FinalReport

Baoxu(Dash) Shi

November 21, 2014

Evaluations

We evaluate the performance of HDTM on Wikipedia with the following metrics:

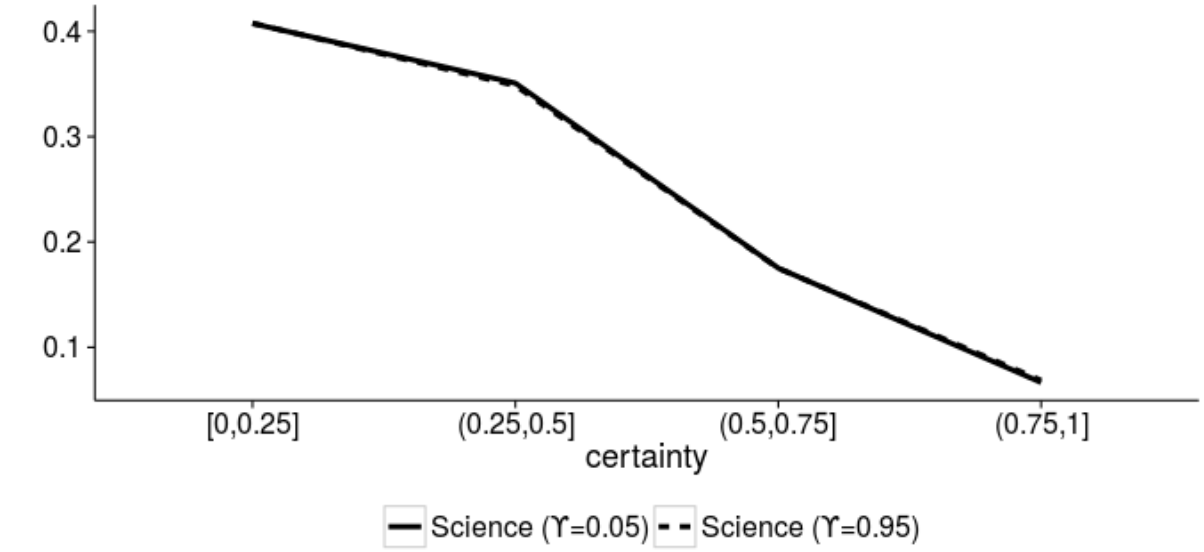
- Log likelihood (convential way to evaluate the effectiveness)
- Jaccard Coefficient (use category graph to evaluate the effectiveness)
- Certainty (The confidence of assigning parent to children based on HDTM)
- Height (The height of output hierarchy)
- ECDF (Empirical cumulative density function, alternative to height)
- DeltaCon (Graph similarity)

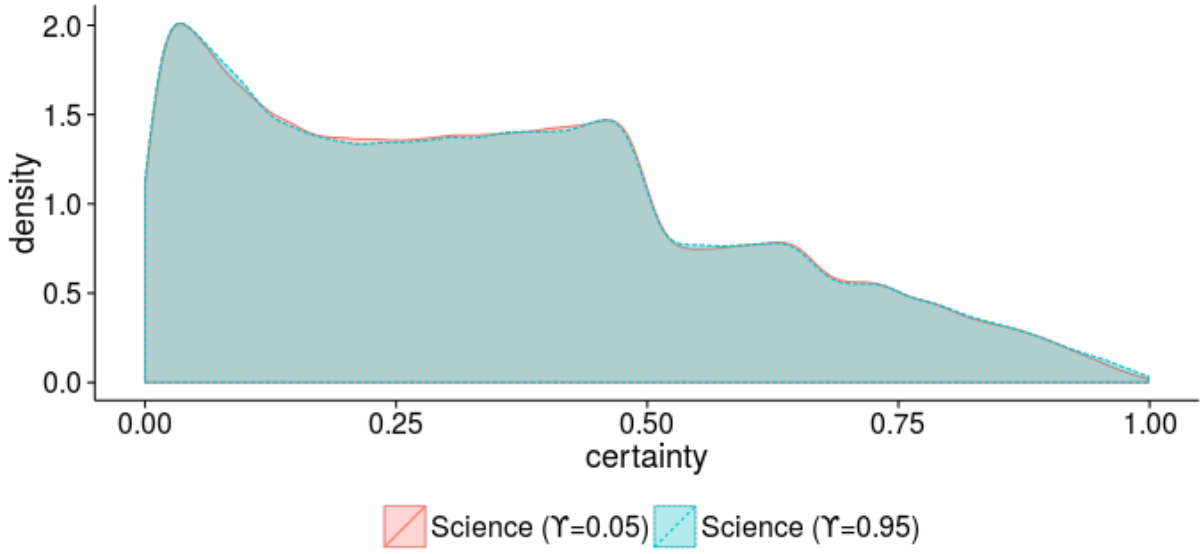
Log likelihood

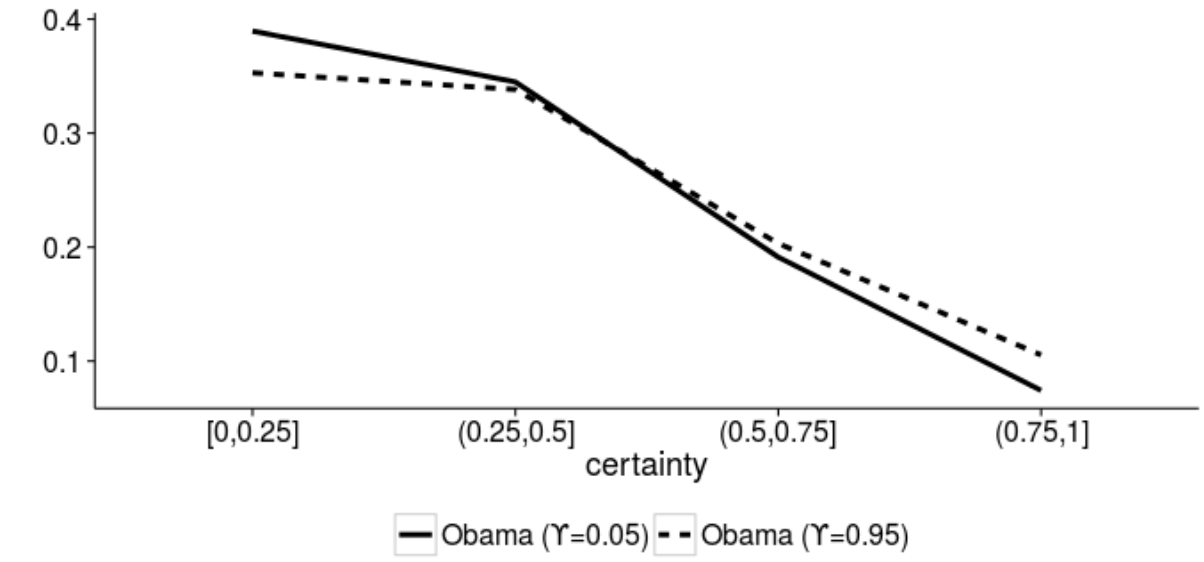
Root	γ	Log likelihood
Science	0.05	-1.98045×10^{14}
Science	0.95	-1.92137×10^{14}
Obama	0.05	-2.04097×10^{14}
Obama	0.95	-2.92894×10^{14}

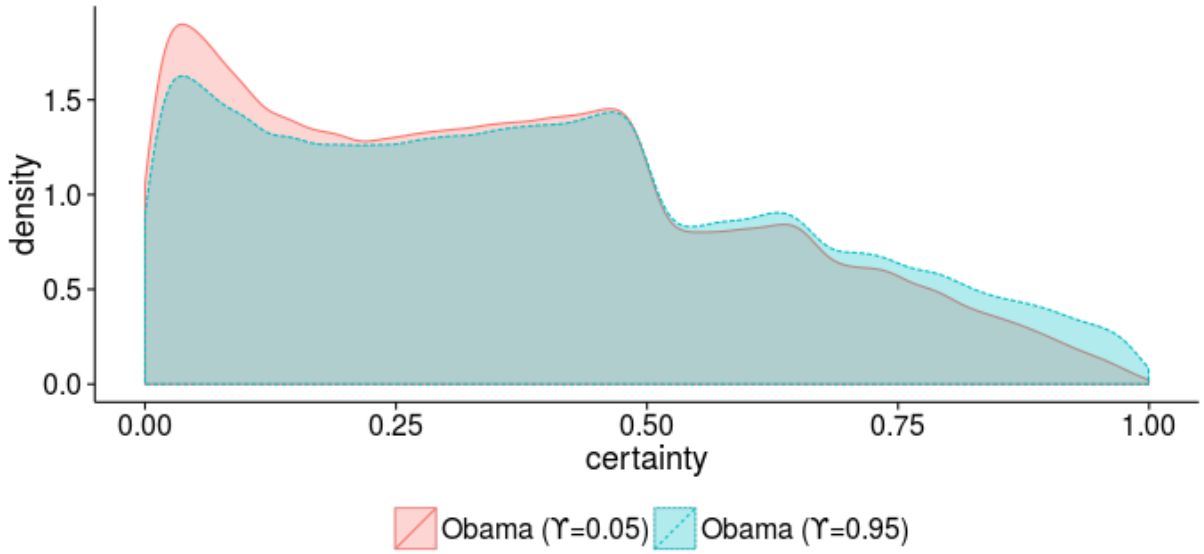
From previous discussion we make a conclusion that likelihood score is negative correlate to γ depth. This still holds on large graph, e.g. Wikipedia document graph. When Obama was picked as root, higher γ has a lower likelihood score. When Science was chosen, the correlation is minor.

Certainty


0.4
0.3
0.2
0.1
[0,0.25]
(0.25,0.5]
(0.5,0.75]
(0.75,1]
certainty
Science (ϒ=0.05)
Science (ϒ=0.95)


2.0
1.5
1.0
0.5
0.0
density
0.00
0.25
0.50
0.75
1.00
certainty
Science (ϒ=0.05)
Science (ϒ=0.95)





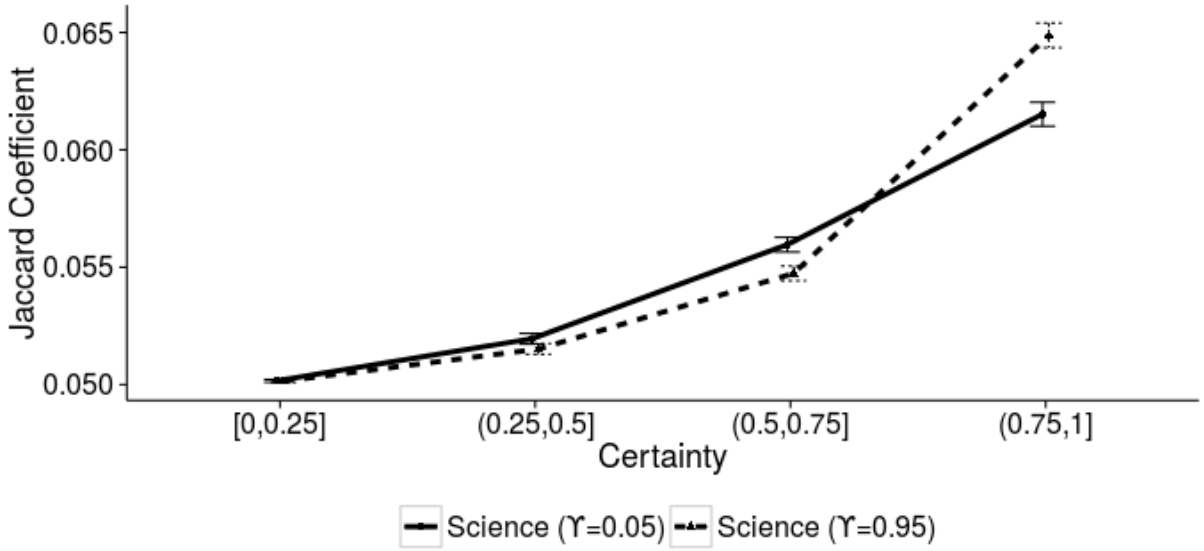
Recall that every document d needs to draw a path $\mathbf{c}_d$ during inference stage. The number of potential paths is determined by in-degree of d. Therefore, drawing $\mathbf{c}_d$ can be viewed as drawing from a multinomial distribution. We use the following formula to denote the certainty of selecting $\mathbf{c}_d$ from such distribution:

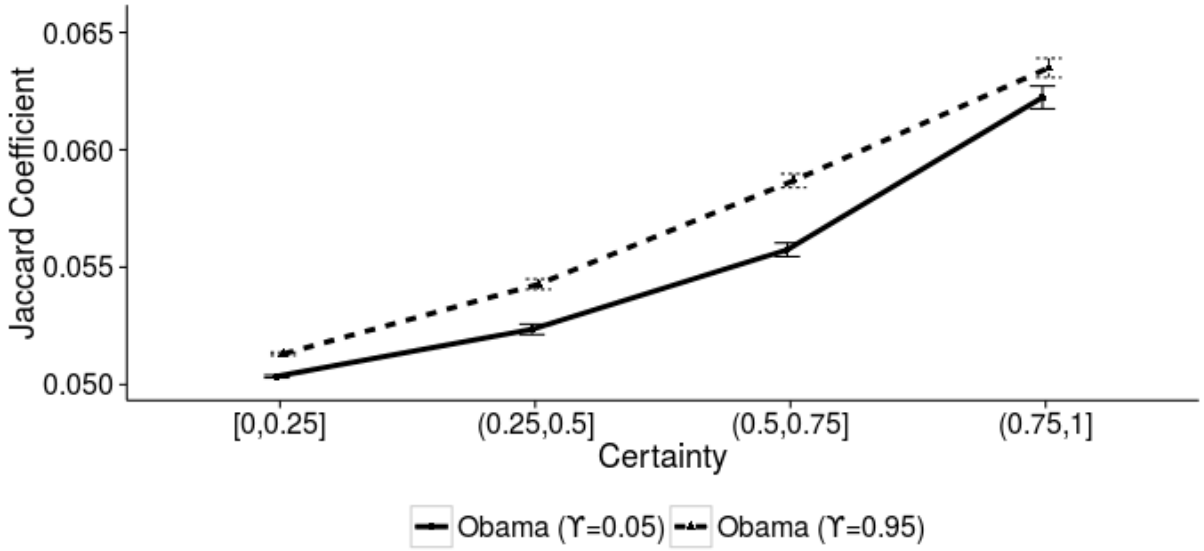
$$certainty_d = \frac{\frac{n_i}{n} - \frac{1}{deg^-(d)}}{\frac{n_i}{n}}$$

n represents the total number of trails, n_i is the number of trails that i exists in $\mathbf{c}_d$, in-degree is denoted as $deg^-(d)$. To eliminate the bias of in-degrees, we measure certainty by normalized difference of raw probability and probability of random guess.

When root is a document with general topic, the change of γ can not significantly affect the distribution of certainty, in fact the distribution is almost identical. Whereas with a specific topic root, larger γ will change certainty distribution, namely the number of lowest certainty nodes are reduced and the number of high certainty nodes are increased.

Jaccard Coefficient





Instead of calculate intra-cluster content likelihood, we choose to evaluate the similarity between parent document and its children. Because in some cases, the documents under a certain topic are diverse, but each of them has a strong conceptual association to parent topic. An example would be {Honest Leadership and Open Government Act} and {Alexi Giannoulias}. These two articles are directly connected to {Barack Obama}. The first one is an Act that Obama offered amendments to, the other one is an Illionis politician who endorsed by Obama during his 2006 campaign for Illinois State Treasurer. These two documents have less common but they all bonded to Obama, so they will be put in the same topic under Barack Obama according to HDTM.

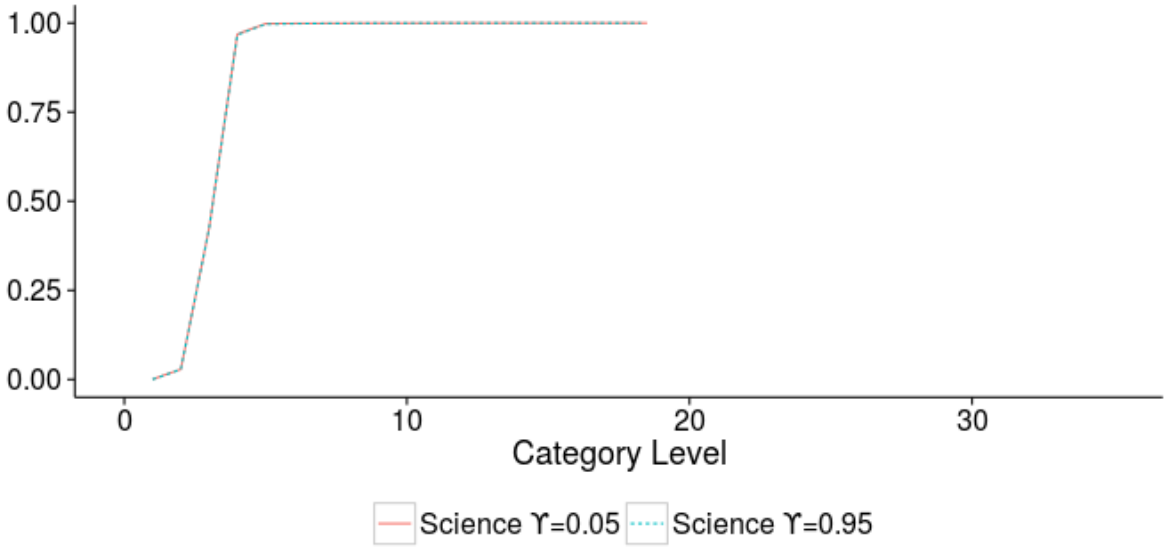
Such similarity is not based on document's content but category set of documents. The decision is based on the fact that category set represents human justification on document classification. Since we are comparing two sets, Jaccard Coefficient is chosen as the metric. For each document d, it belongs to a set of categories C, so the Jaccard coefficient is calculated by

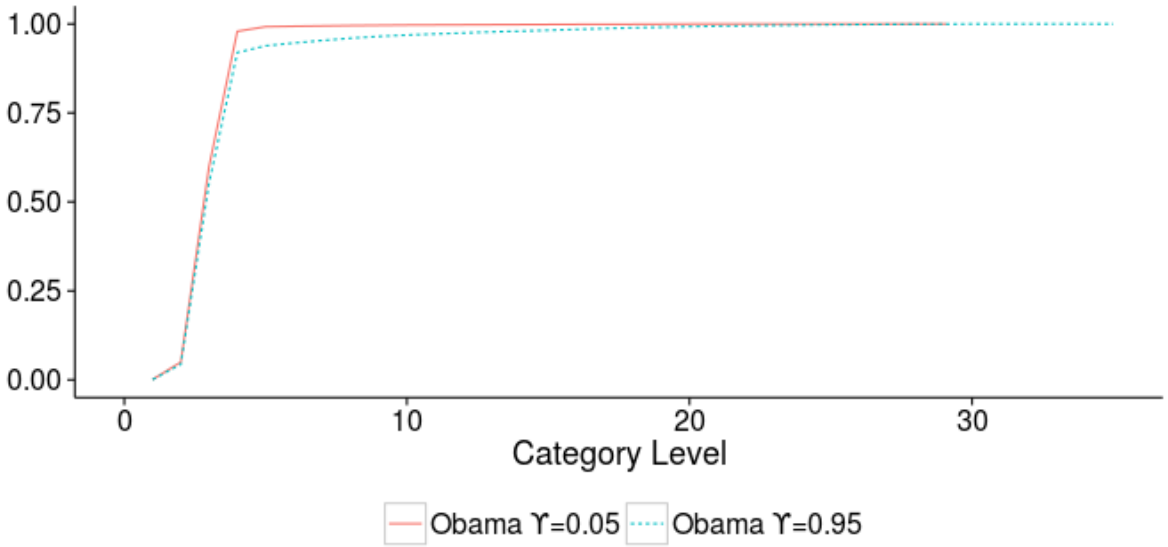
$$J_{score} = \frac{|C_d \cap C_{pa}|}{|C_d| + |C_{pa}| - |C_d \cap C_{pa}|}$$

Where d is a document, pa is d's predecessor on path $\mathbf{c}_d$, C is the corresponding category set.

When a specific topic is chosen as the root, Jaccard Coefficient is correlated to γ. Whereas there is no correlation if the root is a general topic.

ECDF





When hierarchy is a binary tree, we can use height to indicate the sparseness of hierarchy. But when dealing with real world data set which has more complex hierarchy structure, it is safer to measure the sparseness by topic size. But topic size still have drawbacks, namely it can not reveal more details about node distirbution.

Hence, cumulative distribution function (CDF) is chosen because it can display probability density among levels.

With CDF, an interesting result is discovered. Although sparseness is increasing with the increase of γ, the density on first few levels, in this experiment is first three levels, changes little.

DeltaCon

Discussion

Recall HDTM uses RWR to draw sampling paths. In RWR, γ is restart probability parameter. The larger the γ is, the less likely a long path will be drawn from original graph. That means the size of result graph correlate to γ. This is true when data set is small or the underlying hierarchy is shallow. When data set is large, {*e.g.*}, *Wikipedia document graph, it is more complex because the generative model we use in HDTM.*

For each inference, a sampling path is drawn at the beginning. Then HDTM draws topic proportion and does Latent Dirichlet Allocation based on sampled path. Path sampling favors shorter path by nature because a node near root will have higher affinity score. On the other hand, LDA favors longer path, which implies more specific topic and higher likelihood score. γ in HDTM plays an important role in balancing these two procedures. Namely when γ is small, the affinity score distributed more evenly. Under such circumstance, path drawing has less favor on shorter path because the difference of RWR scores on different hierarchy levels is subtle comparing to likelihood. Therefore RWR has less influence on deciding sampling path. Whereas a high γ results a power-law like distribution. In that case, a path with small length is more likely to be chosen. However, that only applies for very short paths. Because most of the affinity score lies on top levels, nodes lying on top levels will favor shorter path. Sampling result of nodes that do not have connection to top levels will be dominated by LDA.

This mimics how human curate hierarchy: First, a root node which represents desired perspective is chosen, then we structure the hierarchy according to relevance. A document strongly related to root should be placed on higher level, while those who are not relevant being classified by its content. This explains why higher γ has better Jaccard score.

When root is a general topic, γ does not have important influence on final result. The reason is that hierarchy rooted by general topic is a simple hierarchy that goes from general to specific. Document links and content, in this case, imply the same information. Therefore no matter how γ changes, the result will not differ much.