Emily Hillenbrand

Metis Data Science
Project 04 - Unsupervised Learning and NLP

How Can This Chemical Hurt Me?

04 November 2019

Overview

The International Union of Pure and Applied Chemistry (IUPAC) standardized nomenclature of molecules is the universally accepted system for naming chemicals. A molecule's name is built on its makeup of substituents. Using a database of Safety Data Sheets (SDS), I want to find a way to predict a chemical's hazards by finding relationships between those hazards and the substituents of the molecule.

Potential Applications

A. **Identify hazards associated with "found" chemicals:** When working in a chemistry lab, especially in a university, it sometimes happens that you find orphaned/legacy/unknown chemicals, or sometimes the bottle's label falls off or becomes corroded. There are steps you can take to identify the functional groups of its contents, but sometimes a chemical's name is too complicated or context-dependent to allow you to find an SDS. If you can use those functional groups to identify some likely hazards, you can decrease the risk of those hazards until you are able to find a more permanent solution.

B. **A more complete way to identify hazards of newly created chemicals:** In chemical synthesis, by its very nature, there is no SDS associated with the chemical you've just made. There isn't a clear process in place for labeling these chemicals properly. With a working predictive model based on a molecule's functional groups, chemists can

decrease some amount of risk by inferring their new molecule's hazards and labeling it properly .

Milestones

1. Scrape data from SDS database using CAS numbers from Kaggle dataset

   I want to use the Kaggle dataset so that I have a limited number of molecules to go through. There are hundreds of thousands of unique chemical names, but the dataset gives me a selection of about 67,000 unique numbers on which I can build my model.

2. Parse chemical names and associated hazards to create a single document for each molecule

   Create n-grams from molecule functional groups. I expect this to be the most complicated part.

3. Create MVP

   Do an analysis: dimensionality reduction or clustering

4. Improve and finalize model