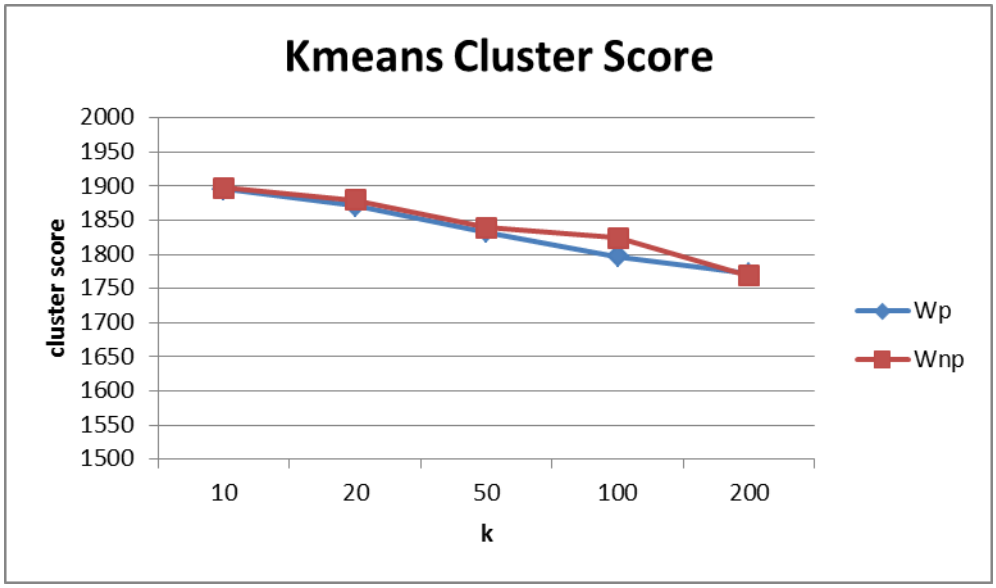
Xiaobo Zhang

4/14/2015

HW4 Report

1.

a) KMeans cluster score:



b) noticeable topic:

k-200 is best

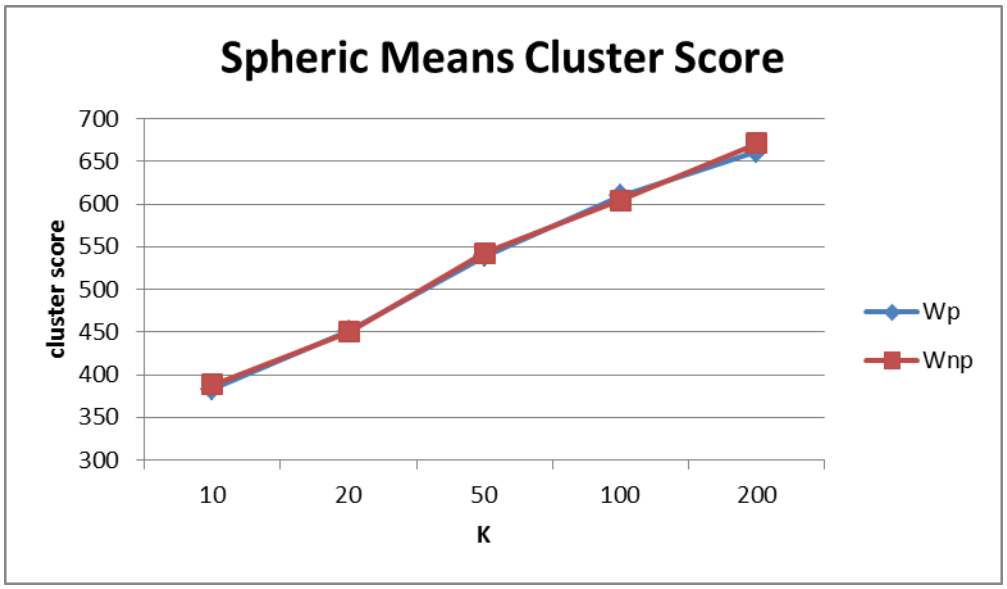
['meal', 'dinner']

['tell', 'change', 'charge', 'oil', 'doctor', 'surgery']

['amazing', 'definitely',' awesome'…,'Fun'…., 'wonderful', 'stars……'enjoy', 'real'……., 'cool' ……'fantastic', 'sweet',…..]

['nothing', 'money', 'later', 'finally', 'rude', …..'worst', 'oh', 'server'…..'maybe', 'ill',…. 'wrong'…….]

c) Spherical means cluster score:



k-200 is best

Noticeable topic:

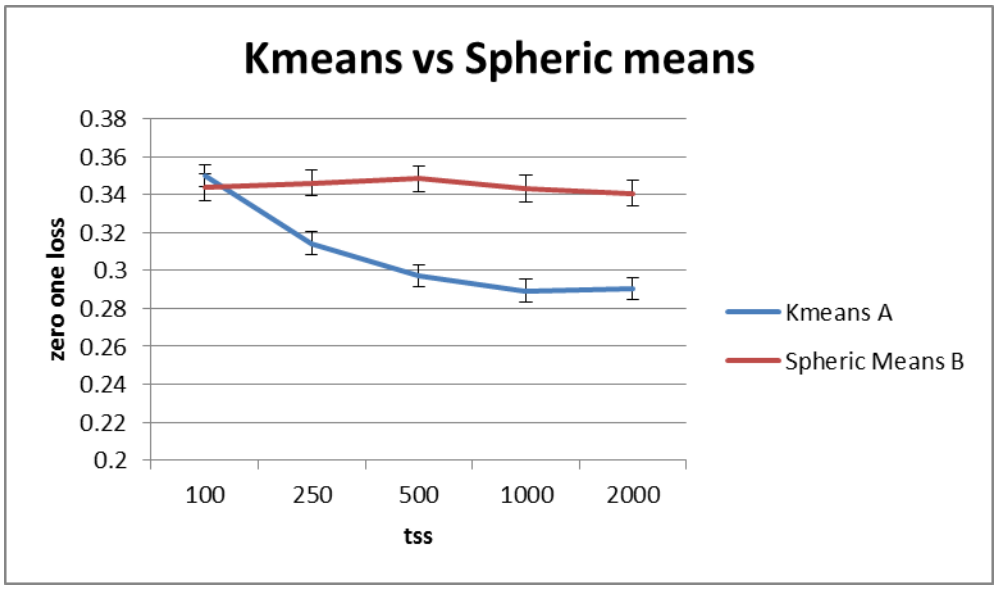
['yes', … 'chinese', 'simple', 'egg', 'rolls', …. 'sour', 'tastes', 'teriyaki', 'damn', 'spring', 'enchiladas',

'brown', 'kinda', 'daily', 'lease', 'happens', 'mall', 'tofu', 'crisp', 'generous', 'english', 'strip', 'heres', 'stupid', 'sliced', 'guacamole', 'chunks', 'starting', 'yuck', 'suck']
['wine', 'glass', 'bartender', 'list', 'bottle', 'plates', 'stated', 'split', 'bianco', 'prefer', 'slaw', 'cheeses', 'mixed', 'tap', 'classic', 'cocktail', 'cocktails', 'occasion', 'mustard', 'initial', 'apologize', 'consistent', 'insisted', 'extensive', 'sampler', 'sample', 'relaxed', 'includes']

d) The most noticeable difference for KMeans algorithm, the best k with lowest score, and for spherical means algorithm the best k with highest score.

2.

a)

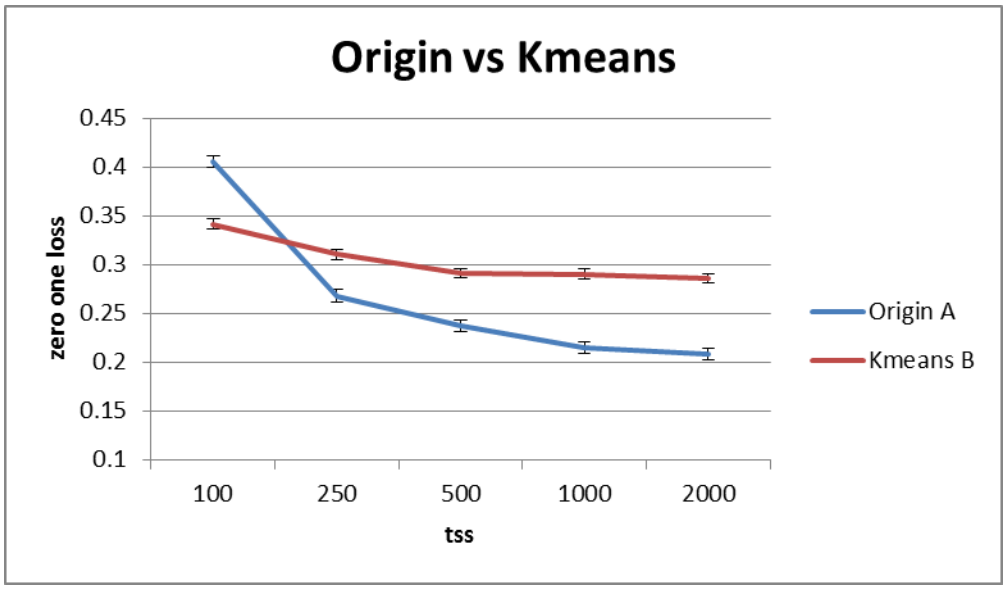


b) Hypothesis: Approach A (KMeans topic feature) will change cluster score faster than Approach B (Spherical means topic feature) when the tss size increase.

c) Perhaps the change of distance (greater than 0) will be larger than change of similarity (less equal to 1), so the score change will be more clear on KMeans topic feature.

3.

a)

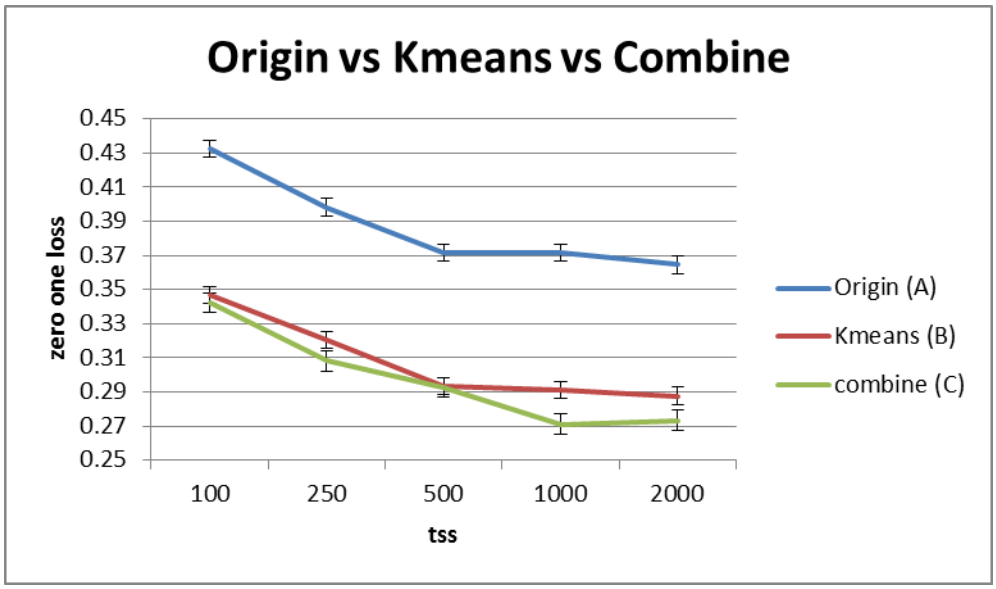


b) Hypothesis: Approach A (Origin Feature) has much lower 0/1 loss score than approach B (KMeans topic feature) when the tss size increase.

c) Perhaps the model size of Approach A (2000 single word attributes) will be larger than Approach B (100 topic feature), so the score of Approach A will be better than Approach B as the tss size increase.

4.

a)



b) Hypothesis: Approach B (KMeans topic feature) score is much smaller than Approach A (Origin feature) score.

Perhaps random choose 100 frequent word is not enough to represent the entire model, and perhaps KMeans topic feature include more words than origin NBC to represent the model will be more correctly.

c) Hypothesis: The score of the combined feature is dominant by KMeans topic feature.

The Approach KMeans topic feature has enough topic features to represent the model, so the combined model is dominant by topic features.