A data-driven approach to predicting Chemical Reaction Kinetics

Maneet Goyal, Keren Zhang

Feb 16 2018

Background

The products of chemical reactions are dependent on various factors, such as temperature, pressure and molecular properties, which underscore the complexity of these processes. Meanwhile, it is imperative for us to figure out the impact of these factors, and to be able to predict the results of these reactions based on varying reaction conditions. However, oftentimes it is extremely difficult to elucidate the underlying mechanisms and foretell what occurs in a reaction system. The tide of using data science to tackle such problems has risen recently, thanks to the advent of many large on-line databases and tools that facilitate a wide scale acquisition of relevant data in a time efficient manner.

Goals

At the outset, the objectives of our project are to:

1. Successfully retrieve reaction and molecular data via PubChem and ChemSpider web APIs, web scraping using Beautiful Soup, and NIST Chemical Kinetics database and then integrate them into feature vectors. Here, a reaction's feature vector is formed by appending the reactants' properties along with the normalized reaction temperature. Reactants feature vector will attempt to be a diagnosis and dissection of reactants into functional groups and chemical bonds present and quantify them into entries of vector. Here is a representation: [Number of C-C, C=C, C#C, -OH, -COOH, -CHO, (-Cl + -Br + -Fl), etc.] up to 20-30 elements depicting major functional groups and will be populated by parsing SMILE representation of the candidate reactants. Some binary element will also be there depicting, for instance, presence of radicals, ions, etc.
2. Use ensemble based methods (involving K-Nearest Neighbor and Random Forests) to develop a classification model for predicting Reaction Order and neural networks and linear/non-linear regression to develop a regression model for predicting Activation Energy.
3. And finally make an attempt to evaluate our results using chemical engineering knowledge. For instance, try to look for the factors that differentiate first order reactions from others and propose a chemical reasoning behind the same.

Definitions of Success

Achieving over 75% accuracy in reaction order, and a relative error of under 20% in activation energy while performing cross-validation. Goals 1 and 2 are critical to success. Goal 3 is secondary but quite important because it allows our chemical engineering knowledge to interpret the patterns of results.

Deliverables

The *3 models/ensembles* (Python functions, hyper-parameter JSON, if applicable) which predict the proposed parameters and a *comprehensive dataset* on gas phase reactions. Others include plots showing the accuracy of our models, an OOP based Python package for reproducing our data collection/processing results, and a schematic which presents the main work-flow of our complete analysis. Last but not least, we will perform an analysis of the results using in an attempt to explain what similarities are there between reactions exhibiting a certain reaction order and activation energy.

Challenges

Being able to predict the proposed parameters via a general and robust model may require years of experimental, theoretical or even computational (simulation based) research. Data analytics, on the other hand, provides us with tools to process gigantic datasets and bestows opportunities of using statistical methods to build surrogate models that help correlate the input descriptors/parameters and outputs, and recognize patterns in reactions. However, the scale of data we may encounter and the speed & variety of analysis required could be overwhelming for traditional data analysis tools like MS Excel. Moreover, traditional tools assume that the data is already available in a tabular format. Clearly, this is not the case here where we had to resort to using Python to scrape the required data off the HTML sources. Although Excel also allows scripting to process data programatically, but the off-the-shelf Python libraries make the development time shorter.