Tech Review

Stanford Topic Modeling Toolbox

Introduction

The focus of this tech review paper is to overview the "Stanford Topic Modeling Toolbox"1 and its applications as a modeling tool for the social scientists and other personals that want to use for their analysis. The software has implemented Labeled LDA.

Stanford Topic Modeling Toolbox

The application was developed by the Stanford Natural Language Processing Group 2 and became available for the first time in September 2009. Daniel Ramage and Evan Rosen are the responsible for the application development.

The main propose of the application is to give to social scientists and any person that want to incorporate the tool in their research to create datasets with the following features described by the group: "*Import and manipulate text from cells in Excel and other spreadsheets. Train topic models (LDA, labeled LDA and PLDA) to create summaries of the text. Select parameters (such as the number of topics) via a data-driven process. Generate rich Excel-compatible outputs for tracking word usage across topics, time, and other groupings of data.*" 3.

Requirements and Installation

To run the application, it is necessary to have at least java 6SE, and download the tmt-0.4.0.jar from the project website 4. The jar file is executable with double click. It also can be run by command line.

Dataset and Tokenizing, Data Model

The dataset come from CSV or TSV files exported by Microsoft excel, it is necessary to upload the files at the Topic Modeling Toolbox. After the file is loaded it has to walkthrough a process with a lot of stages and transformation.

The tokenizing process means that the terms in the dataset will be transform in topic model so it can be analyzed. During the process, if it uses SimpleEnglishTokenizer(), it will remove

punctuation from the end the words. Another important feature is CaseFolder that lowercase the words, so they are in the same standard. Anything that is not number or not characters are removed by WordsAndNumbersOnlyFilter. The MinimumLengthFIlter will remove any word less than 3 character.

In resume the methods that can be uses to tokenizing the data are: SimpleEnglishTokenizer, CaseFolder, WordsAndNumbersOnlyFilter, MinimumLengthFilter.

Other methods that can be implemented is the *TermMinimumDocumentCountFilter(n),* where n is the number of the document, and the method removes terms in less than "n" documents. TermDynamicStopListFilter(K), where k is the k most common term in the vocabulary, so this method will cut those words. Those methods must be run after the document is tokenizing and it is a great feature of the application, because this can be used to find the similarities between the documents.

Another important feature is the DocumentMinimumLengthFilter(w) to remove document with length less w, where w is the length of the document, this is important to remove the empty documents out of the data set.

The next step is to create a training topic model, for that it will be necessary to set the parameters for the LDA model using LDAModelParams(numTopic= j, dataset = dataset), if the user wants to smooth the topic terms it has the possibility to use Dirichlet adding LDAModelParams(numTopic= j, dataset = dataset, topicSmoothing=SymmetricDirichletParams(M)) where M is the Dirichlet parameter.

After that run TrainCVB0LDA(params, dataset, output=modelPath, maxIterations=1000) to train the model to fit the document.

Other type of models can be used: Labeled LDA model LabeledLDADataset(text, labels) and PLDA model.

Conclusion

The "Stanford Topic Modeling Toolbox"1 is a application the reach what was proposes for with the LDA, Labeled LDA and PLDA models implemented at the tool. It has the versability of install as standalone application, run by command like (which make possible to interact with other

program), and can be loaded on excel spreadsheet that is very used in many organizations and institutions around the globe.