What are People Saying about Net Neutrality?

Adison Wongkar
(adison@stanford.edu)

Christoph Wertz
(cwertz@stanford.edu)

Introduction

The issue of Net Neutrality has recently made it into the headlines. The FCC has established an open inbox where people can submit their comments on Net Neutrality and has made the dataset publicly available. Given the 800,000+ individual comments in the dataset, it becomes an immense challenge to understand the major topics addressed by what people are saying and to be able to quickly group comments with similar main ideas together. This project aims to do unsupervised clustering and topic modeling to effectively learn the main ideas of what people are saying in their comments. The main challenge in performing such analysis is that the textual comments are often redundant and may use variation of terminologies to describe the same concept. Our goal is to partition these unlabeled examples into clusters and discover natural categories in an unsupervised manner using LDA. Topic alignment, F1 measure, and perplexity evaluation are used to diagnose and optimize our choice of parameters.

Dataset, Features, and Preprocessing

We use the FCC published dataset[FCC ECFS] which includes comments entered in ECFS (Electronic Comment Filing System) before 18 July 2014. The dataset contains 801,781 documents, each of which contains **id** (unique id of the filing) and **text** (main body of the comments), along with other metadata. We performed basic data cleaning to remove non-readable artifacts that come from OCR or noise in text extractions from attachments (like .ppt). Data Cleaning is done by: (1) normalizing accented terms, (2) dropping illegible artifacts described above. Data Preprocessing and Tokenizing is done by: (1) normalizing all terms to lowercase, (2) using English tokenization rule, (3) filtering out terms w/ < 3 characters, (4) removing common English stop words, and (5) normalizing spelling variations on some important terms such as “ISP”, “pay-to-play”, “Commissioner/Chairman Tom Wheeler”, “common carriers”, etc.

Topic Modeling

LDA (latent Dirichlet allocation)[Blei 2003] is a generative probabilistic model for collections of discrete data such as text corpora. It models documents as a random mix of latent topics, which is characterized by a distribution over terms, and infers latent topic structure that is most likely to generate the observed corpora. We use smoothed LDA that assumes the following generative process for each document **w** in corpus D: (1) Choose N ~ Poisson(ξ), (2) Choose θ ~ Dir(α), (3) Choose β ~ Dir(η), (4) For each of the N words w_n: (4.a) Choose a topic z_n ~ Multinomial(θ), (4.b) Choose a word w_n from $p(w_n|z_n, \beta)$.

We use Stanford Topic Modeling Toolbox v. 0.4[TMT] to perform LDA training and inferring on our dataset. Unless mentioned specifically, we use the default topic & term smoothing = 0.01 and filtering out X = 30 most common terms. And for inference, we use Gibbs Sampler method with 1500 max iterations (at the end of which we observed convergence as evidenced by stabilized log probability estimates).

Result Diagnostics

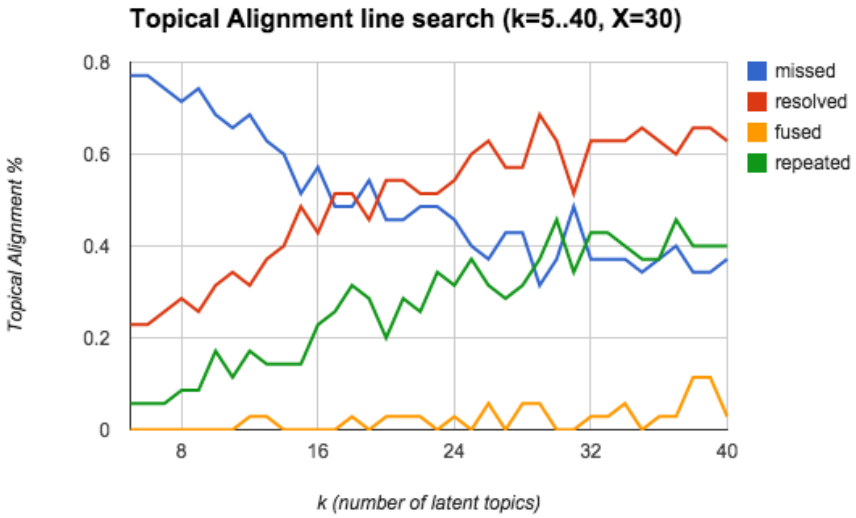
Choosing the parameter k (number of latent topics) is extremely critical, as is X (number of most common terms filtered out). We use various diagnostics to find the optimal choices of k and X.

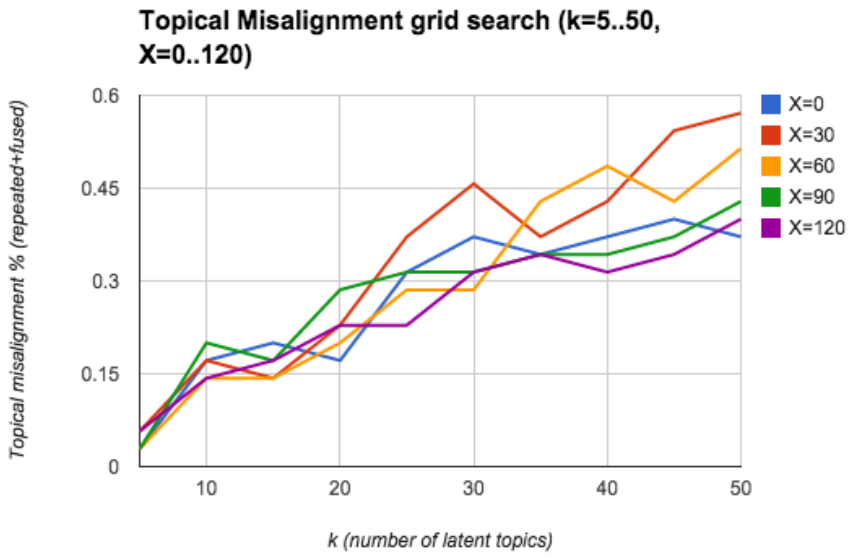
Topical Alignment

To evaluate the fitness of our parameters, we apply topical alignment technique similar to Chuang et al. [Chuang 2013] which calculates matching likelihoods for topic-concept pairs (where the concepts are identified by human/expert as reference). But due to our low number of reference concepts compared to k (number of latent topics), we calculate the matching likelihood of concept-terms to topics and compute **resolved**, **missed**, **repeated** as the probability that a concept term appears (in one or more topic), is missing (in all topics), and is repeated (in two or more topics). And similarly, we compute **fused** as the probability of two terms from separate concepts appearing in the same topic. The repeated + fused value is a proxy of how misaligned are the reference concepts with the discovered topics. We used expert/reference concepts published by Sunlight Foundation[Sunlight] below, for a total of 6 concepts and 35 concept terms (which we express as regex), as shown below.

Concept 1	**Concept 2**	**Concept 3**	**Concept 4**	**Concept 5**	**Concept 6**
slowlane(s)?	common	important(ly)?	work	access	monopoly
fastlane(s)?	commoncarrier(s)?	vital(ly)?	competition	choice	competition
paytoplay	classif(y|ying|ied)	economy	startup(s)?	entertainment	comcast
wealthy	reclassif(y|ying|ied)	essential	kill(s)?	fee(s)?	verizon
divide	authority	resource(s)?	barrier(s)	content	warner
netflix		cornerstone	entry	extort	
				extract	

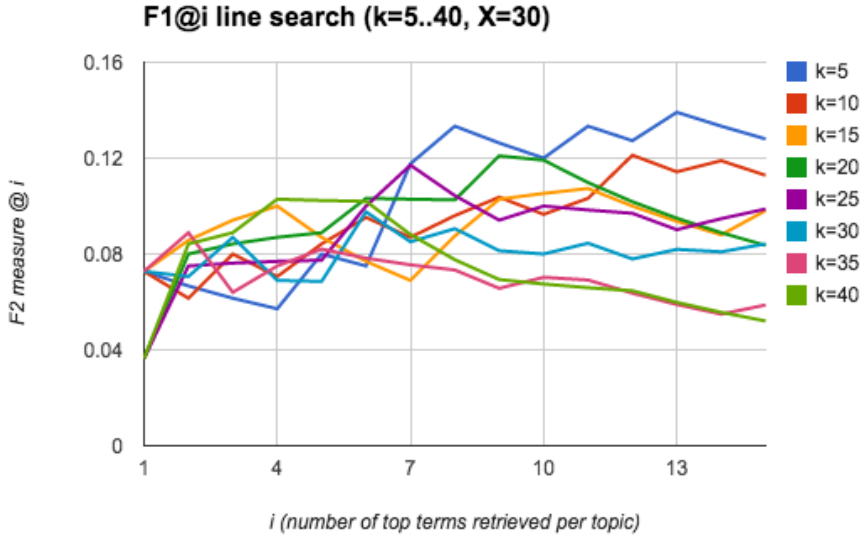
We performed both line search with k={5..40} and grid search with X={0, 30, 60, 90, 120} and k={5..40} and plot the topical [mis]alignment graphs. From line search, we see the rise of repeated topics with larger k, especially k>16. And the resolved/retrieval rate is about 50% at k~16. From grid search, we observe better performance (relatively lower misalignment) for X={0, 30} for k<20.





Concept Terms Retrieval (F1)

F1 score (the weighted harmonic mean of precision and recall) is often used in IR to measure document classification performance. We evaluated how well the terms in our topics perform to retrieve the reference concept terms. In particular we observe the F1 score for various choices of k={5..40} with our default parameters X=30 and term/concept smoothing=0.01. In computing precision and recall at i, we considered the retrieved set at i to be the union of top i terms in all discovered topics. We found that k=5, 10, 15 performs better compared to larger choices of k, even though we are considering proportionally larger retrieved terms with larger k.



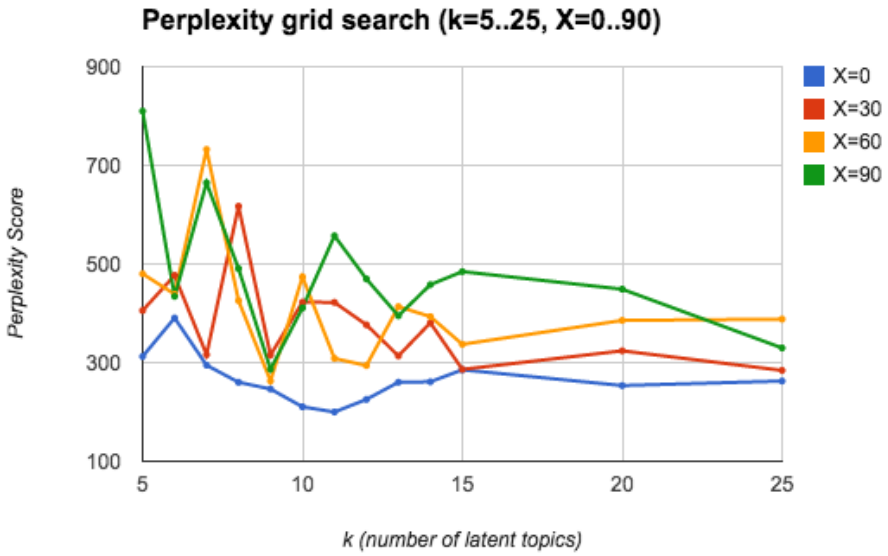
Perplexity Evaluation

Perplexity is a way of evaluating topic model in NLP. We perform perplexity analysis by: (1) randomly splitting the dataset into: 80% training docs (for training LDA model), and 20% hold-out docs (for evaluating perplexity on unseen data); (2) finding parameters that minimize the model's perplexity on held-out data. Lower perplexity is seen as a measure of goodness of fit for LDA parameters based on the held-out test data. Perplexity score on each document is determined by: (1) splitting the document in half, (2) estimating per-document topic distribution on the words in the first half of the document, and (3) computing the "surprise factor" (i.e. the number of equiprobable word choices, on average) of

encountering words in the second half of the document. More formally perplexity is defined below, assuming the test documents D_{test} contains M documents, each of which contains words $\mathbf{w}_d$ for a total of N_d words. The $p(\mathbf{w}_d)$ is estimated from Dirichlet distribution learned from D_{train}.

$$perplexity(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\}$$

Below is the perplexity score grid search with X={0, 30, 60, 90} and k={5..25}. We see that X=0 consistently produced lowest perplexity, followed by X=30. The lowest perplexity seems to be around k=10.



Discussion

From all the diagnostics above, we see that the optimal parameters would be around k=5..15, and X=0..30. And after applying our human judgment to the results, we settled on k=10, X=30 (and default topic/term smoothing=0.01) as our chosen parameters. This agrees with results we see in perplexity evaluation where perplexity is at minimum around k=10. Choosing X=30 also made the most sense because if we remove no terms, the most frequent terms that virtually exist in all documents such as “net neutrality”, “FCC”, and “internet” would be repeated in most of the topics as one of the top terms. On the other hand, if we remove too many terms, we would lose the most important terms that define the topics. Below is the summary of the discovered topics (we name our topic titles) and list of its top keywords, collapsing similar topics and discarding terms we think are junk:

Reclassify ISPs as common carriers	**Economic Impact**	**Objecting to idea of fast/slow lane**	**Objecting to idea of pay to play**	**Internet as an essential freedom**
common carriers service broadband telecommunications **reclassify** pay faster act communications rules content open free new	use choice **business** speed services slow destroy able **economic** experience **level playing field** important strong	utility public **tom wheeler** service commissioners time lanes private declare economy life live modern want **fast** allow **slow lanes** telecommunications street free broadband	**important services** new better **pay-to-play** innovative principle succeed live treat equally travels worries data dear subscribers **protect open** rule principles cancer	companies people **free access** service like **open** pay cable just don government content information **freedom**

		commission communications federal open rules	commissioners telecommunications proposed authority urge service	

Conclusions

Discovering latent topics in a large text corpora, especially in an unsupervised manner (without label), requires significant subjective manual analysis. LDA is useful in uncovering the most likely latent topic structure, but its usefulness is largely dependent on choosing the "right" parameters. In this project we performed automated diagnostics with line and grid search on k and X. We compared the diagnostics from the perspectives of topical alignment, concept retrieval F1 score (a popular IR method), and perplexity (the probabilistic "surprise" factor, a popular NLP method). Using all three diagnostics help us to give a more complete evaluation on how the parameters impact the LDA performance. While there are minor variations in the results, we found that the parameter optimization is generally effective and all three diagnostics pointed to the same range of k={5, 10, 15} and X={0, 30} being optimal. It would be interesting to perform a comparative analysis with datasets with different topic structures or within a supervised/labelled dataset.

Future

There are a few tasks that could be logical next steps to this project:

1. Diagnosing topic/term smoothing parameters, through grid search and topical alignment. Chuang et al.[Chuang 2013] observed that small changes in smoothing parameters can significantly alter the ratio of resolved and fused.
2. Exploring and extending current approaches to a more principled and automated unsupervised way to discern junk topics from legitimate ones, and to rank topic significance using Topic Significance Ranking (TSR)[AlSumait 2009]. In many cases an increase in junk topics is an indication of excess latent topics, so this is very useful information to have.
3. Applying Hierarchical LDA (hLDA), a non-parametric topic model that can select the number of latent topics as part of the model training procedure, and comparing it with our choice of k. Note however that hLDA still requires tuning of its smoothing parameters.

Acknowledgment

The authors would like to thank Dr. Andrew Ng and CS229 teaching staff for the great lectures on Machine Learning, as well as Dr. Dan Jurafsky for the project idea, the discussions, and the feedbacks he gave at various times during the project.

References

[Blei 2003] Blei et al., Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. Vol. 3.

[AlSumait 2009] AlSumait et al., Topic Significance Ranking of LDA Generative Models. Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science Volume 5781, 2009.

[Chuang 2013] Chuang et al., Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. ICML, 2013.

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

[TMT] Stanford Topic Modeling Toolbox: http://www-nlp.stanford.edu/software/tmt/tmt-0.4/

[Sunlight] What can we learn from 800,000 public comments on the FCC's net neutrality plan? 9/2/2014. http://sunlightfoundation.com/blog/2014/09/02/what-can-we-learn-from-800000-public-comments-on-the-fccs-net-neutrality-plan/#data