~~DNA~~ URL Next Generation Sequencing

**Juan Acosta**

Abstract

Background:

DNA Next Generation sequencing (NGS) or high-throughput sequencing is a term utilized to describe different DNA sequencing methods such as Illumina, Roche 454, and SOLiD, among others. These methods are used to determine the exact sequence of the four basic proteins (adenine, guanine, cytosine, and thymine) in DNA and RNA strands. Moreover, it can also be used to track mutations such as those of Salmonella-K and other pathogens.

The Hierarchical Dirichlet Process (HDP) is a nonparametric Bayesian method for clustering data, which is widely used in machine learning and probabilistic topic modelling. This unsupervised analysis of grouped data can be modified to become streamable and effectively eliminate the need for multiple passes through the data while optimizing the inference algorithm to produce real-time predictions on massive and streaming data.

Methods:

I propose an innovative approach to preprocess URL’s with DNA NGS techniques and an online variation of the HDP to predict domain sub-resource permanence through multiple visits to the same website. The process consists of 5 steps: browser user data collection, hashing and encryption of URL’s to eliminate personally identifiable information, deconstruction of the URL structure into basic components like those of DNA, application of the online variation of HDP on the deconstructed URL’s, and finally a visual analysis of the topic allocation to the different domains and domain sub-resources. The browser data used was obtained from both bots and real users of the Sparrow browser (ViaSat) as a proof of concept that the basic distribution of domain sub-resources through multiple visits of the same domain is sufficiently like that of DNA protein chains in multiple generations of a single species and the distribution of domains to that of DNA basic structures across species. Finally, the topic distribution across domains was graphically displayed for easier pattern recognition and result analysis.

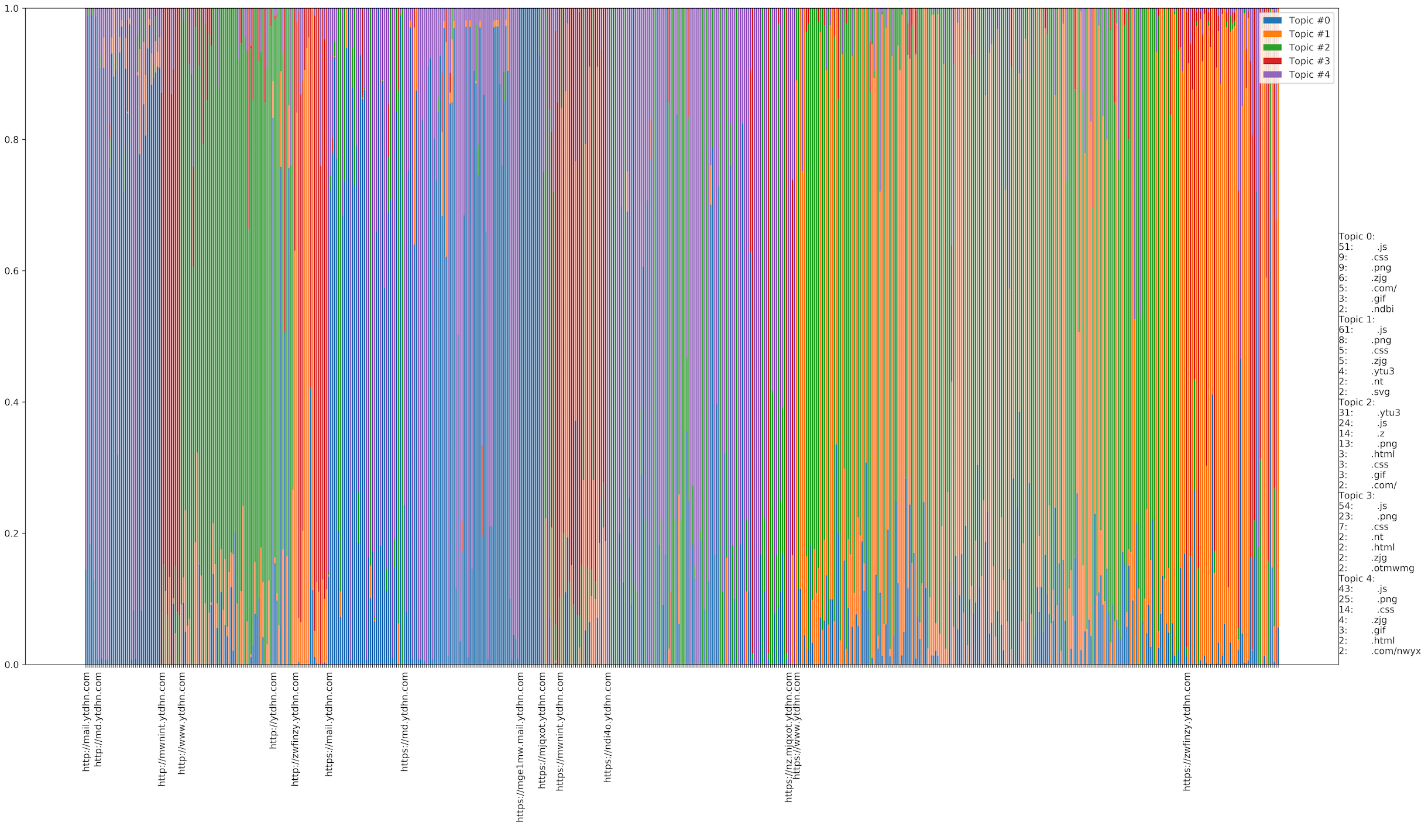
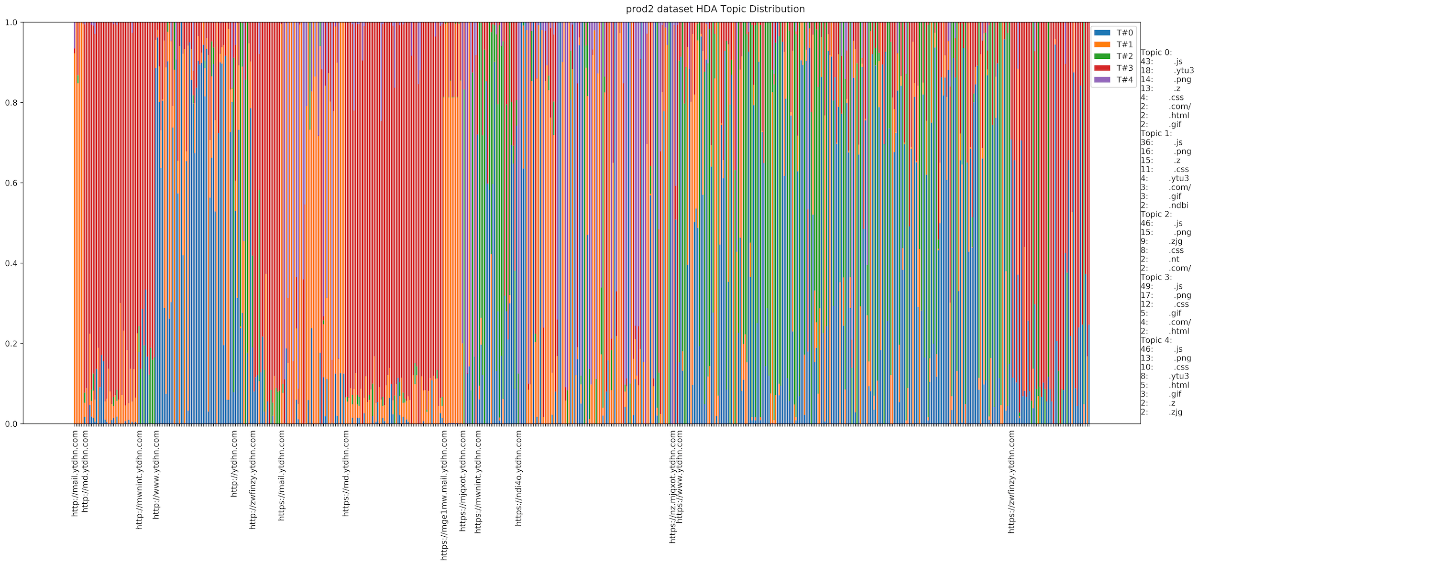
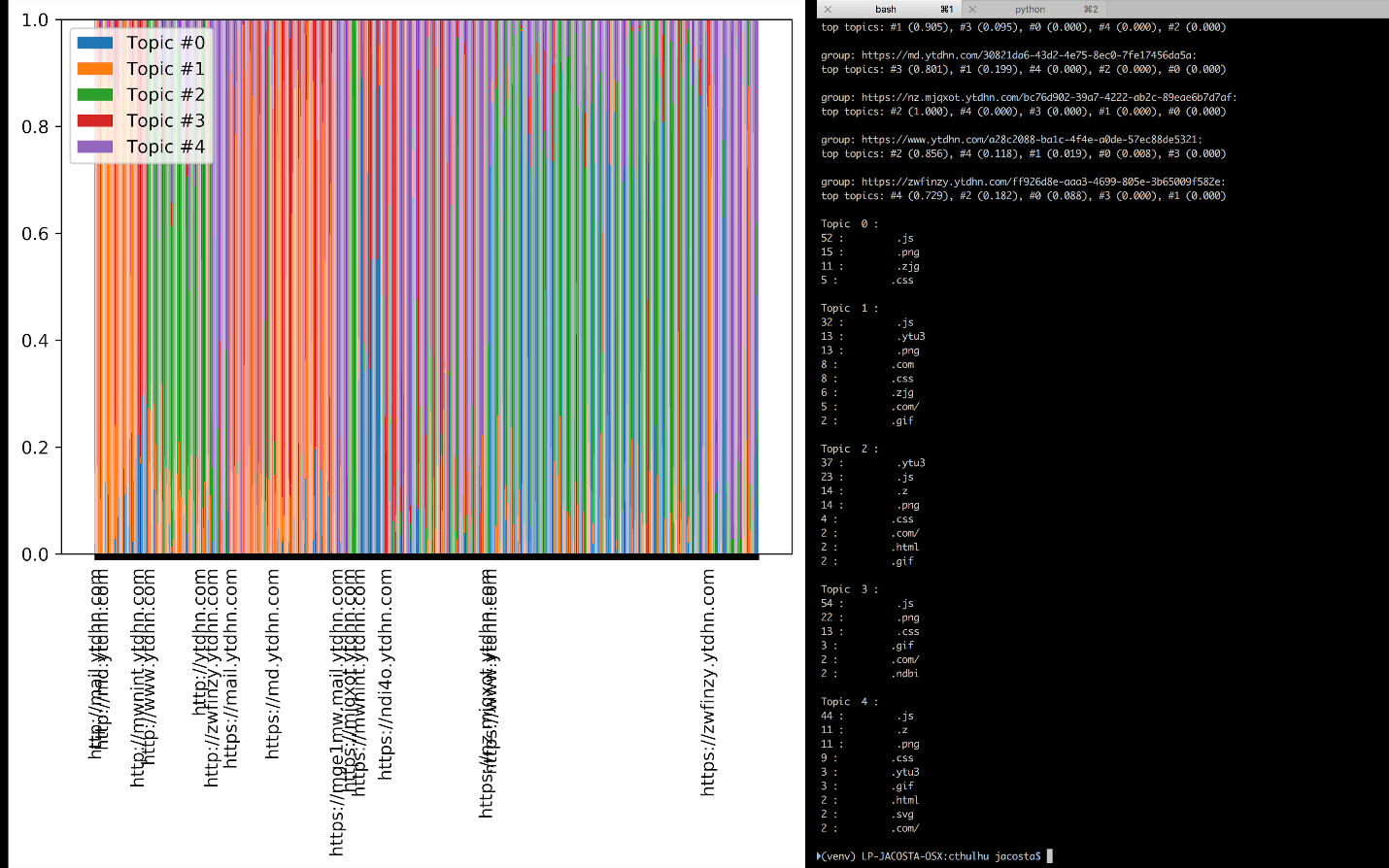
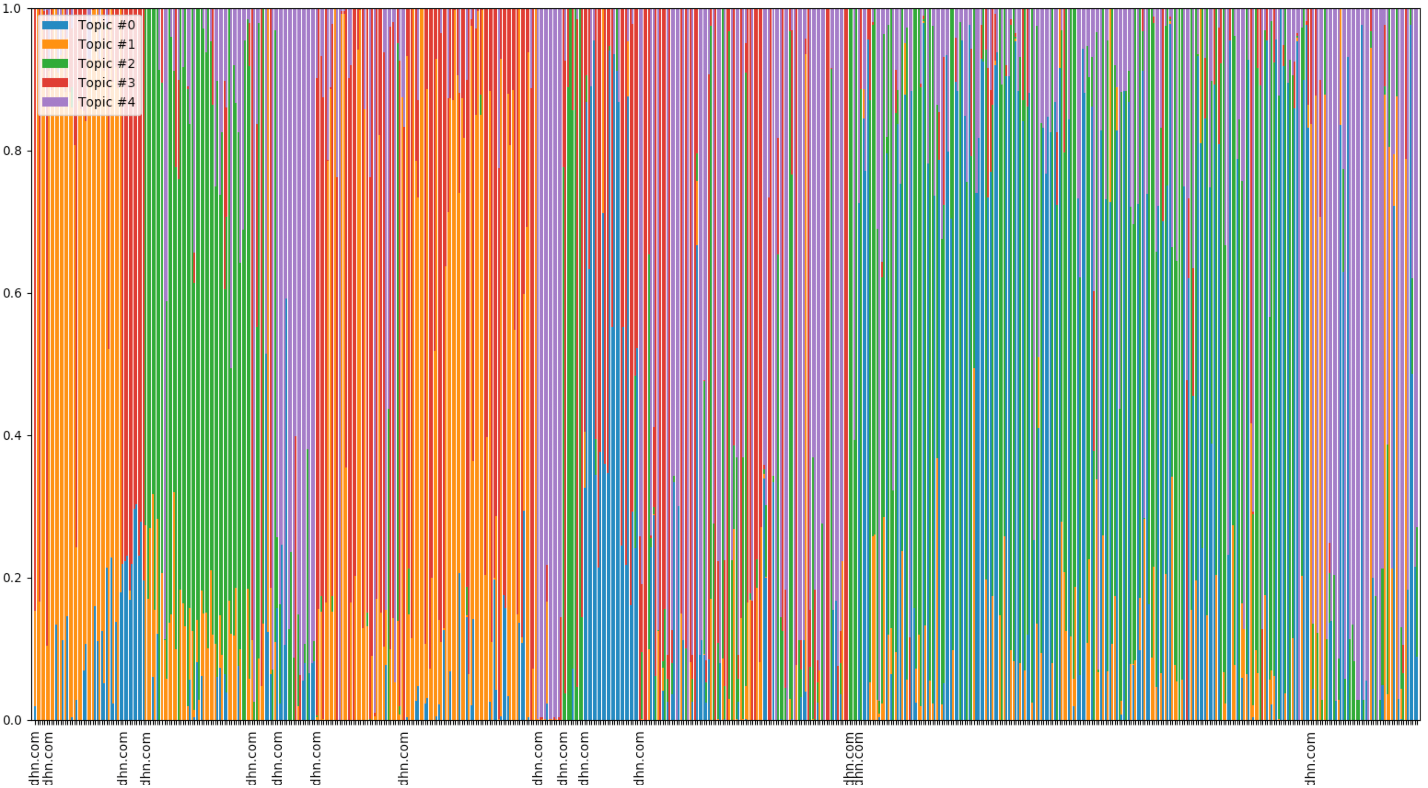
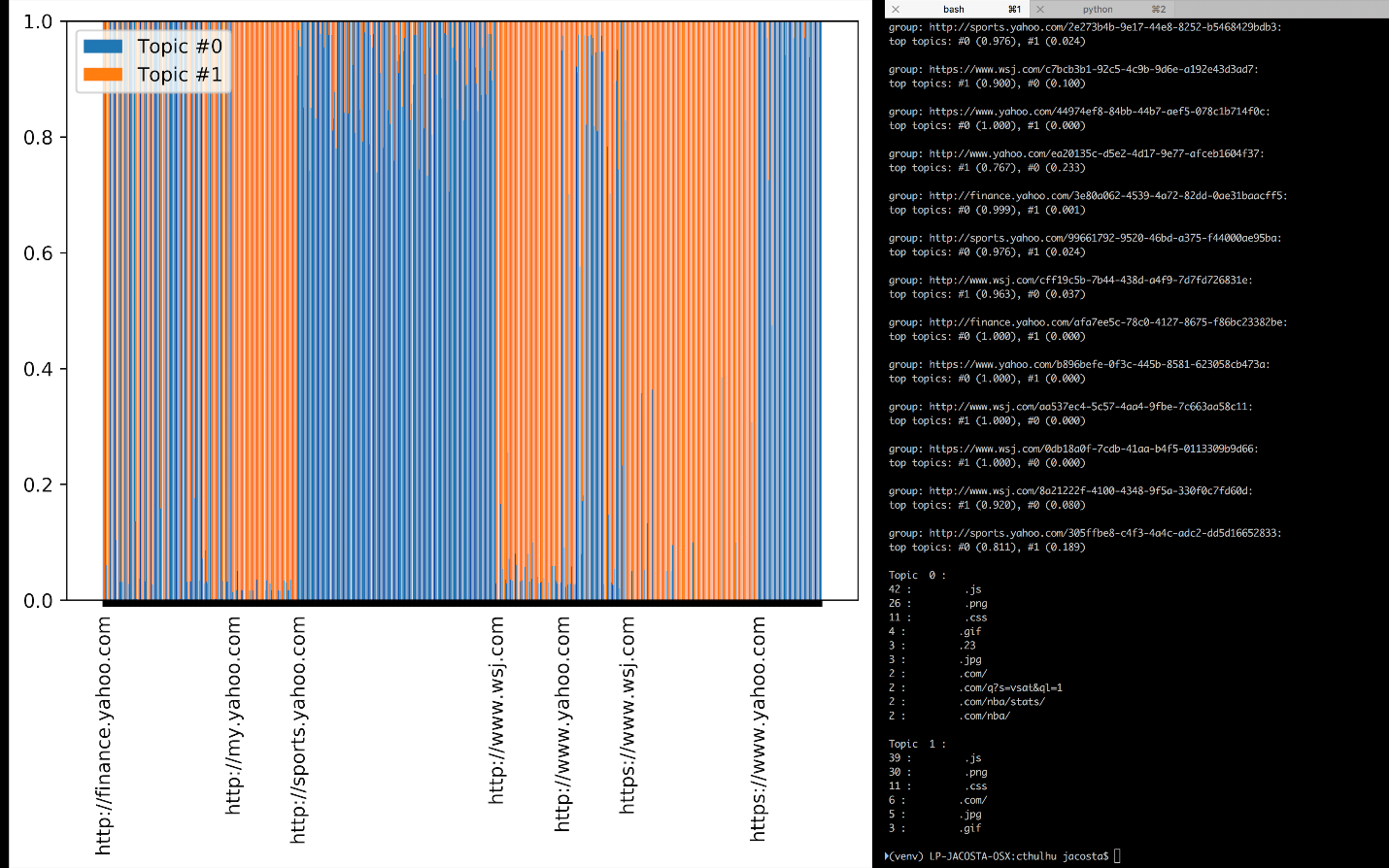
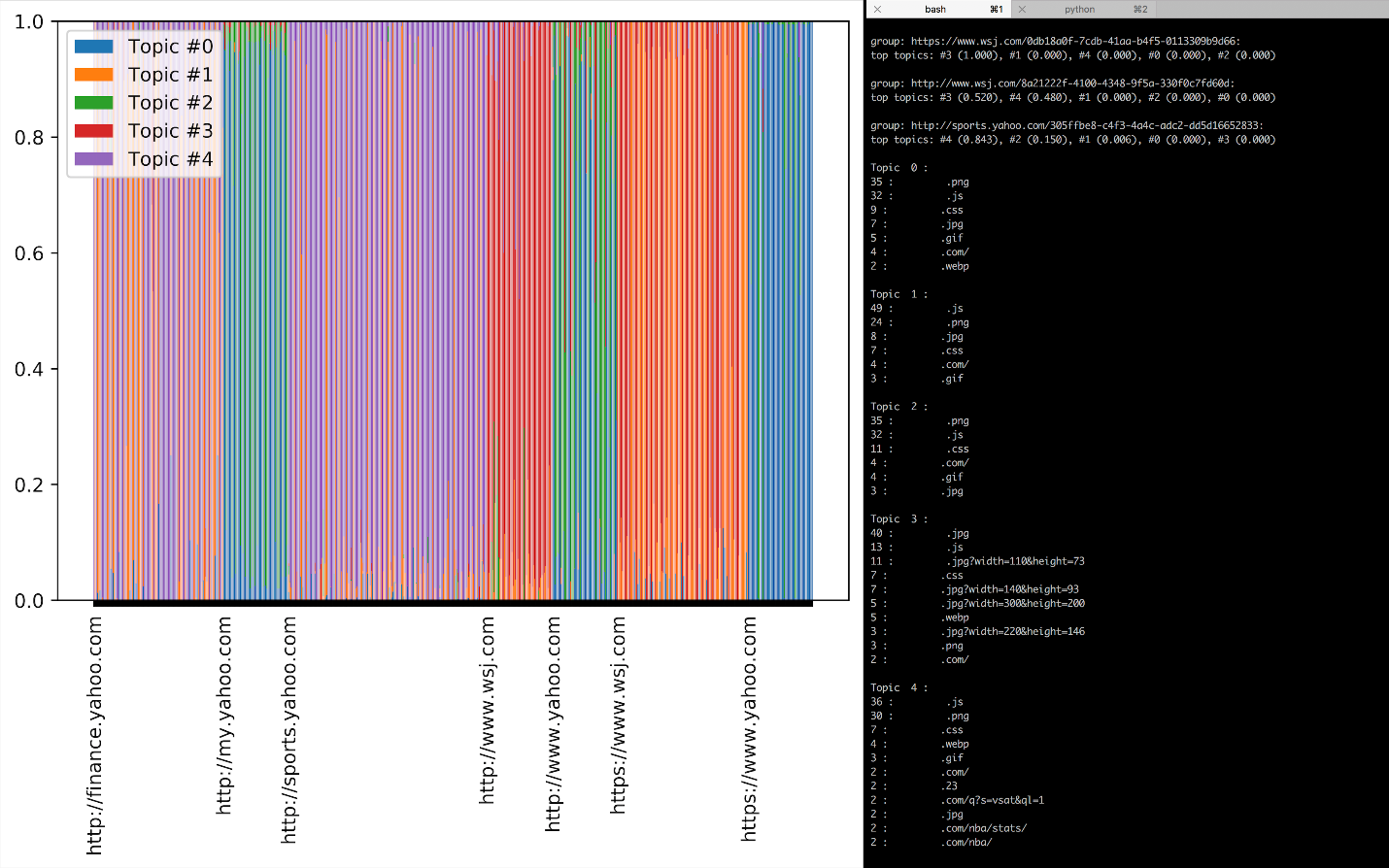
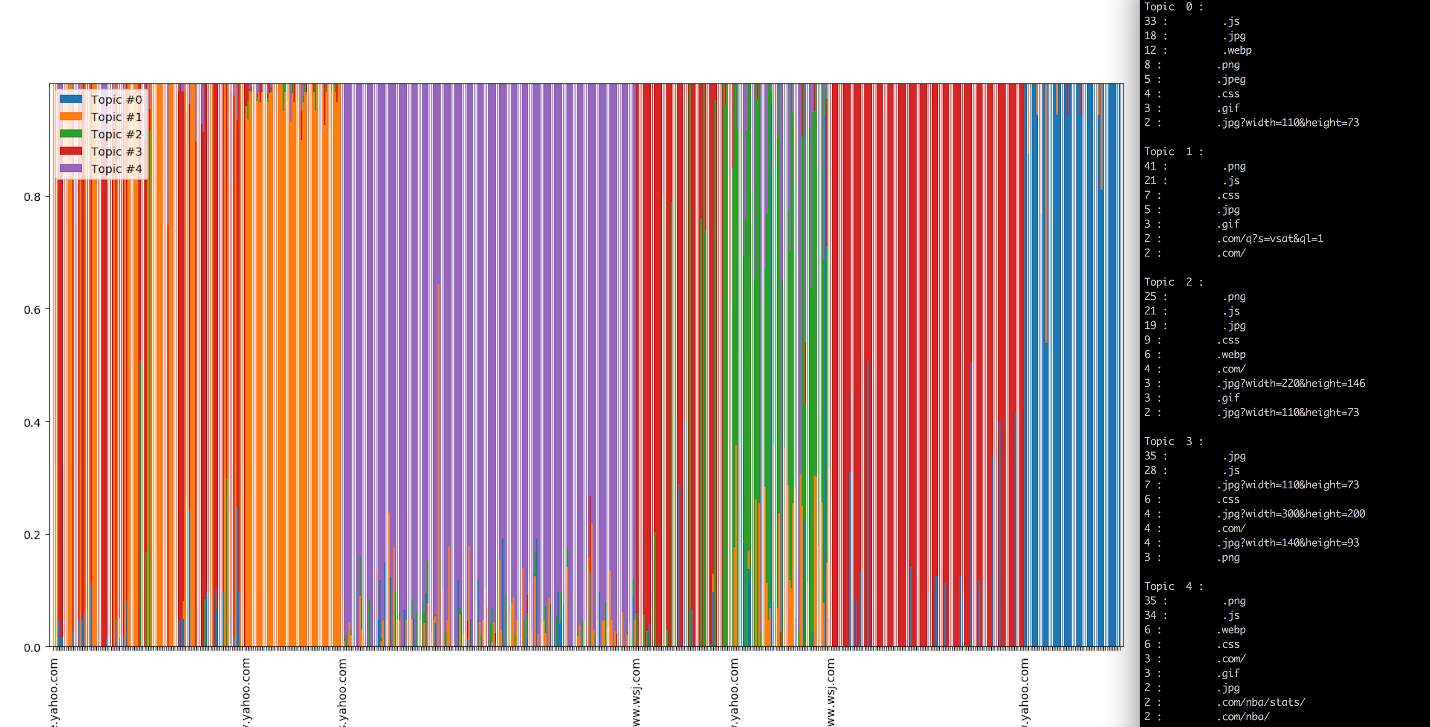
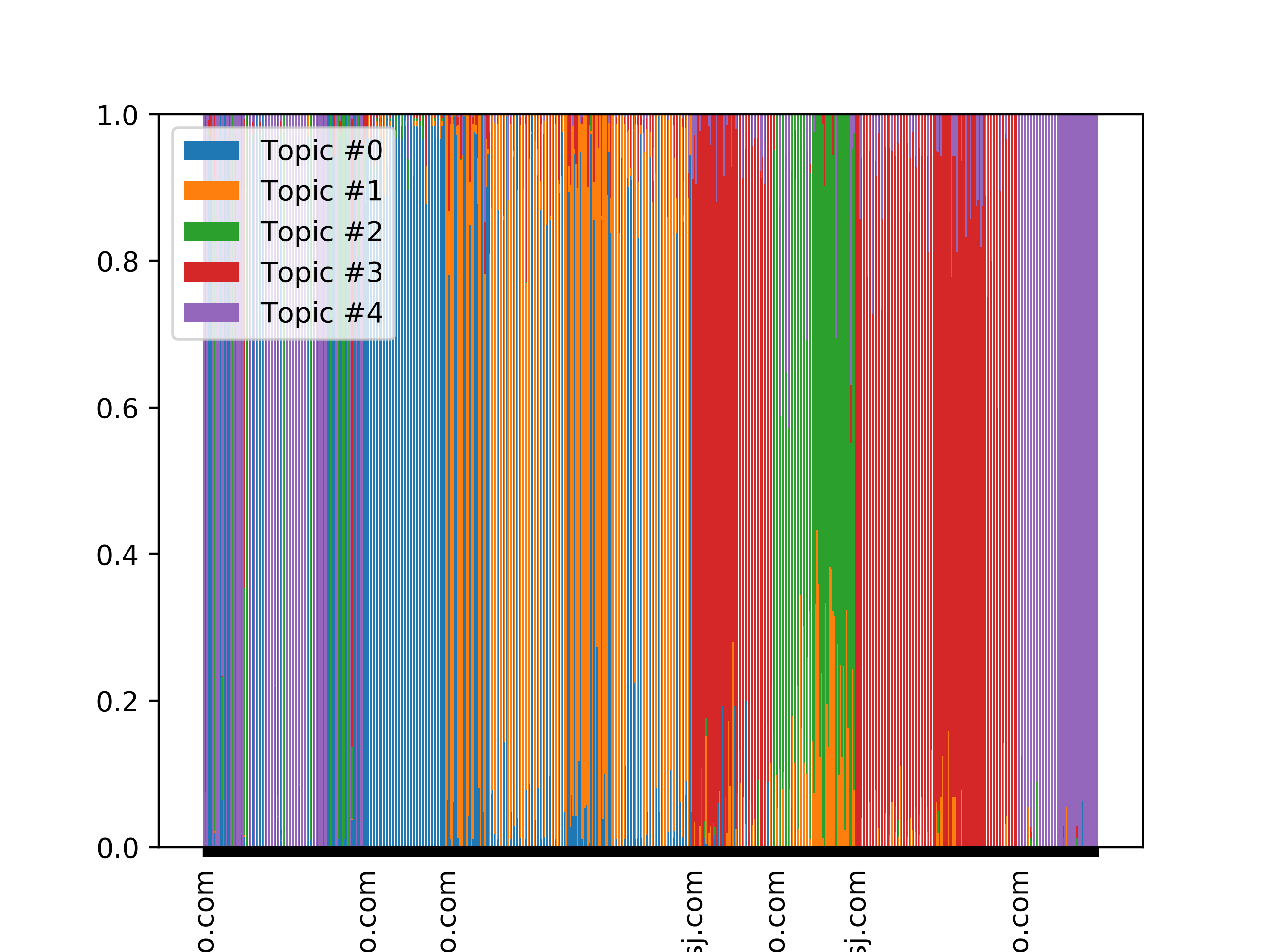
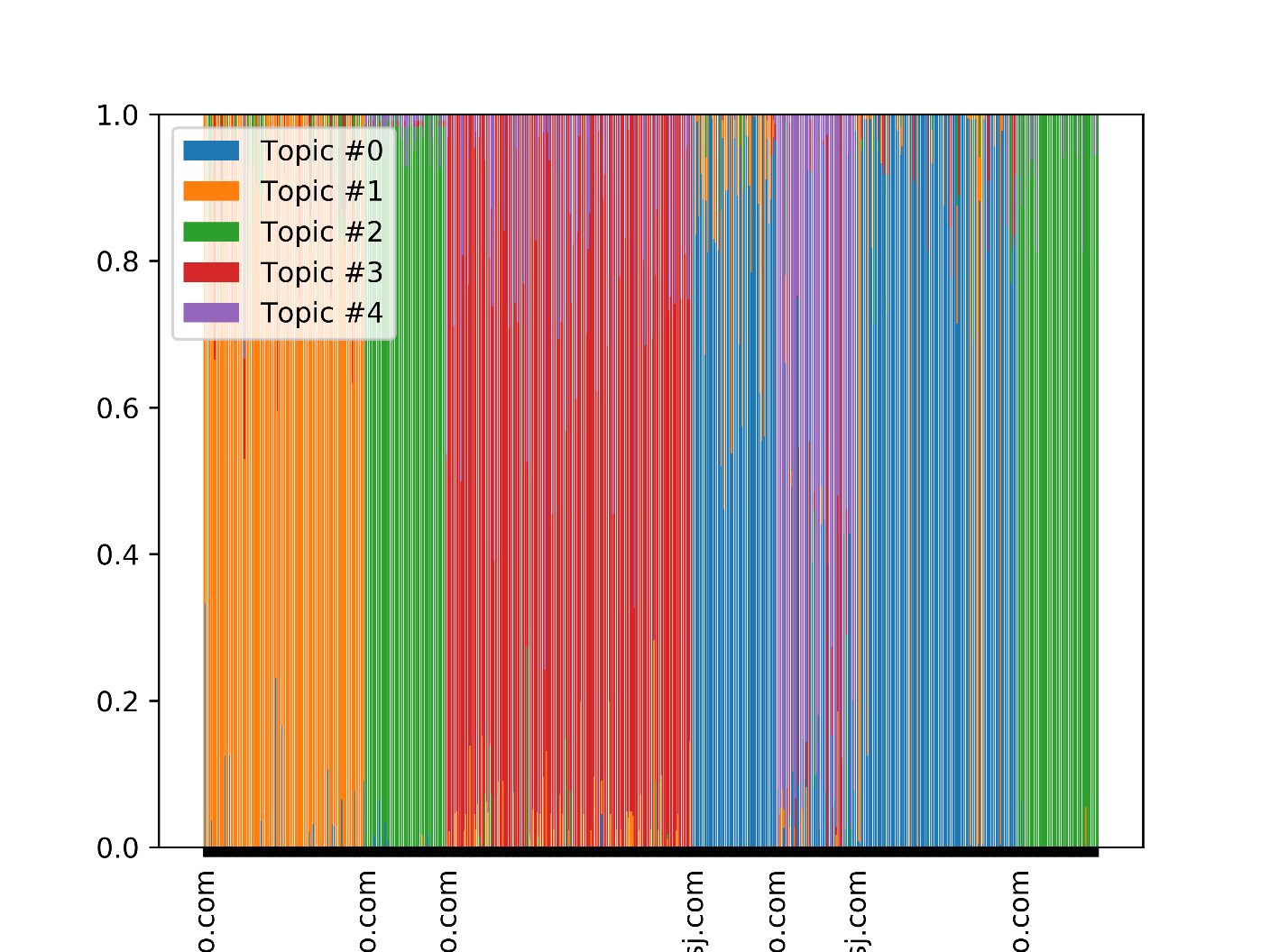
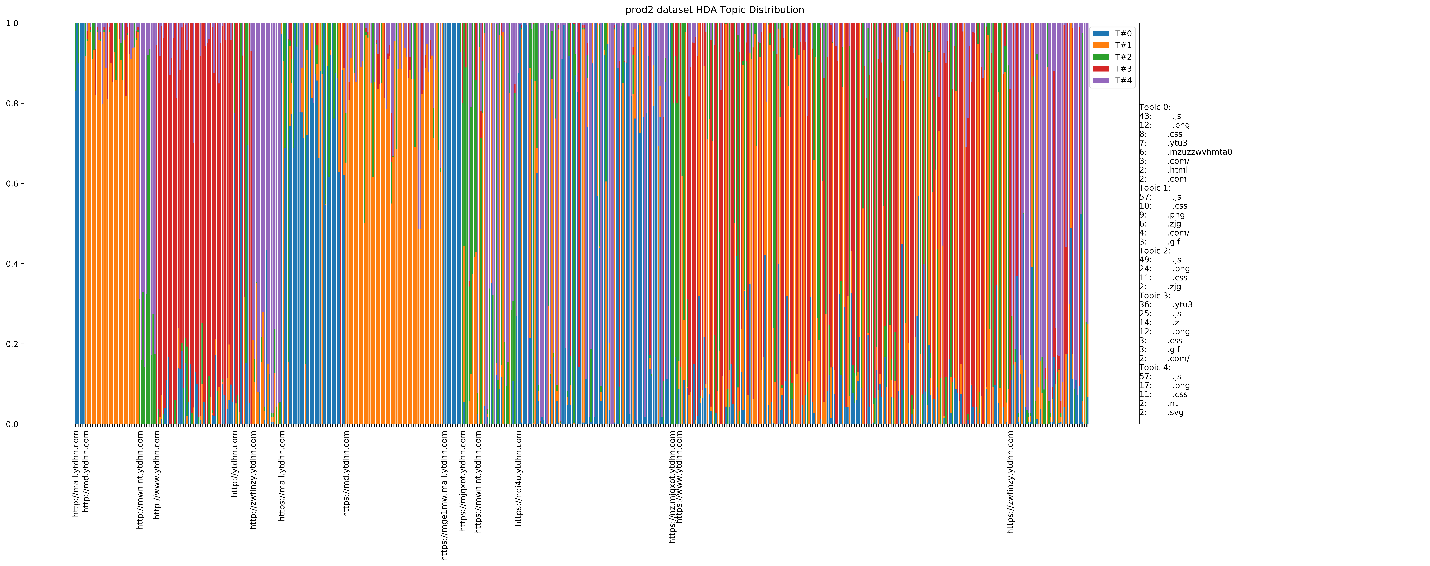
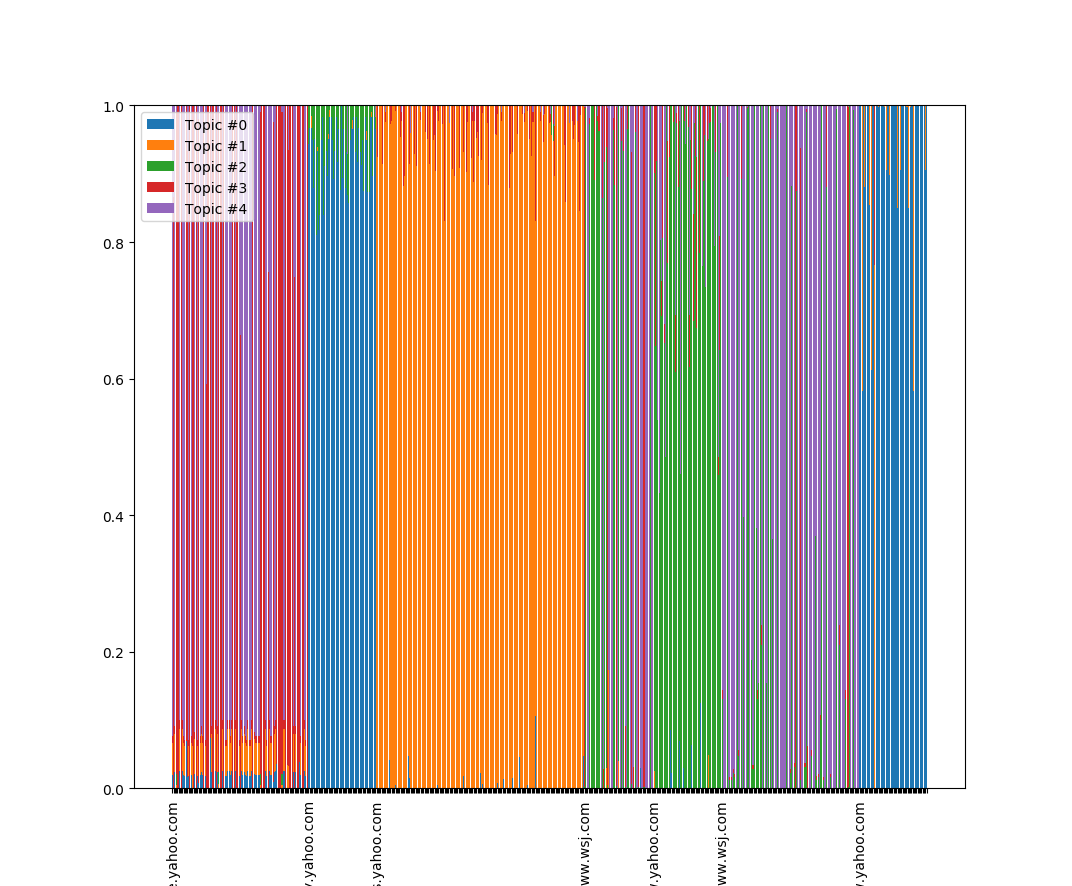
Results:

The output graphs displayed patterns on the topic allocation from the HDP, managing to differentiate between different types of sub-resources under a specific domain. Furthermore, the topic modelling across different domains with similar sub-resource permanence behavior displayed similar topic allocations.

Conclusion:

The use of DNA NGS techniques and the online variation of the HDP could provide for a better method to predict sub-resource behavior among the vastly different techniques implemented across different websites and in my opinion merits further exploration. This could prove to be crucial in the world wide and out of world implementation of faster satellite internet.

Appendix



References

Hoffman, M., D. Blei, F. Bach. *Online inference for latent Drichlet allocation*.

Wang, C., J. Paisley, D. Blei. *Online Variational Inference for the Hierarchical Dirichlet Process*.

Zhao, W., J. Chen, R. Perkins, Y. Wang, Z. Liu, H. Hong, W. Tong, W. Zou. *A novel procedure on next generation sequencing data analysis using text mining algorithm*.