aiaioo labs

TEXT ANALYTICS

THE TARGET AUDIENCE FOR THE COURSE:

- Engineers and analysts who use text analytics tools and want to get a deeper understanding of how to solve text analytics problems using primitives available in their tools
- Data Scientists interested in developing text analytics skills
- Professionals who develop software for:
  - ✓ social media analytics
  - ✓ unstructured big data analytics
  - ✓ text analytics
- Managers and Leads managing Text Mining Projects

The Wikipedia defines *natural language processing* as the field of computer science, artificial intelligence and computation linguistics that deals with the processing of human languages.

Text analytics is the process of deriving information from text. So, *text analytics* is a kind of data analysis that is performed using *natural language processing* techniques.

For many text analytics tasks, statistical or machine-learning-based natural language processing techniques as opposed to rule-based natural language processing techniques provide the best results.

In this workshop, we examine a number of text analytics tasks, and by working through text analytics problems, we learn the statistical natural language processing techniques and machine learning tools that might be applied to solving the same.

We also learn to measure the quality of a solution and to compare models, so that we can select the most suitable models and assemblies using hard quantitative evidence.

The workshop is suitable for experienced program and project managers, solution architects, data professionals and data analysts who need to work with text. It is also suited for programmers who wish to use open source tools for natural language processing.

For those aspiring to develop natural language processing algorithms, this workshop is a pre-requisite for an advanced course in machine learning and statistical natural language processing.

GOALS

Expected takeaways from the course

The course is designed to equip attendees with:

a) The ability to model text analytics problems as machine learning problems, and choose the right algorithms and tools to apply to solve them.

b) The ability to use open source NLP and ML toolkits to build solutions

c) Knowledge of the properties of different machine learning and text analysis algorithms.

d) Knowledge of resources and metrics that can be used in evaluating ML tools.

e) A jumping off point to several relevant key research topics in NLP and research papers related to these topics to serve as a launching pad for further reading/research.

TOPICS

1. INTRODUCTION TO STATISTICAL METHODS FOR TEXT ANALYTICS

Learn why statistics can be of use in natural language processing. What do you mean by statistical methods and why and when would you use them to analyze text? Learn some tools of probability theory and see how they can be of use for text analytics.

You will also learn how to reduce a text analytics problem to a machine learning one. You'll also get your first introduction to features and begin to understand how machine learning and analytics are related:

This section introduces you to the following text analytics problems and shows you how machine learning algorithms can be used to solve them. You'll learn what these terms mean and what approaches are typically applied to them:

a) Topic classification
b) Sentence segmentation
c) Tokenization
d) Part-of-speech tagging
e) Named entity recognition
f) Relation extraction

Additional case illustrations (all the above are already case illustrations).

- Function word identification
- Spell checking

Laboratory:

- Use open source tools to perform topic classification

2. BASICS OF TEXT CLASSIFICATION

Learn how a machine learning classification algorithm is different from a rule-based algorithm for text classification, and learn to measure the effectiveness of a classifier. Learn to break up your data into training, development and test sets and train and evaluate machine learning algorithms. Learn to gauge the reliability of a measurement using n-fold cross validation:

a) Rule-based classification.
b) Classification using Machine Learning Tools.
c) How a Naïve Bayesian classifier works.
d) Application to text categorization
e) How to measure how well a classifier works.

Case illustrations (using classification to solve problems in text analytics):

- Document classification
- Language identification
- Sentence segmentation
- Sentiment analysis

Laboratory:

- Learn to build classification models
- Build models for language identification
- Build models for sentiment analysis
- Measure the performance of various models

3. MULTICLASS AND MULTILABEL CLASSIFICATION

Learn the difference between multiclass and multi-label classification. Learn to identify problems where you need to use multi-label classification. Learn to measure the performance of a multi-label classifier. Learn different classifier configurations that can be used to solve the multi-label classification problem.

Case illustrations (using classification to solve multi-label problems in text analytics):

- Document classification

Laboratory:

- Learn to build multi-label classification models

4. INTRODUCTION TO ADVANCED TRAINING TECHNIQUES

Learn to train a classifier using partially labelled data.

Creating data sets is expensive, so it helps if you can use unlabelled data to improve the performance of a classifier. Learn to train a classification algorithm with very little labeled textual data. Learn to improve a classifier or change its characteristics using unlabelled data. Learn about a very useful procedure called expectation maximization.

Case illustrations (training a classifier for the topic classification of documents using partly labelled data):

- Document classification

Laboratory:

- Use open source tools to train a topic classifier using unlabelled data
- Learn to build a subcategory remover

5. CLUSTERING

Learn about various clustering algorithms and learn to use K-means clustering on text. K-means is a very well-known clustering algorithm. How to use it with text is less clear.

a) Agglomerative Clustering.
b) K-Means Clustering.
c) Soft K-Means Clustering.
d) Distance metrics for text
e) Distance metrics for probability distributions
f) Taking expectation maximization to the extreme

Case illustrations:

- Document clustering by topic

Laboratory

- Cluster a set of documents by topic using the K-means algorithm, and learn a nifty trick to end your day.

6. MAXENT CLASSIFIER

Learn about when you might want to use a maximum entropy classifier for text classification.

a) How a maximum entropy classifier works.
b) Under what conditions a maximum entropy classifier is worth the greater training cost.
c) When not to use a maximum entropy classifier.

Case illustrations:

- Document clustering by topic

Laboratory

- Classify documents using a maximum entropy classifier

THIS WORKSHOP IS OFFERED IN PARTNERSHIP WITH AIAIOO LABS:

Instructor: Cohan Sujay Carlos, CEO of Aiaioo Labs

Cohan is the CEO of Aiaioo Labs and comes with 15 years of experience in enterprise software, 8 of them working in statistical natural language processing. Cohan and his team have published work in natural language processing conferences including IJCNLP.

Aiaioo Labs is a research lab for natural language processing and machine learning.

Venue, Dates, Fees, Registration:

June 13th, Sat, from 9:30 Am to 5:30 PM

In Bangalore (Bangalore South, Bannerghatta Road, Executive Training Venue)

Venue and Directions will be communicated to confirmed registrants

Program Info: http://compegence.com/workshops/open/text-analytics/

Registration Link: http://compegence.com/register/tm201506.html

For additional Information, please contact:

workshop@compegence.com / +91-99805-40426