular Research Topics in Marketing Journals.	2014	Yung-Jan Cho et.al	Focus on quality of data to get more accurate result.
3.	Natural Language Processing – A solution for knowledge extraction from patent unstructured data	2014	Achille Souilia et.al	This paper analyzes dataset and maintain the count of keywords to predict the trending patent projects.
4.	Topic discovery and future trend forecasting for texts	2016	Jose L. Hurtado et.al	This paper describes methods to find out future trends from a given set of document.
5.	Using NLP on news headlines to predict index trends	2018	Marc Velay et.al	This paper predicts trends from news headlines by using Natural Language Processing

Table 1.



.

Fig. No. 1 System Architecture

II. LITERATURE SURVEY

Natural language processing is rapidly evolving technology that is widely used in analyzing human speech. As we are using NLP in our system, we have studied 5 different papers and journals in order to understand how more accurate results can be obtained. The Table 1 summarizes the studied papers.

III. SYSTEM ARCHITECTURE

The fig. no. 1 depicts the system architecture along with overall working of the system. There are 7000 records in The dataset used in this system. Out of which 5000 records will be used as training records and remaining 2000 will be used as testing records. The training dataset as well as testing dataset will o through the steps shown in fig. no. 1.

IV. METHODOLOY

After brief study of various papers we led to two conclusions. First, we had implemented Bag of Words Technique which one of the best technique for data pre-processing and for digging out trending words we had implemented Latent Dirichlet Allocation (LDA) model. As our data is a homogeneous kind of data LDA stood mostly among top. Now lets forward towards detail study of working models. Our LDA model requires text data in the form of pure words (No punctuation, wide spaces, etc). But, the input of dataset is not in the pure form. It contains various spaces, punctuation, different cases of letters and many things. To make out pure words from the

sentences we implemented Bag of Words. Bag of Words is a technique which is easy to understand and implement. It extract out features (words) from given input. Lets understand how bag of words works internally.

In the bag of words technique, all the words are known as token and it implements tokenization to identify frequency of each words. For example, “It is awesome thank you”, “It is great thank you” and “It is not bad not good”. Consider each word as different document and remove all unnecessary punctuation and words. We get, it, is, awesome, thank, you, great, not, bad, good. Now we create vectors, it converts text that can be use by ML. Frequency of words in all documents will be like:
It = 1, is = 1,…… good = 0
1) [1,1,1,1,1,0,0,0,0]
2) [1,1,0,1,1,1,0,0,0]
3) [1,1,0,0,0,0,2,1,1]
The word token is also named as gram and tokens with a pair of two is called bigram model. The process of converting NLP text into numbers is called Vectorization. Multiple ways to convert NLP text into vectors are:
1. Total number of time each word appears in a document.
2. Total number of time each word appears in a document out of total words in the document.

Now, moving further towards LDA, Latent Dirichlet Allocation (LDA) is a “generative probabilistic model” of collection of words. It generates topics from documents and each topic consists of multiple words. LDA is method which automatically detects topics from the documents. From above example words we have, awesome, thank, you, great, not, bad, good, etc.
From above words, LDA might classify the blue words as Topic A, label as “Greetings”. Red words as topic B, label as “Feelings” and purple words as Topic C, label as “negativity”. On word level defining topic provide two benefits:
1. We can reduce content spread of sentences from frequency.
2. We can represent in proportion how much topic relate to word.

V. RESULTS

1. Data set:

We choose the dataset provided by NLPS. We can obtain a number of research papers published in NLPS conference. A research paper consists of a title, an abstract and the main text. Other data such as figures and tables were not extracted from the PDF files. Each paper discusses a novel technique or improvement. In this analysis, we will focus on analyzing these papers with natural language processing methods. the file contains some metadata such as id's and filenames, it is necessary to remove all the columns that do not contain useful text information. By looking at the number of published papers per year.

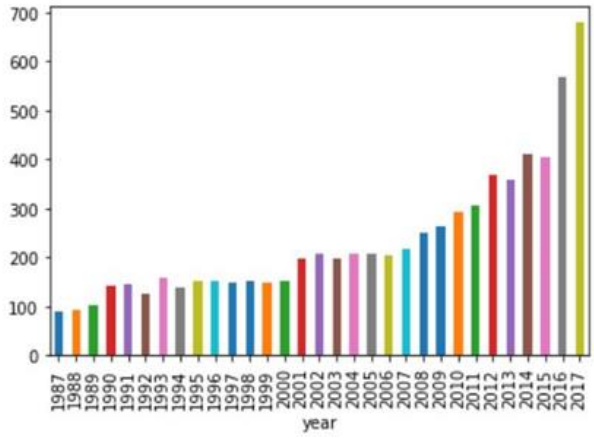


Fig. No. 2

2. Word Cloud:

We will perform some simple pre-processing on the titles in order to make them more amenable for analysis. We will use a regular expression to remove any punctuation in the title. Then we will perform lowercasing. In order to verify whether the pre-processing happened correctly, we can make a word cloud of the titles of the research papers. This will give us a visual representation of the most common words.

3. Analyzing trends with LDA:

The main text analysis method that we will use is LDA. LDA is perform topic detection on large datasets, determining what the main 'topics' are in a large unlabelled set of texts. We'll then plot the ten most common words based on the outcome of this operation. As a check, these words should also occur in the word cloud. Then will be analysed using LDA. It will give the top ten topics in dataset. This will helps to find trends in machine learning.

VI. CONCLUSION

In this project we have proposed a system which will discover future trends for NIPS using the title of a paper. To discover future trends, we have used bag of words for removing stop words and do tokenization. The tokenized data will then be given as input to the LDA model which is the best model for discovering the most frequently occurring words.

VII. REFERENCES

[1] Aiello, Luca & Petkos, Georgios & Martín Dancausa, Carlos & Corney, David & Papadopoulos, Symeon & Skraba, Ryan & Goker, Ayse & Kompatsiaris, Ioannis & Jaimes, Alejandro. (2013). Sensing Trending Topics in Twitter. IEEE Transactions on Multimedia. 15. 1-1. 10.1109/TMM.2013.2265080.

[2] Cho, Yung-Jan & Fu, Pei-Wen & Wu, Chi-Cheng. (2017). Popular Research Topics in Marketing Journals, 1995–2014. Journal of Interactive Marketing. 40. 10.1016/j.intmar.2017.06.003.

[3] Nagar, Anurag & Hahsler, Michael. (2015). Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams.

[4] Souili, Achille & Cavallucci, Denis & Rousselot, Francois. (2015). Natural Language Processing (NLP) – A Solution for Knowledge Extraction from Patent Unstructured Data. Procedia Engineering. 131. 635-643. 10.1016/j.proeng.2015.12.457.

[5] Hurtado, Jose & Agarwal, Ankur & Zhu, Xingquan. (2016). Topic discovery and future trend forecasting for texts. Journal of Big Data. 3. 10.1186/s40537-016-0039-2.