For this assignment I did three things:

1) LDA on the 10 documents provided by Jevin

2) Calculate Culture Hole for 5 toy documents which was divided into 2 groups

3) LDA on the 5 toy documents

I will describe the process for each of them below.

**1) LDA on 10 documents**

I used the LDA and sklearn package. Topics as follow:

\*Topic 0

- aperture contact scrubbing secure middle

\*Topic 1

- rotated claim secure trunk relative

\*Topic 2

- base layer internal conducting extends

\*Topic 3

- claim flexible late rotated spacers

\*Topic 4

- rotated spraying mixed feet ones

**2) Culture Hole for 5 toy documents**

doc1: apples pears strawberries raspberries watermelon honeydew

doc2: cucumber tomato eggplant celery carrot apples

doc3: airplane is flying in the sky bird is flying in the sky cloud is in the sky

doc4: the sky is sometimes blue sometimes black sometimes burning

doc5: dog chases cat computer processes data sometimes

Doc1 and doc2 is group1 and doc3 – doc5 is group2.

I first calculated the Shannon Entropy for group1 and group2: 3.59, 5.49

Then the Q(pi || pj): 8.20, and Q(pj || pi): 12.36

Thus the Eij: 0.44 and Eji: 0.44

And then Cij = 0.56 and Cji = 0.55

**3) LDA on 5 toy documents**

\*Topic 0 - is computer watermelon honeydew cat

\*Topic 1 - sometimes strawberries burning flying carrot

From the results, I think Culture Hole makes more sense. Since the 2 groups do not have any overlapping words, the difficulty of group1 understanding group2 would be comparable to that for group2 to understand group1. This theory was confirmed by the similarity between Cij and Cji.

For LDA, if in a perfect world, I would expect the first topic has vegetable and fruit words, and the second topic has other things. But in this case these words were mixed in both topics, which is hard for me to understand and explain.