# Overview

In this project, you will learn how to find aspects and opinions in product reviews. You will also learn how to perform basic text pre­processing before you can do opinion mining.

We are providing you a digital camera dataset in two forms: tagged and untagged (see readme). You need to use the [Opinion Finder](http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/) tool to do pre­processing steps (you also try out the [GATE](https://gate.ac.uk/download/) [tool](https://gate.ac.uk/download/)). Reference paper ["Mining and summarizing customer reviews"](http://www.cs.uic.edu/%7Eliub/publications/kdd04-revSummary.pdf).

# Getting Started

The coding is based on Java & python 2.7. Download and install python 2.7: sudo apt­get install python2.7. For windows, you can get it from [here](https://www.python.org/download/releases/2.7/).

# Exercises

**Exercise 1** Take the untagged dataset and perform POS tagging using the Opinion Finder tool.

**Exercise 2** From the POS tagged dataset extract all the Nouns. Call this the list of candidate aspects.

**Exercise 3** Select all the frequent aspects using Association Rule Mining algorithm with 1% support (see Resources section). (Submit the top­10 most frequent aspects in decreasing order of support.)

**Exercise 4** Use the identified frequent aspects in Exercise 3 to find candidate opinion words. These are the adjectives or adverbs which are near the frequent aspects identified in Exercise 3 (within five words on both sides of a frequent aspect). (Submit the list of top­10 most frequently used opinion words.)

**Exercises 5** Compute precision, recall and F1 score for the aspects identified obtained step in Exercise 3. Use the tagged dataset to know the ground truth aspects.

**Exercise 6** Compute the number of positive and negative opinion words you obtained in Exercise 4. You need to use the provided [list](http://www.cs.uic.edu/%7Eliub/FBS/opinion-lexicon-English.rar) of standard opinion words in English.

# Resources

Python has good support for basic text processing and mining. You may find interactive IDE such as [Anaconda 2.4](https://www.continuum.io/blog/developer-blog/anaconda-24-release) (see [tutorial](http://conda.pydata.org/docs/py2or3.html)) quite useful (download for [Linux](http://repo.continuum.io/archive/Anaconda2-4.1.1-Linux-x86_64.sh)*/* [Windows](http://repo.continuum.io/archive/Anaconda2-4.1.1-Windows-x86_64.exe)).

For basic text pre­processing using python see [here](https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words). Some more info about text mining see [here](http://text-analytics101.rxnlp.com/p/compiled-tutorials.html).

For association rule mining using FP­Growth algorithm

[https://github.com/enaeseth/python­fp­growth](https://github.com/enaeseth/python-fp-growth) (recommended) <http://www.borgelt.net/fpgrowth.html>

Tutotial: <https://www.youtube.com/watch?v=vPcJEFFWN_k>