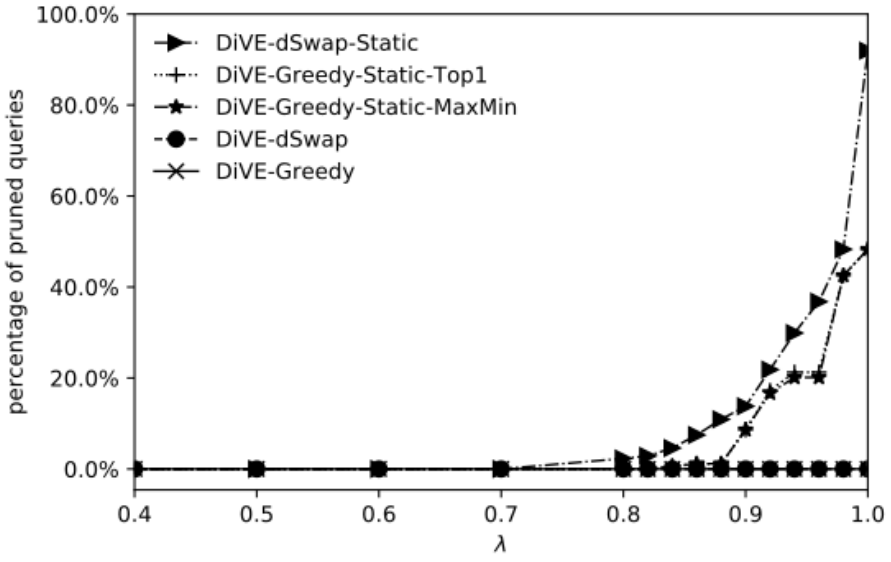
Greedy Static Max-Min vs Greedy Static Top-1

Greedy static Max-min

1. Get two most distant views as the initialization set S.
2. Execute query views of the initialization set and get the maximum importance score.
3. Use √2 as the maximum static bound, this maximum will not be changed until the end.
4. Use maximum importance score from the initialization set as the minimum bound.
5. Compute utility U of all views in X using √2 as the max bound and use actual diversity score.
6. Sort views in X based on the highest utility score, store in the list L.
7. Calculate Umax and Umin
8. Prune all views which have Umax < max(Umin)
9. Execute the first query view in L and calculate the real objective function F(S) as the *current_F(S)*.
10. If the importance score of executed query > current minimum bound, then update the minimum bound.
11. Always repeat step 7 and 8 after the minimum bound is updated.
12. Execute the next query view in L, calculate the real objective function F(S), if this F(S) higher than the current, then replace the current.
13. Repeat step 10 and 12 until there is no F(S) that higher than *current_F(S)*.
14. Go to the next iteration.

Greedy Static Top-1

1. Get two most distant views as the initialization set S.
2. Use √2 as the maximum static bound, this maximum will not be changed until the end.
3. Compute utility U of all views in X using √2 as the max bound and use actual diversity score.
4. Sort views in X based on the highest utility score, store in the list L.
5. Execute the first query view in L and calculate the real objective function F(S) as the *current_F(S)*.
6. If U of views in L < *current_F(S)*, those views will be pruned.
7. Execute the next query view, calculate the real objective function F(S), if this F(S) higher than the current, then replace the current.
8. Repeat step 6 and 7 till there is no F(S) that higher than *current_F(S)*.
9. Go to the next iteration.



The pruning performance of Greedy Max-min and Top-1 is very close. However, still swap has better performance compared to Greedy. The reason why we need swap is because Greedy starts with small number of views in the initialization. Only two views as the set S. Hence, while calculating the utility score of each views in X using static maximum bound and actual diversity score, there are a lot of views have same utility score. Due to of this, the chance of pruning is low.

However, Swap has bigger number of views in the initialization (e.g. five views). While calculating utility score of views in X, using static maximum bound and actual diversity score, the chance of views have same utility score is lower than Greedy.

Maximum bound of Euclidean distance between two probability distributions

Example maximum condition for two bins case:

$\sum a = \sum b = 1,\ a, b \geq 0$

$(\sum a)^2 + (\sum b)^2 \geq \sum a^2 + \sum b^2$

$(\sum a)^2 + (\sum b)^2 \geq \sum a^2 + \sum b^2 - \sum 2ab$

$(\sum a)^2 + (\sum b)^2 \geq \sum (a^2 + b^2 - 2ab)$

$(\sum a)^2 + (\sum b)^2 \geq \sum (a - b)^2$

$1 + 1 \geq \sum (a - b)^2$

$\sqrt{2} \geq \sqrt{\sum (a - b)^2}$

For the general case, Euclidean distance d is defined as following: $d = \sum (x - y)^2 = \sum x^2 + \sum y^2 - 2\sum xy$. Given that in probability vectors all values are nonnegative, d is max when the last term is zero, then $d = \sum x^2 + \sum y^2$.

All values are between 0 and 1 (sum up to 1), $\sum x = \sum y = 1$. In such a vector, its theoretical maximum is attained when all its entries are 0 except one which is 1, it is when $\sum x^2 = \sum x$ and $\sum y^2 = \sum y$. It also follows from the above description, that then $\sum xy$ can very easily happen to be zero (since in each vector there is just single nonzero element).

Maximum bound of Kullback-Leibler (KL) distance between two probability distributions

For distributions which do not have the same support, KL divergence is not bounded. Look at the definition: $KL(P||Q) = \int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$

If P and Q have not the same support, there exists some point x' where $p(x') \neq 0$ and $q(x') = 0$, making KL go to infinity. Even both distributions have the same support, when one distribution has a much fatter tail than the other. Then:

$$KL(P||Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$$

when

$$p(x) = \overbrace{\frac{1}{\pi}\frac{1}{1+x^2}}^{\text{Cauchy density}} \qquad q(x) = \overbrace{\frac{1}{\sqrt{2\pi}}\exp\{-x^2/2\}}^{\text{Normal density}}$$

then

$$KL(P||Q) = \int \frac{1}{\pi}\frac{1}{1+x^2} \log p(x)\,\mathrm{d}x + \int \frac{1}{\pi}\frac{1}{1+x^2}[\log(2\pi)/2 + x^2/2]\,\mathrm{d}x$$

and

$$\int \frac{1}{\pi}\frac{1}{1+x^2} x^2/2\,\mathrm{d}x = +\infty$$

In conclusion, Kullback-Leibler (KL) is not bounded. For instance, when I implement KL in my code. In some case while the bin does not has its pair in the reference (which means 0), the result are two posibilities: error devided by 0 or *log* 0 which is undefined.

Max-sum and Max-min diversification

Max-sum is bi-criteria objective function to maximize the sum of the relevance and dissimilarity of the selected set, which can be defined as follows:

$$F(S) = (1-\lambda) * I(S) + \lambda * f(S, D) \tag{1}$$

Where, $I(S) = \sum_{i=1}^{k} \frac{I(V_i)}{I_u}, V_i \in S$ and $f(S, D) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j>i}^{k} D(V_i, V_j), V_i, V_j \in S$

Meanwhile, Max-min diversification is the bi-criteria objective function that maximize the *minimum* relevance and dissimilarity of the selected set. Based on the work of Gollapudi (An axiometic approach for result diversificaiton), this objective function can be defined as follows:

$$F(S) = (1-\lambda) * \min_{u \in S} w(u) + \lambda * \min_{u,v \in S} d(u, v) \tag{2}$$

While Max-min diversification is to maximize the minimum of importance score, I am not sure this approach is relevant or not for our work.

The End.