**Group 8**

**Identify firms in the Oil sector and firms not in the Oil sector.**

There are 18 firms in the oil sector and 482 firms not in the oil sector.

**Are these tokens associated with the oil sector?**

The tokens are associated with the oil sector. However, the farther down on the list, the less frequent the bigrams are in the different firms. Therefore, some are not closely related to the oil industry and can be applied to other industries as well.

**Compute the average return for the oil portfolio and the peer group for the year 2014** For this one we were unsure whether to compute the average return for each peer firm or for the average of the five peer firms of an oil-firm, so we did both. Here is the average return for the latter:



**Evaluate the performance of your algorithm.**

RMSE for the bigrams 0.03892031. It represents the average distance between the predicted values, in this case the peer values, from the model and the actual values in the dataset, in this case the firm values. A lower value indicates a better fit, that this model has a good fit.

**Evaluate whether using uni-grams (versus bi-grams) performs better/worse?**

RMSE for the unigrams 0.04564829. The bigrams have a higher root mean squared error, meaning that it has a worse performance estimation than for the bigrams. However, it is still a relatively low value and can be a good fit.