EMAILS CLUSTERING AND CLASSIFICATION

Parakh Srivastava
srivastavaparakh2017@gmail.com
Machine Learning Intern
AITS, https://ai-techsystems.com
India.

ABSTRACT
Interfacing through the continuously rising amounts of data in technical, medical, scientific, engineering, industrial and monetary fields and their renovation to logical form for the human user is one of the main requirements. To quickly discover and analyze complex patterns and requirements, we need the efficient techniques and need to learn from new data will be necessary for information-intensive applications. One of the solutions for this is that classification and clustering of largely available data. To partially fulfill the industry requirement, in this paper I proposed a two-level approach for clustering large data set of emails and then classifying mails into positive, negative or neutral. In this paper, a novel approach to clustering of the emails with PLSA and classification with different algorithms, are considered. In particular, the use of Topic Modelling and POS tagging are investigated.

Keywords
NLP, POS, PLSA, Sentiwordnet, Clustering, Classification.

1. INTRODUCTION
Over the previous decade, academic and commercialized databases have been extending at exceptional rates. Capture advanced perception from such databases is hard, expansive and time-consuming if done manually. It is hopeless when data exceeds definite limits of size and complexity. For this reason, during the previous years the automated analysis and visualization of huge multi-dimensional datasets has been the center of attention on scientific research. The fundamental aim is to observe rules and relationships in the data, thereby gaining attain to invisible and potentially valuable knowledge. There are number of applications related to different fields in the computer science and IT industry. One of the most valuable and trending is, Text Processing also known as Natural Language Processing. NLP have many use cases in the market, such as, Speech Recognition, Language Translation, identifying various languages and voice translation, and many more.
We will be using NLP for topic modelling on the email dataset, which is an unlabeled data, and further classification algorithm will be used for checking the accuracy of our labelling of the data.

2. ACRONYMS
NLP- Natural Language Processing
PLSA- Probabilistic Latent Semantic Analysis
POS- Part of Speech
PLSI- Probabilistic Latent Semantic Indexing
IT- Information Technology
BSD- Berkeley Software Distribution

3. RELATED WORK
Data analysis underlies many computing applications, either in a design phase or as part of their on-line operations. Data analysis procedures can be dichotomized as either exploratory or confirmatory, based on the availability of appropriate models for the data source, but a key element in both types of procedures is the grouping, or classification of measurements based on either (i) goodness-of-fit to a postulated model, or
(ii) natural groupings (clustering) revealed through analysis. Cluster analysis is the organization of a collection of patterns into clusters based on similarity.

3.1 Clustering
Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data. Progressively robust computer hardware has made it feasible to examine and probe databases whose complexity, dimensionality and amount of data contravene the limits in which manual analysis is feasible. The motive is to divulge patterns, relationships or regularities that grant us to gain new knowledge and intuition on the data.
Clustering is always done on an unsupervised learning algorithm, that is, it does not have labels for the data. There are number of techniques/algorithms for doing clustering. Some of them can be named as K-means clustering, Hierarchical Clustering, Topic Modelling, etc.

3.2 NLP- Natural Language Processing

A subfield of Computer Science, concerned with the interactions between computers and human(natural) languages, in particular how to program computers to process and analyze large amounts of Natural Language data.

POS Tagging- Given a sentence, determine the part of speech (**POS**) for each word. Many words, especially common ones, can serve as multiple part of speech. For example, "book" can be a noun ("the book on the table"); Some languages have more such ambiguity than others. Languages with little inflectional morphology, as English, are particularly prone to ambiguity.

Stemming- The process of reducing inflected (or sometimes derived) words to their root form. (e.g. "close" will be the root for "closed", "closing", "close").

Word Segmentation- Separate a chunk of continuous text into separate words. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language. Sometimes this process is also used in cases like Bag of Words (BOW) creation in data mining.

3.3 PLSA- Topic Modelling

Probabilistic latent semantic analysis (**PLSA**), also known as **probabilistic latent semantic indexing** (**PLSI**, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis, from which PLSA evolved.

Compared to standard LSA which stems from linear algebra and downsizes the occurrence tables (usually via a singular value decomposition), probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model.

Model- Considering observations in the form of co-occurrences of words and documents, PLSA models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

with 'c' being the words' topic. Note that the number of topics is a hyperparameter that must be chosen in advance and is not estimated from the data. The first formulation is the *symmetric* formulation, here and are both generated from the latent class in similar ways (using the conditional probabilities P(d|c) and P(w|c)) whereas the second formulation is the *asymmetric* formulation, where, for each document, a latent class is chosen conditionally to the document according to , and a word is then generated from that class according to . Although we have used words and documents in this example, the co-occurrence of any couple of discrete variables may be modelled in exactly the same way. So, the number of parameters is equal to . The number of parameters grows linearly with the number of documents. In addition, although PLSA is a generative model of the documents in the collection it is estimated on, it is not a generative model of new documents. Their parameters are learned using the EM algorithm.

3.4 XGBOOST- Classifier

XGBoost is an open source software library which provides a gradient boosting framework for C++, Java, Python, R, and Julia It works on Linux, Windows and macOS. From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library". It runs on a single machine, as well as the distributed processing frameworks Apache Hadoop, Apache Spark, and Apache Flink. It has gained much popularity and attention recently as the algorithm of choice for many winning teams of machine learning competitions.

3.5 WordNet

WordNet is a lexical database for the English Language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser, its primary use is in automatic text analysis and artificial intelligence applications.
The database and software tools have been released under a BSD style license and are freely available for download from the WordNet website. Both the lexicographic data (*lexicographer files*) and the compiler (called *grind*) for producing the distributed database are available.

APPROACH

Text data is difficult to interpret without applying NLP to it. Thus, started with NLP for getting the root words out of the mails. Then, applied Sentiwordnet algorithm for getting the sentiments from the extracted root words.

There are number of algorithms to deal with Topic modelling, like- PLSA, LDA, LSA, etc. Here, I have used PLSA algorithm and clustered the mails into three topics. Names of the topics can be self-assigned by looking deeper into the mails' root words, but this task was out of the context of the project.

Then we classified the mails into positive, negative or neutral, with the help of the extracted words. There, came the need of a classification algorithm to train and test the model and give out the accuracy of the model.

Finally, we achieved the accuracy of 75-80%, which can be further increased if take large number of samples to train the model, and then test it.

CODE ANALYSIS

All the code has been done on the Python Script (.py). There are number of libraries used in the project, namely- pandas, numpy, nltk, etc.

Code execution time for 1000 mails (sample), is around 2-3 minutes (on 4GB RAM laptop). There are number of tasks, that are repeated for executing other tasks, which was needed at the cost of the project.

MODEL EVALUATION

There are number of model evaluation techniques that can be used for evaluation our XGBClassifier model, and tell us how well we are able to predict the nature of the particular mail.

Classification Report

This evaluation metric gives us the detail about Precision, Recall, F1-Score and Support.

The knowledge about these four are sufficient for further analysis, whether we need it or not.

Code: classification_report (true, predicted)

	precision	recall	f1-score	support
0	0.74	0.74	0.74	34
1	0.72	0.69	0.71	26
2	0.83	0.85	0.84	40
avg / total	0.77	0.77	0.77	100

Accuracy Score

Accuracy Score will give the overall accuracy of the model, rather than accuracy for each class of the Output variable.

```
Accuracy Score: 0.77
```

Confusion Matrix

Confusion matrix is a n*n matrix, where n is the number of classes in the output variable.

Diagonal from the left corner of the matrix will tell about the correctly classified data counts.

```
[[25  3  6]
 [ 7 18  1]
 [ 2  4 34]]
```

Data Preparation

After reading the data, our task is to make the mails more suitable for the upcoming algorithms. Rooted words are extracted from the mails, by using NLP.

Stemming, Tokenization, Bagging the words (bag of words model), are some of the Data Preparation steps that are taken to make the data fit for the algorithms.

Summary

Good accuracy has been achieved, which can be further uplifted if the number of mails taken as sample increases.

Further, different models can be tried out for making the accuracy more effective in the particular case.

4. RESULTS

We have clustered 4000 words in each topic and the fourth topic is a random one, which contributes for all the remaining topics. We can cluster a larger or a smaller number of words too, but that depends on a particular choice.

There are 1000 emails taken, after NLP, we obtained root words from the mail body. Thereafter, we applied PLSA and got our three topics' words, which we encoded as numerals.

Then, we made a group of words, tagged as Positives and Negatives, and comparing the words with emails' words, we classified them as Positives, Negatives and Neutrals.

Then, we applied XGBoostClassifier for getting the better accuracy for our model, after splitting the data into training and testing sets.

At last, we got the accuracy of 75-80%.

0	Class	Category
wall str…	0	Positive
wall str…	0	Positive
houston …	1	Positive
congratu…	0	Positive
friday b…	2	Positive
follow s…	0	Positive
forward …	2	Neutral
legal as…	0	Neutral
posit re…	0	Negitive
dear hea…	2	Negitive

5. CONCLUSION

For text data topic modelling and classification, PLSA and XGBoostClassifier are one of the best machine learning techniques. In this proposed work we implement two level architecture for clustering and classification. First, the collected data, is further sampled to do the task. Then, we applied Clustering for topic modelling, POS tagging. Number of clusters we made is 3. Then, classification has been done on the self-labelled data, with the accuracy of between 75-80%. Further, if we take a greater number of mails for training the machine, it will give much better results.

6. REFERENCES

[1] Udemy Course- https://www.udemy.com/machinelearning/

[2] Coursera Course- https://www.coursera.org/specializations/data-mining

[3] Kaggle Kernels- https://www.kaggle.com/kernels