May 24, 2017 Report

Thresholding

In the previous report, we mentioned that we tested the thresholded close-topic method on a couple of examples with a few different threshold values. This day, we tested with few extra examples and with the same thresholds: 0.75, 0.8, and 0.9. Almost always, there was minimal change compared to the normal un-thresholded method.

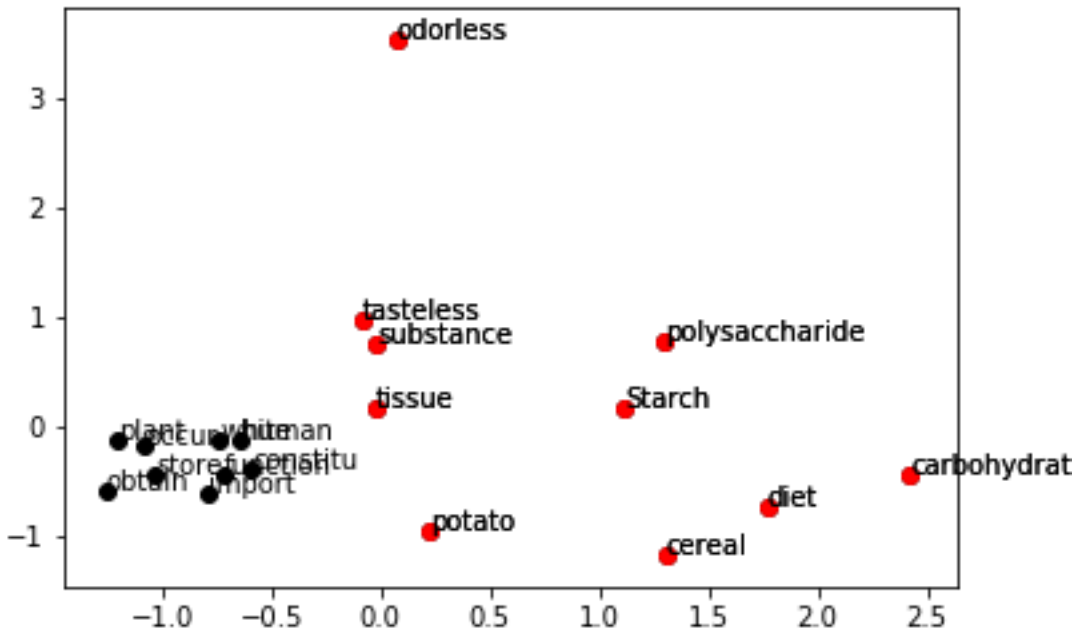
Visualization

We tried applying PCA to visualize the positions of the keywords compared to all the other words in a 2-d space.

Example

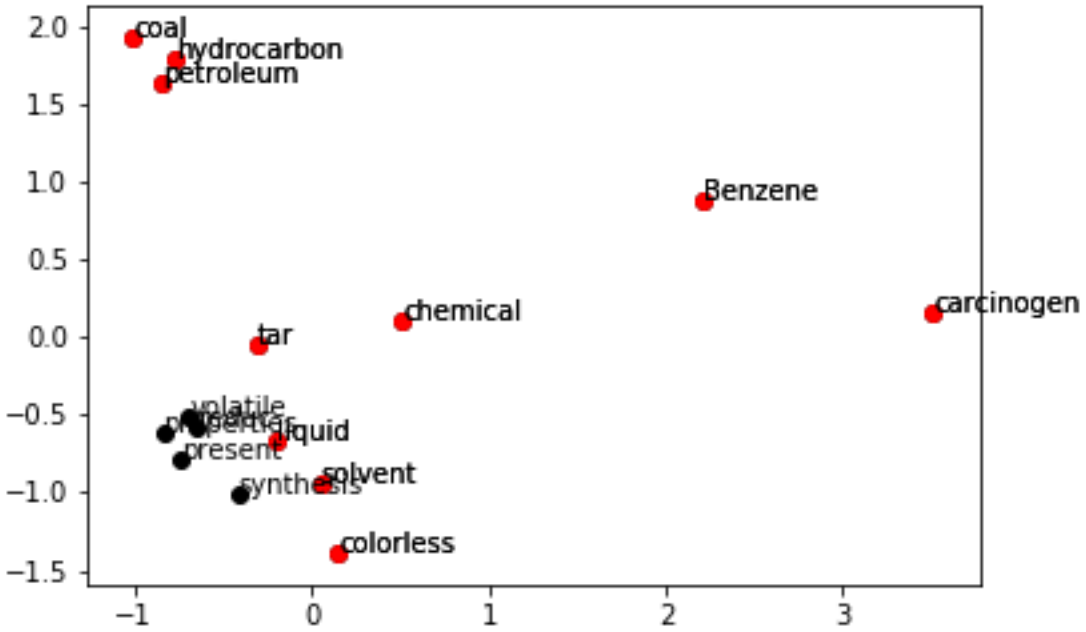
> Starch is an odorless tasteless white substance occurring widely in plant tissue and obtained chiefly from cereals and potatoes. It is a polysaccharide that functions as a carbohydrate store and is an important constituent of the human diet.

These are shown in the scatter plot below:



Another Example

> Benzene is a colorless volatile liquid hydrocarbon present in coal tar and petroleum, used in chemical synthesis. Its use as a solvent has been reduced because of its carcinogenic properties.



Data collection

We collected 15 random samples of data related to chemistry like atomic number, benzene, chlorine, ozone, quantum number etc. The above shown visualizations and examples, and this day's presentation contain all the examples taken from this data set.

Close-Topic Method

After testing on more and more documents, our intuition of why this is working the way it is improved a bit.

Our model works by choosing a word at random and making a chain that has the closest intra-distances between them. Therefore, if we ask the model to give three keywords, these three keywords will be not be the top three keywords in the top ten keyword lists. What we mean is that, the intra-word distance is independent on the number of keywords requested.

Results

In this day's presentation, we showed 10 slides worth examples comparing between normal close-topic clustering method and preprocessed (capitalization and plural forms removed) closed-topic method. After those, we also gave a few examples contrasting between this close-topic method to our previous clustering + PCA method.

Overall, we have seen a lot of decrease in false positives and an increase in number of quality (useful) keywords.

New Method

To improve our current method, we want to try some of the recent machine learning methods like deep networks for our task. Though we have an idea to create an autoencoder that can do PCA, but a non-linear PCA, we do not have an intuition on why or how it would or should work. Therefore, we currently mailed Siraj Raval (http://www.sirajraval.com) who is currently working with Udacity and has a passion for solving different problems using neural networks.

We also read some theory and watched a couple of lectures on Recurrent Neural Networks to see if we can make them work in this context.