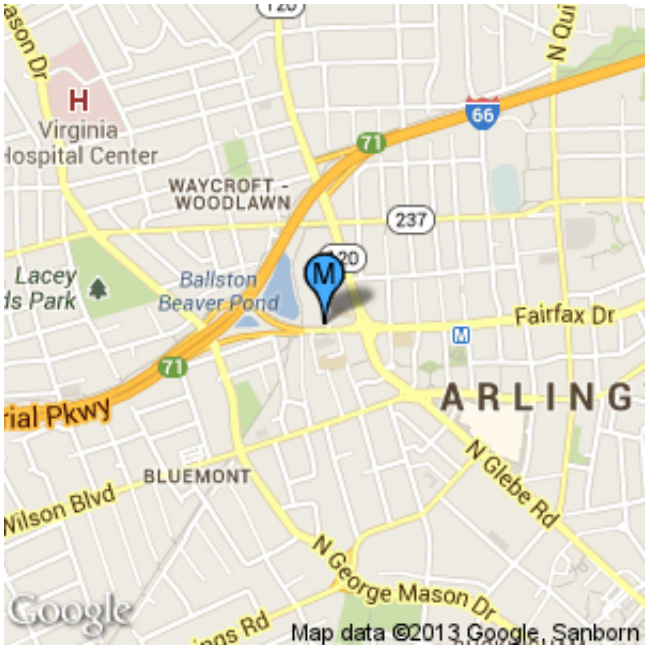
Natural Language Processing Basics with NLTK Workshop

Saturday, July 27, 2013 9:00 AM
Data Community DC

Name	Quantity	Price (per person)	Total	Conf. #
Mark Silverberg	1	$150.00	$150.00	P20735982

Metro Offices - Ballston Office Center

4601 N Fairfax Drive, Arlington, VA 22203



Meetup agenda:

A vast amount of the world's data (and your own) is text and the key to unlocking its value is in a series of standard natural language processing steps that we must make on raw text to ensure that it is understandable to machines. This workshop will teach you how to transform strings to machine usable data and discuss how to apply each step to your application. We'll be using Python's NLTK - a professional grade Natural Language Toolkit that is free and open source.

The price per attendee is $150.

What To Bring:

- Your laptop

- Printed copy of your ticket

Prior to the Event:

Please take our survey so that we can get a feel for the make up of the class. Although, I'd ask only folks who are registered to take the survey-- this will allow us to customize the class to the group.

NLP Workshop Questionnaire

Abstract:

Many of the largest and most difficult to process data sets that we encounter tend not to be from well structured logs

or databases, but rather unstructured bodies of text. In recent years, Natural Language Processing techniques have accelerated our ability to stochastically mine data from unstructured text but require large training data sets themselves to produce meaningful results.

The most popular open source tool for NLP is NLTK and it embraces these techniques, while providing a powerful interface in Python for us to quickly add NLP to our applications.

Outline:

1. Organizing large bodies of text (corpora)
2. Task: tokenization and segmentation;Motivation: cross-document language statistics
3. Task: tagging and stemming; Motivation: information extraction
4. Task: parsing for treebanks; Motivation: discovering crucial concepts

We will provide a virtual machine (vmware) containing an NLP development environment with all required software preinstalled and preconfigured for this workshop.

If you have text that you would like to process, please bring your data along. If not, we will be using corpora from Project Gutenberg.

Instructor

Benjamin Bengfort, Data Science Consultant, Full Stack Data Science

Benjamin Bengfort is a Data Science consultant at Full Stack Data Science, and has used Machine Learning and Natural Language Processing techniques to determine textual complexity in large literary corpora. He is a PhD candidate in Computer

Science, with a focus on NLP, at the University of Maryland, Baltimore County, and has a MS in Computer Science from North Dakota State University.

Organizer's refund policy:

You will receive a refund if:

- the event is cancelled
- you cancel at least 7 days before the event

Hosted by

Tony Ojeda

Benjamin Bengfort