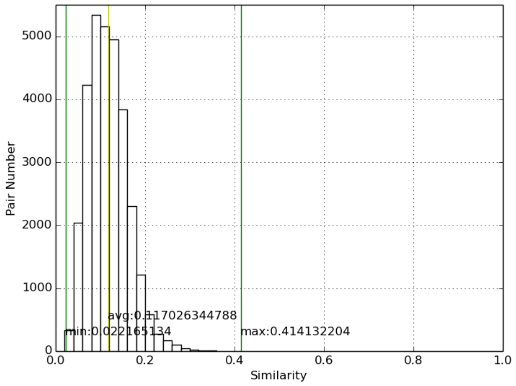
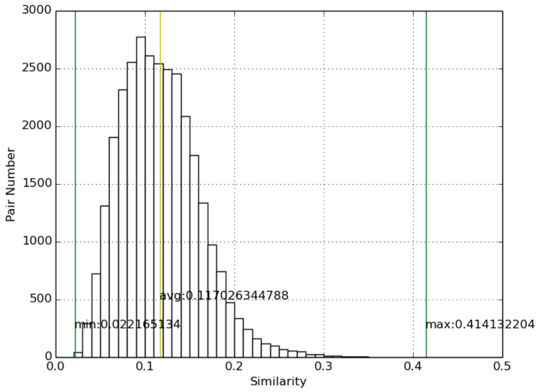
Weekly Report

For the sake of comparing this similarity measure to other common similarity metrics, it is useful to normalize the result to obtain a metric bounded in [0,1]. The normalized similarity measure is given by formula

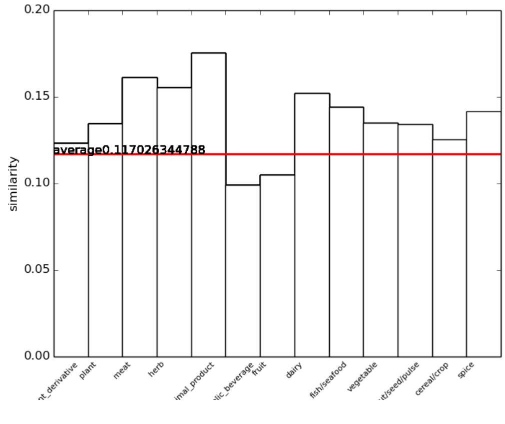
$$sim(i,j) = \frac{\sum_{I_{R_i} \in I} \sum_{I_{R_j} \in J} \frac{2 \cdot |I_{R_i} \cap I_{R_j}|}{|I_{R_i}| + |I_{R_j}|}}{|I| \cdot |J|} \tag{1}$$

If results are plotted, as we can see in the graph below, is very condensed within the range [0,0.5]. The next graph shows a plotting within range [0,0.5].





And the normalized average similarity in category are shown below:



On the other hand, a common way to compare the similarity among vectors is calculating distance metric. In word2vec, the similarity of the vectors are given by their cosine:

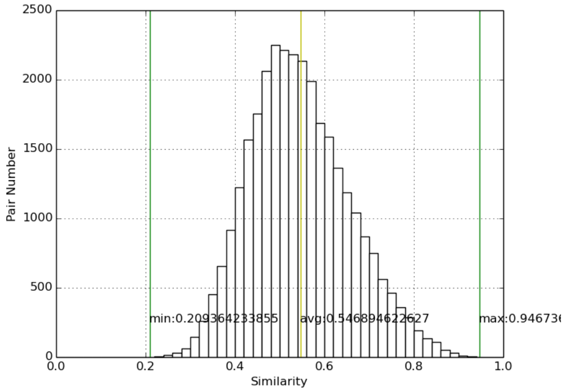
$$cosine\_similarity(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

Since in vector file, their are negative values, the cosine similarity is bounded between $[-1,1]$.

In order to obtain a similarity metric that is bounded between [0,1], we need to transform the cosine similarity to angular similarity by:

$$angular\_similarity(A,B) = 1 - \frac{\cos^{-1}(cosine\_similarity(A,B))}{\pi} \tag{3}$$

By plotting this similarity distribution, we get graph as below:



However, the distribution of angular similarity does not have a shape similar to the normalized similarity distribution that we talked above. To explore if these two similarity measures agree to each other, I try to sort both similarity matrix (the angular similarity matrix and the normalized similarity matrix), and compare if they have same ordering of most similar ingredients in each row. For example, considering ingredient 'beef', according to the normalized similarity, top 15 most similar ingredients are: celery_oil, tamarind, kidney_bean, tomato_juice, tomato, beef_broth, green_bell_pepper, oregano, onion, roasted_beef, garlic, black_bean, cayenne, bay, okra. And according to angular similarity, its top 15 most similar ingredients are: tomato_juice, beef_broth, onion, garlic, roasted_beef, pork' 'black_pepper, meat, bean, green_bell_pepper, bell_pepper, tomato' 'carrot, tamarind, vinegar. The number of common ones are calculated and averaged among all 248 ingredients. And it turns out that these two measures only partially (at best 35