Understanding Comments Submitted to FCC on Net Neutrality

Kevin (Junhui) Mao, Jing Xia, Dennis (Woncheol) Jeong

December 12, 2014

Abstract

We aim to understand and summarize themes in the 1.65 million comments and emails submitted to FCC regarding rule-making on net neutrality. We used the text of these 1.65 million comments, built an Latent Dirichlet Allocation model and summarized top twenty themes of these public comments on net neutrality.

1 Introduction

In seeking to propose new rules regarding net neutrality (anti-blocking and anti-discrimination), FCC opened to public comments and received unprecedented number of about 1.65 million replies from various parties. The purpose of our project is to understand what the public and various parties have to say on net neutrality regulation.

2 Dataset

The data is publicly available on FCC website [3].We included comments from both initial open comment period (May 15 - July 15), and the reply period (July 16th - Sept 10th), comprising of 1.65 million entries after cleaning. The meta fields include name of filer, name of author or lawyer, date of filing, city, state, zip code, text of the comments, etc. The comments are not labeled.

3 Features and Preprocessing

We focused on the text of comments in this project. During preprocessing phase, we split multi-email comments, removed non-alphanumeric characters, removed stopwords, generated unigram, bigram and uni+bigram feature sets, and finally converted each of the three feature sets into the data format that can be recognized by the LDA library.

4 Models

For this kind of unsupervised Thematic Analysis, we may use algorithms such as Nonnegative Matrix Factorization (NMF) [2], Probabilistic Latent Semantic Analysis (PLSA) [4] and Latent Dirichlet Allocation (LDA) [1].

For the purpose of our project, we decided to perform topic modeling using the open source LDA library from Mallet [6]. Mallet uses Gibbs sampling with hyperparameter optimization and can run in parallel with multi-threads. We used 80% of the data for training and reserved 20% data for topic inference. We have experimented with unigram, bigram and uni+bigram feature sets with varying number of topics, from 5 topics to 50 topics, incrementing in 5.

4.1 Latent Dirichlet Distribution

LDA is an unsupervised topic modeling technique. The basic idea of LDA is that each document contains a random mixtures of latent topics from which words are drawn. LDA treats documents as bag-of-words. The LDA generative process is illustrated by Algorithm 1 [7]

Algorithm 1 Latent Dirichlet Allocation Generative Process

Assume we know K topic distributions for our dataset, let V be the number of tokens in our corpus and M be the number of documents.

1. For each document i, a multinomial topic distribution $\Theta_i \in \mathbb{R}^K$ is drawn from a Dirichlet prior with parameters α
2. For each word in the document, a topic z_{ij} is drawn from the multinomial distribution Θ_i
3. Finally, the word w_{ij} is drawn from the multinomial distribution $\Phi_{z_{ij}} \in \mathbb{R}^V$, and $\Phi_{z_{ij}}$ itself is drawn from a Dirichlet with parameters β

Usually the Dirichlet priors on topics and words are symmetric. Details about inference can be found in the LDA publications [1] and [7]

4.2 Perplexity evaluation

We evaluate the perplexity on the test set for each of the unigram, bigram and uni+bigram feature sets. Perplexity here is defined to be

$$perplexity(\mathbf{D}_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} log(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

Where M is the number of documents in the test set. For each document d, it has N_d words and $\mathbf{w}_d$ is the sequence of words in the document.

The plot of relationship between *log-perplexity* and number of topics is given in Figure 1

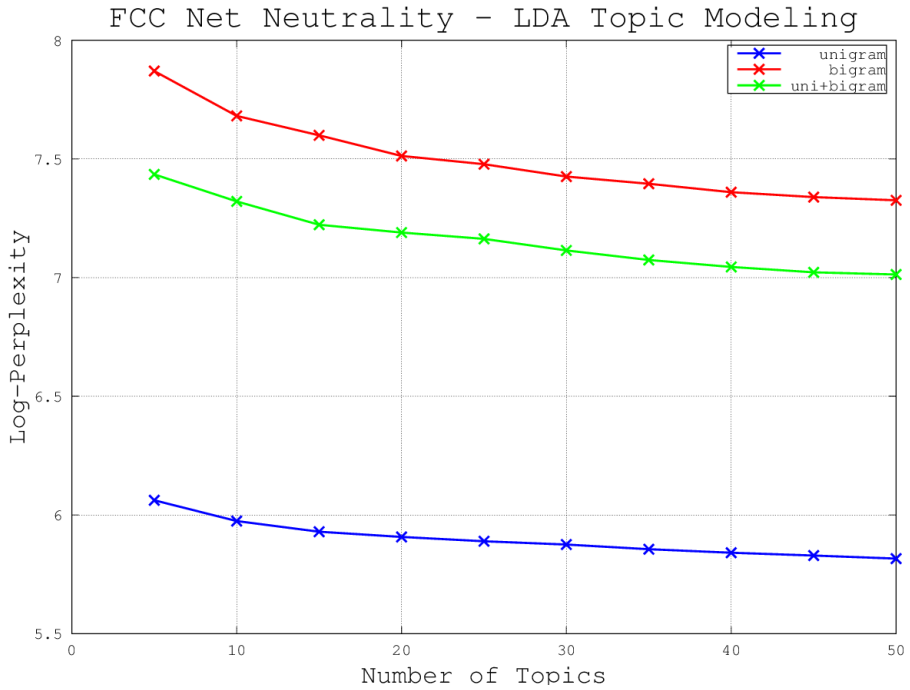

Figure 1: Plot of relationship between log-perplexity and number of topics. For given number of topics, our data set shows that basically unigram has the lowest perplexity score compared with bigram and uni+bigram feature sets

5 Results

The final result we present employs LDA model with unigram features and 20 topics. After we trained a model on the training data, we used the model to infer topics on the testing data. The output is a matrix $X \in \mathbb{R}^{M \times K}$, where M is the number of testing Documents and K is the number of Topics. Each row $X_i^T \in \mathbb{R}^K$ is the probability distribution of topics in document i

5.1 Probability mean of topics based on testing data

The mean probability for each topic is calculate cross all test data. The plot for probability mean of topics and the top-3 topics with highest probability scores are given in Figure 2 and Table 1.

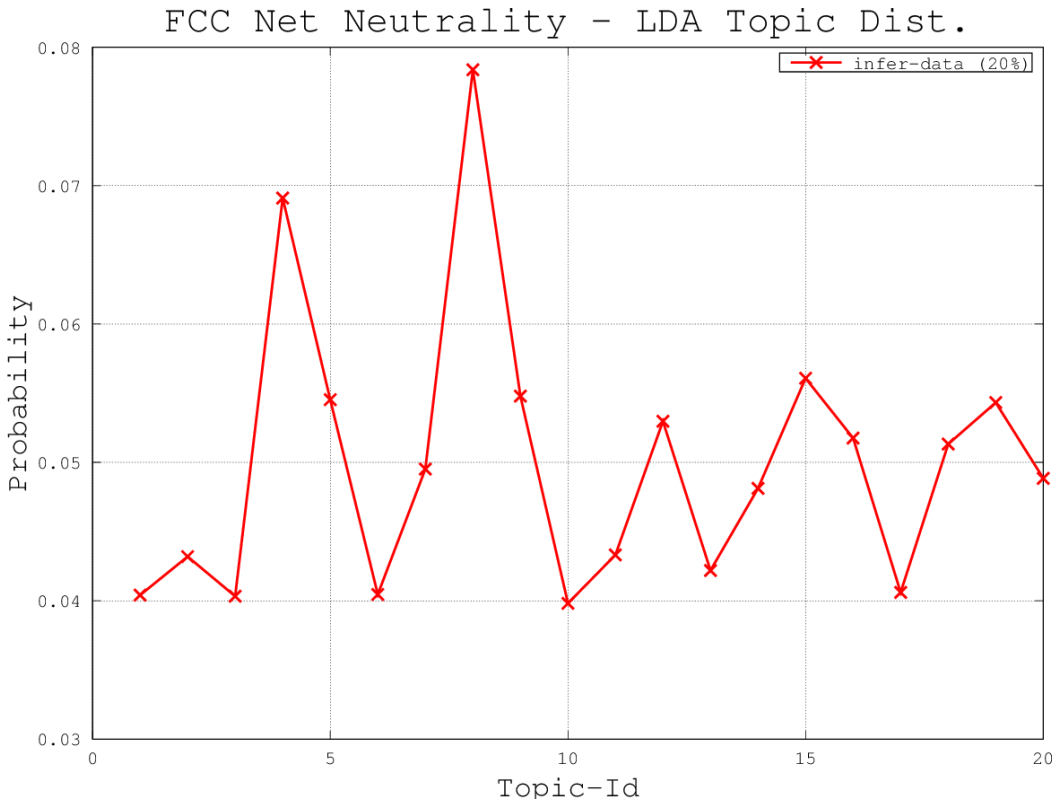

Figure 2: Plot of mean probability of topics based on test data

Table 1: Top-3 topics with highest probability

Rank	Topic-Id	Keywords
1	8	isps can use choice title speed destroy slow business able also others experience economic services first high company work bad
2	4	companies don cable like one content many just good comcast get see even big already thing people world idea make
3	15	economic rule opportunity fewer democracy entrepreneurs certainty must access rules protect powerful investors proposal ensure businesses strong remembered current erecting

5.2 Histogram of primary topics

For each document, we chose the topic with the highest probability as the primary topic for that document. We then count the number of documents for each primary topic to generate this histogram. The plot for histogram of primary topics and the top-3 topics with highest document frequency are given in Figure 3 and Table 2. Please note that the first two topics in Table 2 are the same as the first two topics in Table 1

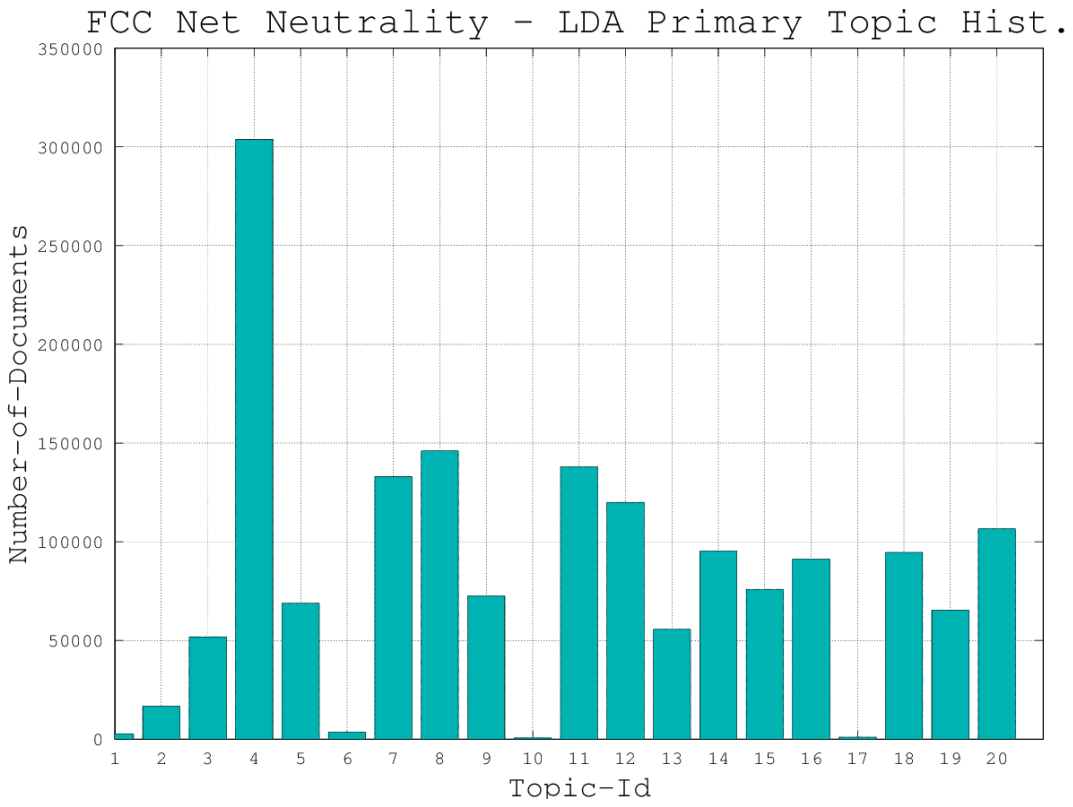

Figure 3: Histogram of primary topics

Table 2: Top-3 primary topics with highest document frequency

Rank	Topic-Id	Keywords
1	4	companies don cable like one content many just good comcast get see even big already thing people world idea make
2	8	isps can use choice title speed destroy slow business able also others experience economic services first high company work bad
3	11	service common providers internet carriers communications broadband classified reclassify act want title proposed fcc federal stop telecommunications chairman past must

5.3 Word cloud of top topics

We extracted the terms from the top topics in Table 1 and Table 2, then plot the word cloud based on term frequency. The top-10 high frequent terms are listed in Table 3

Table 3: Top-10 high frequent terms from top topics

Rank	Term	Term-frequency
1	isps	738,377
2	economic	463,104
3	title	425,251
4	choice	422,214
5	can	421,399
6	use	393,044
7	service	334,203
8	also	318,011
9	speed	271,061
10	business	270,037

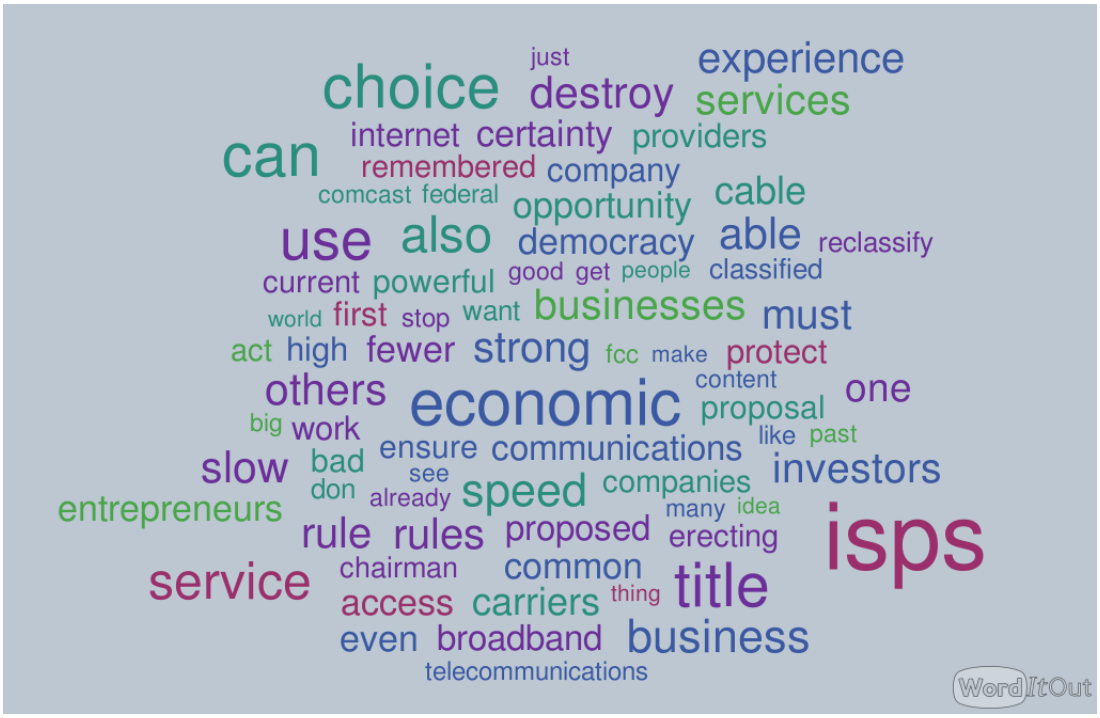

Figure 4: Word cloud for the terms from top topics

6 Conclusion and Discussion

We trained an LDA model to find the top 20 topics from 1.65 million comments on net neutrality. Our findings in Table 1 and Table 2 suggest that:

- People were most concerned with the economic impact of internet service providers having power over internet-based companies.
- People also cited Comcast's dominance and its bundling of cable packages with internet.
- Finally, the third most frequent topic involved the conflict between entrepreneurship in a democratic society and the protection of powerful companies.

7 Future

Some of the topic words, such as 'internet', 'net', 'also' and 'can', were not particularly helpful and in hindsight should have been added to the stopword list. For future directions, one could contrast our LDA results with NFM or PLSA models. We could also make better use of meta-information, use SVM for expert comments detection. Our code is available in GitHub [5]

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. 2003.

[2] Ding Chris, Li Tao, and Peng Wei. Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing. 2006.

[3] FCC. Electronic Comment Filing System. http://www.fcc.gov/files/ecfs/14-28/ecfs-files.htm.

[4] Thomas Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. 2001.

[5] Kevin Mao. Net Neutrality GitHub. https://github.com/kevinmao/fcc-nnc.

[6] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

[7] Alexander Smola and Shravan Narayanamurthy. An Architecture for Parallel Topic Models. 2010.