“LDA on Enron dataset”
Report by Fech Scen Khoo
(9 September 2020)

In this report, we use the LDA model (Latent Dirichlet Allocation) [1] to find an optimal parameter setting for the Enron dataset which will enable us to detect frauds ultimately.

The full Enron dataset in our corpora contains the email folders of 150 employees. There are 465660 documents and about 931190 tokens after preprocessing which involves the removal of stop words and lemmatization. The documents consist of duplicative contents as items such as "sent items" and "inbox" are all included.

There are a number of particular LDA model parameters which we mainly tune in search for an optimal parameter setting, guided by the perplexity in a power-law like behaviour. The parameter *alpha*, α is a prior belief of the topics' probability. The parameter *eta*, η is a prior belief of the word probability. A general presumption is that α and η are smaller than 1. The parameter *update_every* sets the number of documents iterated for each model parameter update, while the parameter *offset* controls how much the first steps are slowed down in the first few iterations.

Through Gensim library, with four CPU cores for LDA parallelization, we set 20 passes over the corpus during training and 1000 iterations through the corpus for inference of the topic distribution. We used the LDA model to study the Enron dataset under the following two settings,

$$\text{setting 1}: \ \alpha = 50/t \ , \eta = (60 \times t)/\ell \ ,$$

$$\text{setting 2}: \ \alpha = 50/t \ , \eta = (40 \times t)/\ell \ ,$$

where t is the number of topics which is unknown a priori and ℓ is the number of tokens after preprocessing. These settings however managed to compute only up to $t = 200$.

In order to learn of an optimal parameter setting in a more effective and less expensive way, we focus hereafter on the emails of the four key individuals known to have involved in frauds that eventually led to the bankruptcy of the company [2]. They are Kenneth Lay, Jeffrey Skilling, John Forney and David Delainey (LSFD).

For all the complete results and plots, including the choices of t, please refer to the Appendix. Results are shown up to 5 decimal places for the training coherence and test coherence, and up to 2 decimal places for the test perplexity.

1 Train and test on LSFD

We first split the shuffled dataset into 90% for training and 10% for testing. In the preprocessing procedure, we form in addition word bigrams and subsequently apply lemmatization. Here, the number of training documents are 12933 and the number of tokens ℓ are 67488. We begin by varying only 2 parameters, α and η.

α	$50/t$	$25/t$	$10/t$
η	$(60 \times t)/\ell$	$(35 \times t)/\ell$	$(20 \times t)/\ell$

We also look into the parameter *chunksize*, which states the number of documents used in each training.

α	$10/t$	$10/t$
η	$(20 \times t)/\ell$	$(20 \times t)/\ell$
chunksize	370	1000

It turns out that the inclusion of a varying *chunksize* does not help in producing an expected power-law like test perplexity. So is the sole inclusion of the parameter *update_every*.

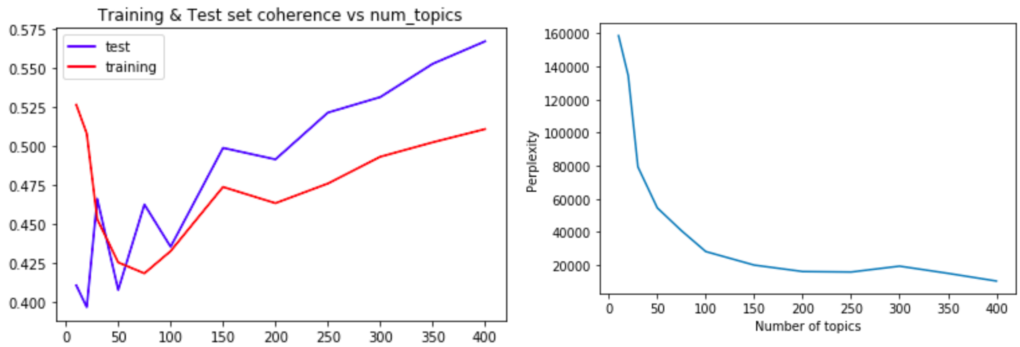
α	$10/t$
η	$(20 \times t)/\ell$
update_every	400

Taking hints from the work in [3], when we include both the parameters *update_every* and *offset*, we begin to observe test perplexity to decrease with the increasing number of topics.

α	$10/t$	$20/t$	$50/t$	$80/t$
η	$(20 \times t)/\ell$	$(10 \times t)/\ell$	$(40 \times t)/\ell$	$(70 \times t)/\ell$
update_every	400	400	400	400
offset	{ 2.4, 4.4, 6.4, 8.4}	4.4	4.4	4.4

Additionally, we have considered in the notation of $\{\alpha, \eta, update\_every, offset\}$:
$\{80/t\ ,\ (70 \times t)/\ell\ ,\ 500\ ,\ \{3, 4, 6.5, 7.3, 8, 10, 12\}\}$, and
$\{80/t\ ,\ (70 \times t)/\ell\ ,\ 700\ ,\ 3\}$.

In setting 11: $\{20/t\ ,\ (10 \times t)/\ell\ ,\ 400\ ,\ 4.4\}$, the test perplexity decreases and plateaus out while the training and test coherences tend to increase indefinitely (fig. 1).



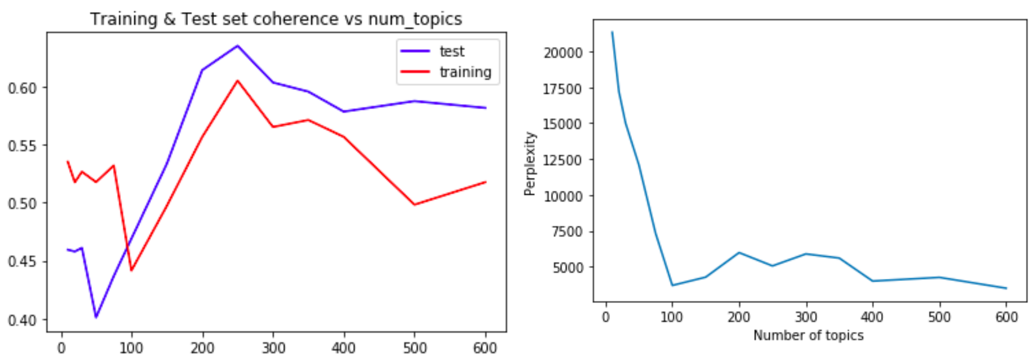
(a) Training and test coherences

(b) Test perplexity

Figure 1: Setting 11: $\{20/t\ ,\ (10 \times t)/\ell\ ,\ 400\ ,\ 4.4\}$

We start to notice a change in setting 13 (see Appendix), in the pattern of the training and test coherences, where they no longer increase linearly, although the test perplexity begins to change wildly.

Unlike perplexity, there is no generic behaviour to look out for in the coherence with respect to the number of topics, although a higher coherence can signal a better interpretability of the results. Intuitively, a climbing coherence does not seem to be indicative. Therefore, when we arrive at the result from setting 18 (fig. 2), we look into the word clouds from the test set. We find that they are not clustered coherently.



(a) Training and test coherences

(b) Test perplexity

Figure 2: Setting 18: $\{80/t\ ,\ (70 \times t)/\ell\ ,\ 500\ ,\ 8\}$

The following setting 21 which includes the tuning of the parameter *decay* saw a drastic peak in the training coherence. *decay* controls how rapid old information is forgotten. This setting nonetheless does not improve inter-

α	$80/t$
η	$(70 \times t)/\ell$
update_every	500
offset	12
decay	0.6

pretability of the test set word clouds.

We observe that the constants we choose in parameters α and η affect the overall perplexity value. In this case, when increased, the perplexity values are lowered (greatly). The *offset* parameter has the role of smoothing out the plateau in the perplexity, when increased.

When the goal is to look for frauds in the texts, especially in official email correspondences, it is a reasonable assumption that the degree of occurrence of frauds related words would be lower. Let us zoom into the lower frequency words. Among the high frequency words are for instance "enron", "please". We consider only words below frequency 2000 under the following scenarios,

setting 23 :
$\alpha = 80/t$, $\eta = (70 \times t)/\ell$, $\mathit{update\_every} = 500$, $\mathit{offset} = 12$, $\mathit{decay} = 0.6$
setting 24 :
no word bigram formation , $\alpha = 80/t$, $\eta = (70 \times t)/\ell$,
$\mathit{update\_every} = 500$, $\mathit{offset} = 12$, $\mathit{decay} = 0.6$
setting 26 :
no word bigram formation , $\alpha = 80/t$, $\eta = (70 \times t)/\ell$,
$\mathit{update\_every} = 500$, $\mathit{offset} = 14$, $\mathit{decay} = 0.6$

and lastly, in addition we ignore words which occur only once under

setting 27 :
no word bigram formation , $\alpha = 80/t$, $\eta = (70 \times t)/\ell$,
$\mathit{update\_every} = 500$, $\mathit{offset} = 14$, $\mathit{decay} = 0.6$.

In general, the word clouds in the test set are not well interpretable, let alone traces of fraud that we can recover. They are realized at the parameters where the test perplexity is low, or when the test or training coherence is high. In fact, in setting 25 (parameters as in setting 24 but with slightly improved parsing of the body contents), at $t = 75$, out of the 20 topics shown, the word probabilities in 18 of them are zero.

2 Coherence of the dataset of Skilling

In this section, we narrow down the investigation to only the emails of a single individual out of LSFD, i.e. Skilling. We will be examining only the coherence of his entire email collections. With much lesser data to analyze, we will see if LDA is able to extract fradulent information here. In this section, we refrain from making word bigrams to feed to the LDA model.

First we remove only the words of frequency 1.

α	$27/t$	$50/t$
η	$(17 \times t)/\ell$	$(40 \times t)/\ell$
update_every	167	167
offset	4	4

Interestingly, these two plots present an inverse behaviour (see Appendix), where in the region of number of topics 50 to 75, the first setting 1 shows a low coherence while in setting 2 it is high.

Next we choose to remove the words of frequency 1 and those of higher than 2000.

α	$50/t$	$70/t$
η	$(40 \times t)/\ell$	$(60 \times t)/\ell$
update_every	167	167
offset	4	4

Thirdly, we remove the words of frequency above 200. This is supported simply by a brief manual inspection that a few scandal relevant words occur below this bound. We can flag words such as “fear”, “angry”, “deceptive”, “jedi”, “partnership”, “account”, “destroy” in the texts [2, 4]. For this approach, we look at a setting $\{70/t\ ,\ (60 \times t)/\ell\ ,\ 167\ ,\ 4\}$.

Taking into account only words of frequency ≥ 5 and ≤ 50, we consider the following settings, for *update_every* $= 50$ and *offset* $= 1.1$,

α	$25/t$	$50/t$	$70/t$	$100/t$	$130/t$
η	$(15 \times t)/\ell$	$(40 \times t)/\ell$	$(60 \times t)/\ell$	$(90 \times t)/\ell$	$(120 \times t)/\ell$

.

The settings that we have also tried at $\alpha = 100/t$, $\eta = (90 \times t)/\ell$ are
- *update_every* $= 50$, *offset* $= \{2.1, 3.1, 5.1\}$
- *offset* $= 2.1$, *update_every* $= \{30, 100, 130, 160, 200, 300, 400\}$
- *offset* $= 2.1$, *update_every* $= 300$, *decay* $= \{0.55, 0.7, 0.9\}$.

When the *decay* parameter is included, the coherence of the dataset drops.

For only words in the range of frequency ≥ 5 and ≤ 30, we consider settings such that $\eta = (90 \times t)/\ell$, *update_every* $= 300$, *offset* $= 2.1$, $\alpha = \{5/t, 10/t, 100/t\}$.

Basically in this section we are seeking for a setting that returns high coherences. The restriction to only a subset of words might have caused a lower word probability we see in the topic distributions. We reach the same conclusion as in the previous section that the word clouds formed are generally not all interpretable and most importantly they do not show corruption information. Despite so, with the parameters in the final setting, at $t = 10$, in one of the clustered topics, the word "ljm" appears. "ljm" was a company created by one of the Enron's employees, Andrew Fastow and used to manipulate Enron. When searched through Skilling's emails, one finds the following short text, which is a positive sign of fraud:

Please note that LJM2 Co-Investment, L.P. ("LJM2") is no longer a "related party" for purposes of disclosure in Enron's proxy and financial statements. Transactions may occur with LJM2 as with any other unrelated third party. Andy Fastow EVP/CFO, ENRON

3 Cross-validation on LSFD

As we are interested in detecting frauds, cross-validation is a fitter route to study the dataset so that we make use of all the available data to train and test. Here we remove the email duplicates and 1099 emails of the same content from different senders which are unrelated to frauds in Enron. Note that in these 1099 emails, there is a consistently occurring word "underhanded". Although the word itself is fraud related, the context in the emails is practically harmless.

We split the data into 90% for training and 10% for final evaluation. Within the training set, we opt for a 15-fold cross-validation, where roughly 93% of it account for training and 7% for testing. In this section, we restore the formation of word bigrams. To choose a better performing setting, we first run the following 3 on our first fold (setting A, B, C), given *update_every* = 200, *offset* = 4,

$$\alpha = 40/t \ , \ \eta = (30 \times t)/\ell$$
$$\alpha = 60/t \ , \ \eta = (50 \times t)/\ell$$
$$\alpha = 80/t \ , \ \eta = (50 \times t)/\ell \ .$$

We decide to also look into a different splitting of dataset, namely training:test in the ratio of 7:3. We choose to make a 10-fold cross-validation, in which about 90% of the initial training set will be used to train and 10% to test. We start with the same settings as before (setting a, b, c). Judging from the resulted higher coherence and lower test perplexity from the second setting (setting b), i.e. $\alpha = 60/t, \eta = (50 \times t)/\ell$, *update_every* = 200, *offset* = 4, in comparison with the above scenario in 9:1 splitting, we choose to proceed with setting b on the remaining 9 folds.

The plots (see Appendix) from all the folds look similar, except in the fold 6, the training and test coherences do not intersect. Further examination in the word clouds of this particular segment of training and test data does not give us compelling results. Finally, we take the average of each the training, test coherences and test perplexity, and look at the word clouds of the initially held-out test set, at number of topics, $t = 50, 75, 300$. As ℓ differs throughout the folds, we take an average of them for the final evaluation. The word clouds are again not interpretable.

We also compute the rate of perplexity change, following the work of [5] which argues that it is a more stable measurement. Using the average perplexity of the folds, the rate of perplexity change is given by

$$rpc = \left| \frac{\text{Ave. perplexity}_i - \text{Ave. perplexity}_{i-1}}{t_i - t_{i-1}} \right| \ , \tag{1}$$

where $i = \{2, ..., 10\}$ in our case. Our plot of the perplexity change rate shares a similar general behaviour as the averaged perplexity. However, they peak at a different number of topics. For the rate of perplexity change, it is between 30 and 50, while for the averaged perplexity, it peaks at 20.

4 Cross-validation on LSFD based on Benford's law

In this section, we try to make use of Benford's law to study words which fall into a certain frequency range. Benford's law is known to be useful in detecting frauds in terms of artificial or crafted numbers. When the data is genuinely true, the numbers should follow a power law. The larger the dataset the more accurate it follows. Nevertheless, let us put the LSFD dataset to test. We filter out the email duplicates and the 1099 non-critical emails, apply no bigrams and use a different lemmatization than before. We group the word frequencies into 1 to 9 according to the first digit of the frequency number of the word.

For the dataset of 12 employees, the bars follow quite strictly the Benford's law (fig. 3). For the LSFD dataset, as we knew that these individuals had committed frauds, the groups of 4 to 9 is an appealing region to investigate (fig. 4). Therefore, with the help of LDA we shall consider only words in this particular region. Similarly, we make a 7:3 split of the data, where 30% of it is used for the final evaluation, and we make a 10-fold cross-validation. For our cross-validation, we allow for word bigrams. We use the setting of $\alpha = 20/t, \eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$.

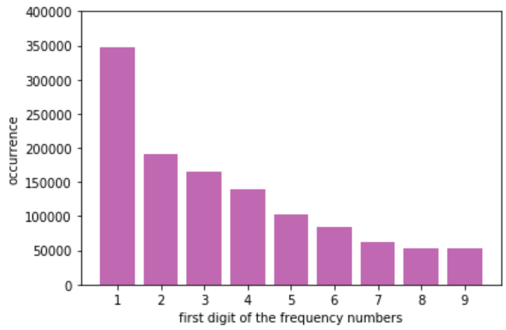


Figure 3: 12 non-fraudulent employees

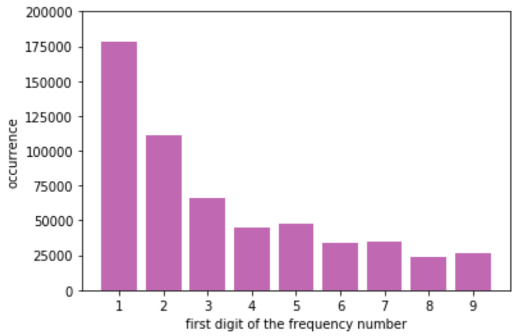


Figure 4: LSFD

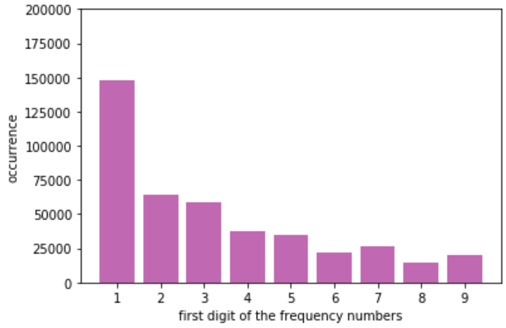


Figure 5: 4 non-fraudulent employees

As in the previous section, we take an average of the results. We also compute the corresponding rate of change in test perplexity. In general, the coherences appear to be quite low here. In the plot of the rate of perplexity change, it shows a plateau at about 100 to 150 number of topics, in contrast to the averaged perplexity. Word clouds of the held-out test set and also of the entire dataset at number of topics, $t = 10, 100, 150, 300$ fail to capture fraudulent information.

5 Summary

We report the effectiveness of LDA model on tackling Enron dataset, with a primary focus on 4 key Enron employees (LSFD) known to have committed the frauds.

We begin with the standard train-and-test approach. There are two important model parameters, *alpha* and *eta* which are associated with the parameter *number of topics*. The test perplexity is not a power-law function of the number of topics until the addition of both parameters *update_every* and *offset*. Neither the parameter *update_every* nor *chunksize* alone together with *alpha* and *eta* can achieve this. Based on the experiments here, the parameters *alpha* and *eta* control the overall test perplexity values, while *offset* can smoothen out small fluctuations in the test perplexity. In this approach, we also examine how LDA performs with words in certain frequency ranges only. In general the parameter *decay* does not play a crucial part in the optimal parameter search. Regardless of the multiple results of a lower test perplexity, the word clouds obtained from the test set are not conclusive.

Next we single out the employee, Skilling's dataset and study under various word frequency windows. The choices of frequencies are approximately driven by a few common words of fraud and some informed knowledge words related to the actual Enron scandals. A higher coherence could point to a higher interpretability of the results in terms of word clouds. With respect to the number of topics, we are inclined to seek for a coherence pattern which is rather high for a few points continuously and falls. One encouraging outcome from this section is the word "ljm" being captured by a word cloud under one of the parameter settings. "ljm" is one of the entities involved in the scandal.

As cross-validation makes use of the data in somewhat different combinations in trainining and testing, it is a better suited approach in an attempt to detect any trace of frauds. We find that a 7:3 splitting of the LSFD data into training and testing eventually gives a significantly lower test perplexity over the 9:1 splitting of data. Subsequently we make a 10-fold cross-validation. The results of the training, test coherences and test perplexity are averaged. We determine a few number of topics from the averaged result and evaluate them on the held-out test set. We also consider the rate of change of test perplexity in extracting the optimal number of topics. The word clouds realized are however not satisfactorily interpretable.

In the final section, we use Benford's law as a structured means to study words of certain frequencies in LSFD dataset. A different lemmatization method is used here. We group the frequencies in 9 according to their first digit. We make a 10-fold cross-validation after a 7:3 data split. Word clouds of the test set fail to achieve our purpose of fraud detection, given the number of topics deduced from the averaged results of training, test coherences and test perplexity, and also from the rate of test perplexity change.

References

[1] David M. Blei, Andrew Y. Ng and Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022.

[2] Dinesh Balaji Sashikanth, Analysis of communication patterns with scammers in Enron corpus, arXiv:1509.00705 [cs.CL], 2015.

[3] Matthew D. Hoffman, David M. Blei and Francis Bach, Online Learning for Latent Dirichlet Allocation, NIPS 2010.

[4] David Noever, The Enron Corpus: Where the Email Bodies are Buried?, arXiv:2001.10374 [cs.IR], 2020.

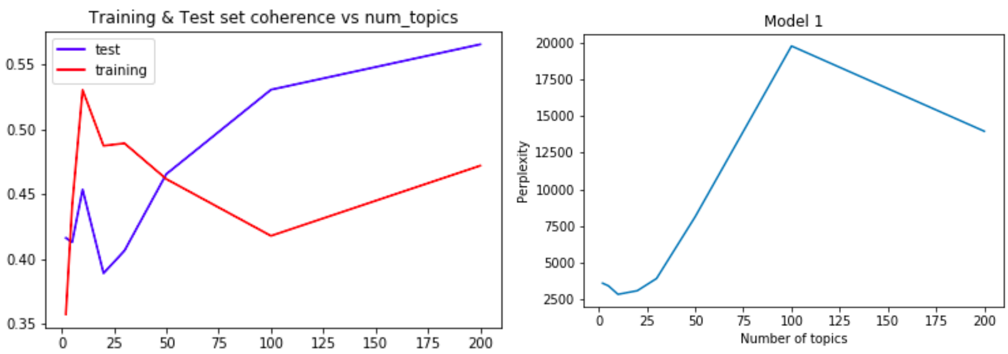
[5] Weizhong Zhao, James J. Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding and Wen Zou, A heuristic approach to determine an appropriate number of topics in topic modeling, *BMC Bioinformatics* **16**, Article number: S8 (2015).

6 Appendix

Train and test on LSFD

Setting 1: $\alpha = 50/t$, $\eta = (60 \times t)/\ell$

t	Training coherence	Test coherence	Test perplexity
2	0.35773	0.41643	3622.64
5	0.44116	0.41310	3448.20
10	0.53026	0.45374	2866.52
20	0.48736	0.38925	3112.39
30	0.48923	0.40663	3942.31
50	0.46176	0.46559	8134.40
100	0.41805	0.53049	19750.25
200	0.47203	0.56539	13954.31

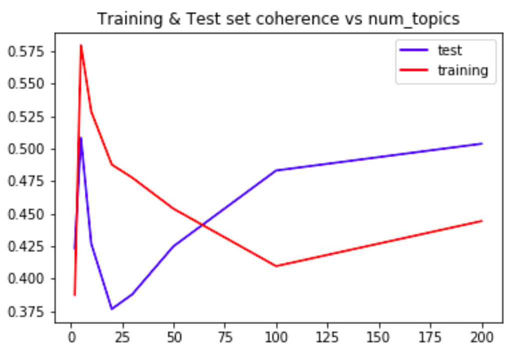


(a) Training and test coherences

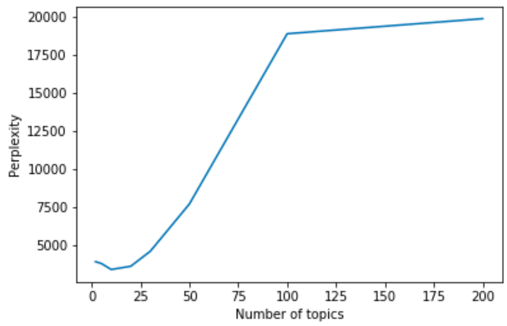
(b) Test perplexity

Setting 2: $\alpha = 25/t$, $\eta = (35 \times t)/\ell$

t	Training coherence	Test coherence	Test perplexity
2	0.38738	0.42318	3966.23
5	0.57968	0.50836	3844.06
10	0.52859	0.42695	3452.97
20	0.48776	0.37650	3664.90
30	0.47763	0.38789	4646.46
50	0.45381	0.42478	7742.11
100	0.40957	0.48318	18885.79
200	0.44428	0.50378	19872.08



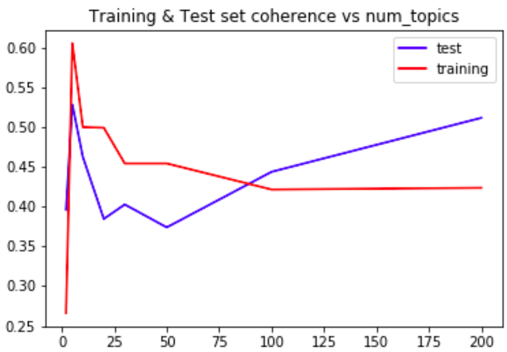
(a) Training and test coherences



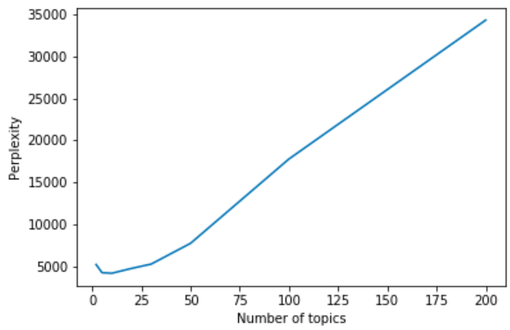
(b) Test perplexity

Setting 3: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$

t	Training coherence	Test coherence	Test perplexity
2	0.26620	0.39611	5187.33
5	0.60579	0.52854	4215.09
10	0.50010	0.46266	4146.88
20	0.49925	0.38458	4720.54
30	0.45434	0.40275	5238.34
50	0.45434	0.37393	7719.02
100	0.42156	0.44383	17783.79
200	0.42367	0.51173	34367.41



(a) Training and test coherences



(b) Test perplexity

Setting 4: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, $chunksize = 370$

t	Training coherence
2	0.61851
5	0.61851
10	0.61851
20	0.61851
30	0.61851
50	0.48035
100	0.44872
200	0.45280

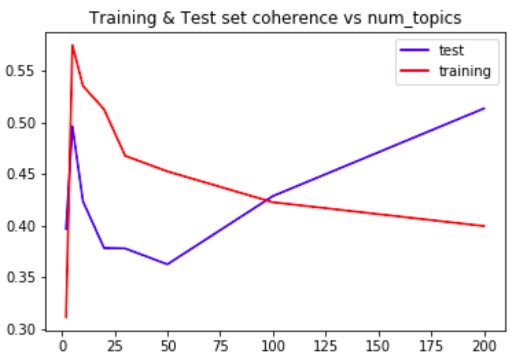
The *log perplexity* in Gensim for $t = 2, 5, 10, 20, 30$ gives “nan”.

Setting 5: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, $chunksize = 1000$

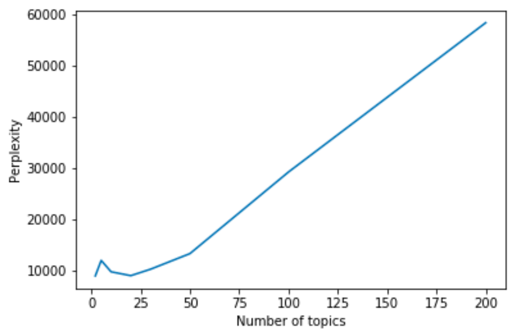
t	Training coherence
2	0.61851
5	0.61851
10	0.61851
20	0.61851
30	0.61851
50	0.45375
100	0.42194
200	0.43683

Setting 6: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, *update_every*= 400

t	Training coherence	Test coherence	Test perplexity
2	0.31130	0.39689	9043.92
5	0.57525	0.49647	12077.04
10	0.53553	0.42358	9857.05
20	0.51286	0.37822	9100.00
30	0.46775	0.37780	10356.59
50	0.45262	0.36244	13405.74
100	0.42264	0.42866	29348.16
200	0.39956	0.51368	58496.38



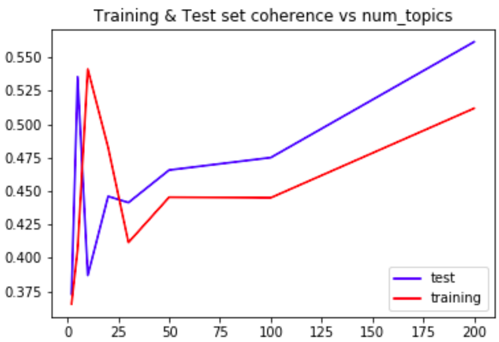
(a) Training and test coherences



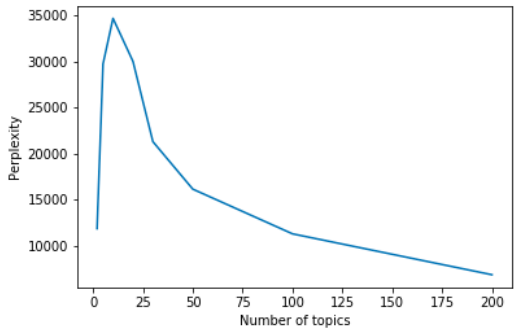
(b) Test perplexity

Setting 7: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, *update_every*= 400 , *offset* = 6.4

t	Training coherence	Test coherence	Test perplexity
2	0.36555	0.37291	11870.83
5	0.40641	0.53551	29744.85
10	0.54126	0.38691	34687.45
20	0.48264	0.44608	30026.53
30	0.41158	0.44140	21310.57
50	0.44542	0.46574	16157.63
100	0.44501	0.47507	11306.70
200	0.51192	0.56166	6867.94



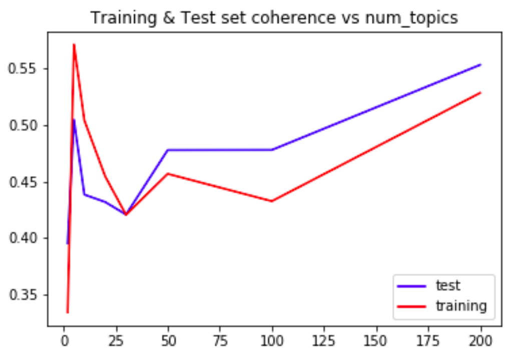
(a) Training and test coherences



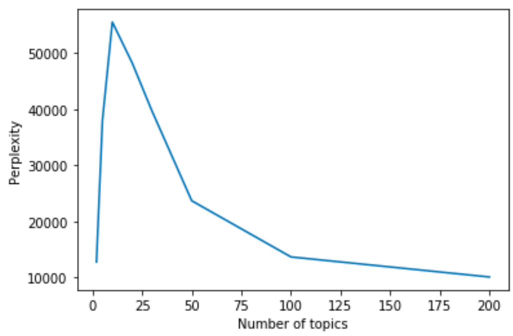
(b) Test perplexity

Setting 8: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, *update_every*= 400 , *offset* = 8.4

t	Training coherence	Test coherence	Test perplexity
2	0.33386	0.39496	12797.81
5	0.57157	0.50473	37959.48
10	0.50417	0.43825	55552.80
20	0.45445	0.43170	48269.54
30	0.42035	0.42055	39681.12
50	0.45670	0.47776	23705.51
100	0.43243	0.47788	13692.17
200	0.52852	0.55342	10118.34



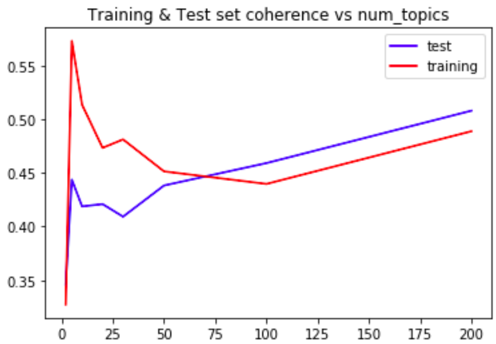
(a) Training and test coherences



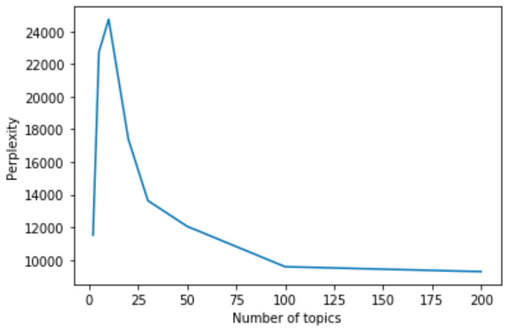
(b) Test perplexity

Setting 9: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, *update_every*= 400 , *offset* = 4.4

t	Training coherence	Test coherence	Test perplexity
2	0.32741	0.34607	11513.12
5	0.57302	0.44377	22731.69
10	0.51352	0.41878	24740.16
20	0.47335	0.42096	17406.72
30	0.48119	0.40926	13633.94
50	0.45142	0.43829	12063.31
100	0.43974	0.45915	9589.04
200	0.48882	0.50794	9290.13



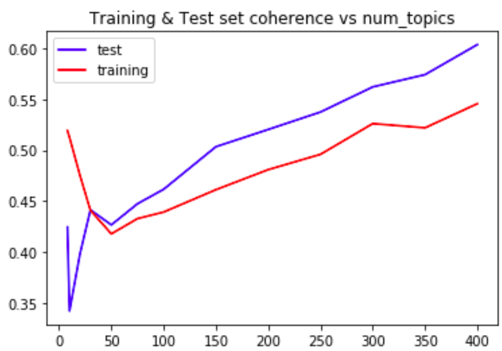
(a) Training and test coherences



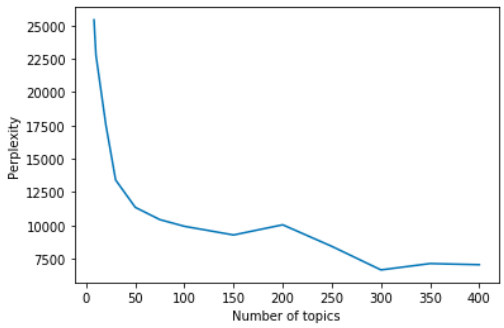
(b) Test perplexity

Setting 9b: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, *update_every*= 400 , *offset* = 4.4

t	Training coherence	Test coherence	Test perplexity
8	0.51918	0.42442	25429.34
10	0.51253	0.34179	22776.26
20	0.47515	0.39799	17582.38
30	0.44078	0.44124	13415.65
50	0.41769	0.42652	11362.94
75	0.43266	0.44727	10441.29
100	0.43911	0.46147	9939.80
150	0.46110	0.50335	9290.23
200	0.48080	0.52030	10053.32
250	0.49585	0.53746	8435.78
300	0.52608	0.56216	6669.51
350	0.52194	0.57421	7146.95
400	0.54563	0.60374	7060.06



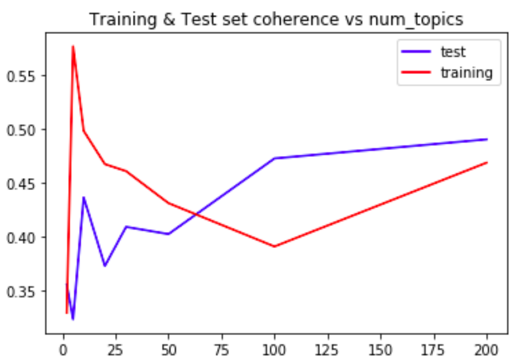
(a) Training and test coherences



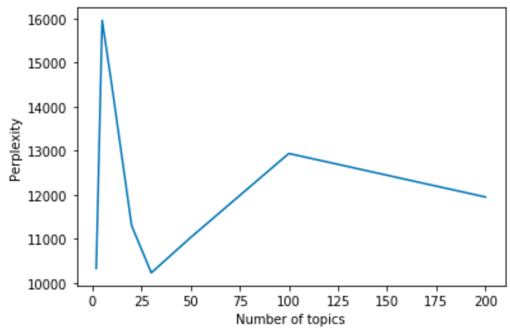
(b) Test perplexity

Setting 10: $\alpha = 10/t$, $\eta = (20 \times t)/\ell$, *update_every*$= 400$, *offset* $= 2.4$

t	Training coherence	Test coherence	Test perplexity
2	0.32889	0.35528	10328.70
5	0.57669	0.32286	15954.23
10	0.49802	0.43617	14451.85
20	0.46714	0.37248	11297.10
30	0.46061	0.40879	10227.70
50	0.43085	0.40207	11027.44
100	0.39054	0.47240	12933.70
200	0.46841	0.49012	11947.04



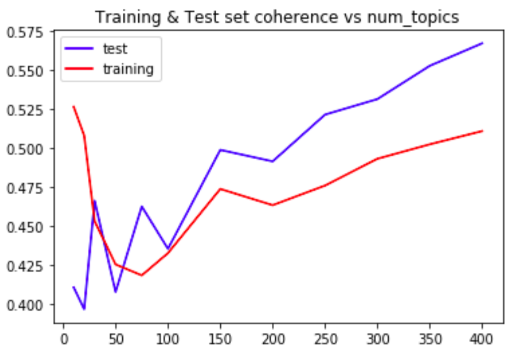
(a) Training and test coherences



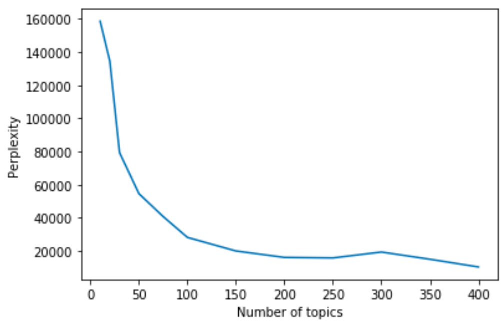
(b) Test perplexity

Setting 11: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, *update_every*= 400 , *offset* = 4.4

t	Training coherence	Test coherence	Test perplexity
10	0.52633	0.41033	158534.94
20	0.50786	0.39635	134444.49
30	0.45287	0.46601	79349.28
50	0.42509	0.40742	54612.55
75	0.41806	0.46232	40825.39
100	0.43234	0.43513	28273.26
150	0.47356	0.49864	20094.06
200	0.46313	0.49126	16212.07
250	0.47574	0.52130	15908.11
300	0.49294	0.53127	19453.52
350	0.50221	0.55264	15112.16
400	0.51072	0.56708	10462.49



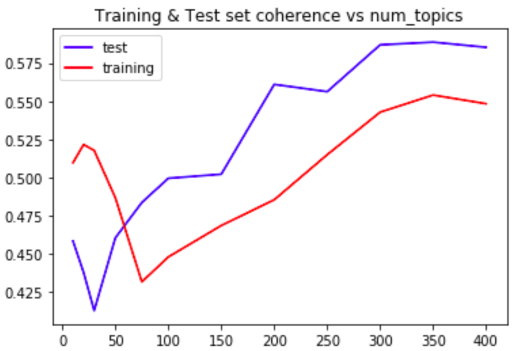
(a) Training and test coherences



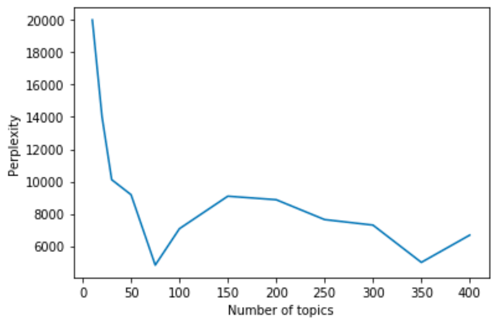
(b) Test perplexity

Setting 12: $\alpha = 50/t$, $\eta = (40 \times t)/\ell$, *update_every*= 400 , *offset* = 4.4

t	Training coherence	Test coherence	Test perplexity
10	0.50971	0.45840	20017.32
20	0.52160	0.43793	14017.02
30	0.51776	0.41287	10113.23
50	0.48677	0.46049	9181.50
75	0.43168	0.48366	4817.34
100	0.44798	0.49951	7073.71
150	0.46858	0.50220	9088.45
200	0.48545	0.56103	8870.59
250	0.51498	0.55628	7638.98
300	0.54285	0.58695	7297.05
350	0.55404	0.58871	4986.60
400	0.54836	0.58533	6676.02



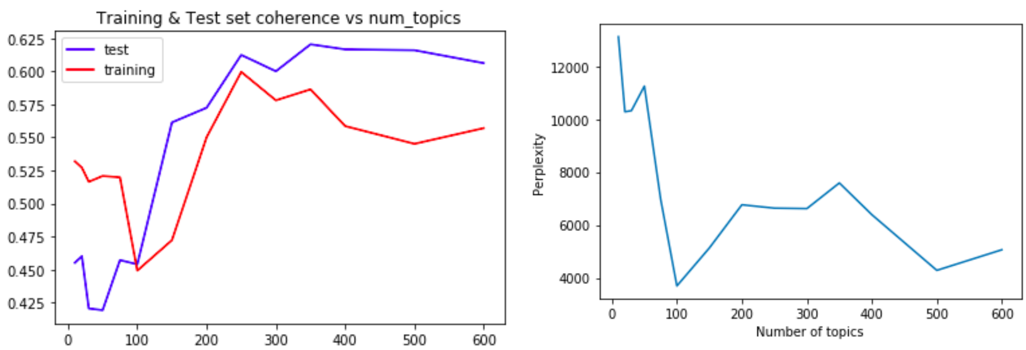
(a) Training and test coherences



(b) Test perplexity

Setting 13: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 400 , *offset* = 4.4

t	Training coherence	Test coherence	Test perplexity
10	0.53196	0.45516	13119.68
20	0.52712	0.46013	10280.57
30	0.51644	0.42060	10321.19
50	0.52085	0.41933	11250.97
75	0.51982	0.45720	6976.40
100	0.44925	0.45401	3695.42
150	0.47227	0.56137	5131.72
200	0.55012	0.57245	6763.55
250	0.59959	0.61243	6635.91
300	0.57807	0.60007	6619.34
350	0.58638	0.62045	7589.25
400	0.55858	0.61664	6389.29
500	0.54517	0.61598	4283.79
600	0.55704	0.60622	5062.42

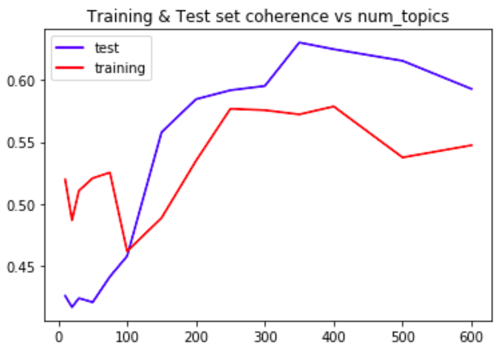


(a) Training and test coherences

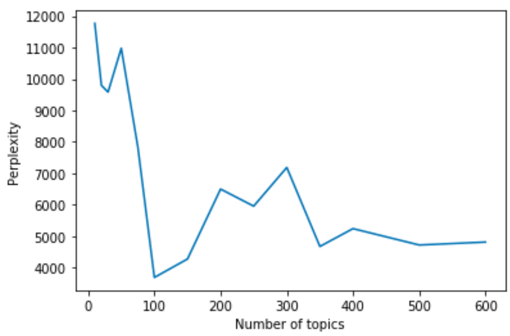
(b) Test perplexity

Setting 14: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 4

t	Training coherence	Test coherence	Test perplexity
10	0.51988	0.42582	11770.39
20	0.48684	0.41661	9797.55
30	0.51073	0.42387	9583.23
50	0.52083	0.42062	10979.36
75	0.52540	0.44127	7840.19
100	0.46177	0.45785	3683.16
150	0.48874	0.55793	4271.13
200	0.53486	0.58462	6496.73
250	0.57693	0.59188	5953.34
300	0.57574	0.59532	7184.66
350	0.57235	0.63042	4669.99
400	0.57877	0.62502	5237.86
500	0.53758	0.61563	4716.76
600	0.54742	0.59294	4809.45



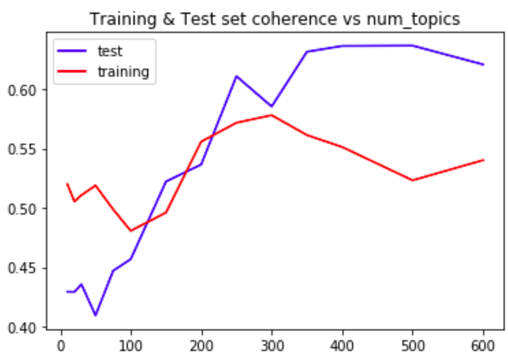
(a) Training and test coherences



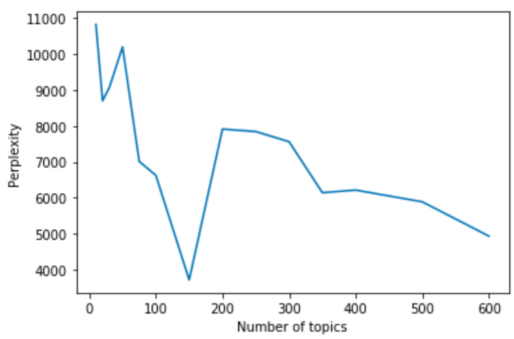
(b) Test perplexity

Setting 15: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 3

t	Training coherence	Test coherence	Test perplexity
10	0.52000	0.42582	10818.95
20	0.50537	0.41661	8699.05
30	0.51059	0.42387	9059.55
50	0.51884	0.42062	10194.63
75	0.49867	0.44127	7011.16
100	0.48057	0.45785	6624.81
150	0.49599	0.55793	3719.03
200	0.55571	0.58462	7911.09
250	0.57166	0.59188	7842.23
300	0.57800	0.59532	7560.76
350	0.56129	0.63042	6142.85
400	0.55122	0.62502	6215.19
500	0.52313	0.61563	5889.50
600	0.54010	0.59294	4935.66



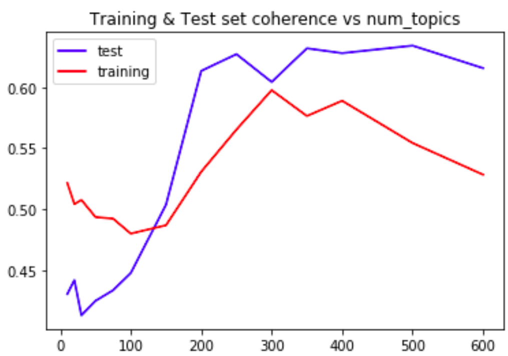
(a) Training and test coherences



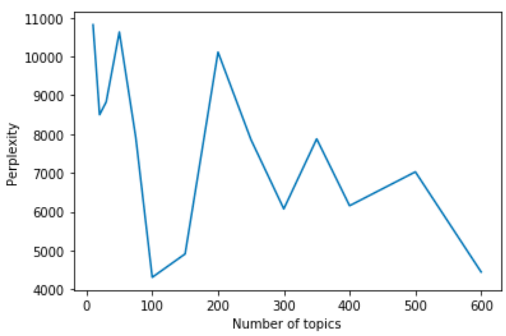
(b) Test perplexity

Setting 16: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 700 , *offset* = 3

t	Training coherence	Test coherence	Test perplexity
10	0.52129	0.43024	10817.93
20	0.50413	0.44165	8500.62
30	0.50746	0.41278	8829.05
50	0.49344	0.42468	10630.56
75	0.49210	0.43339	7911.50
100	0.47986	0.44749	4308.62
150	0.48670	0.50352	4912.52
200	0.53063	0.61336	10112.23
250	0.56519	0.62712	7861.33
300	0.59755	0.60429	6072.79
350	0.57637	0.63190	7876.39
400	0.58887	0.62793	6154.19
500	0.55420	0.63412	7027.03
600	0.52823	0.61566	4445.77



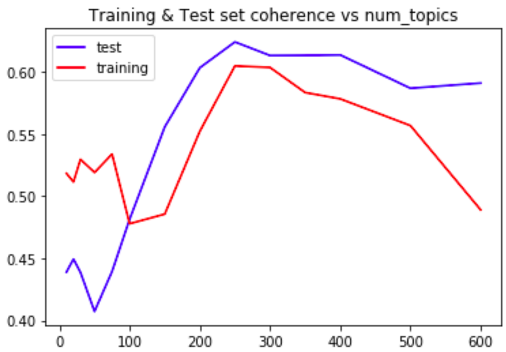
(a) Training and test coherences



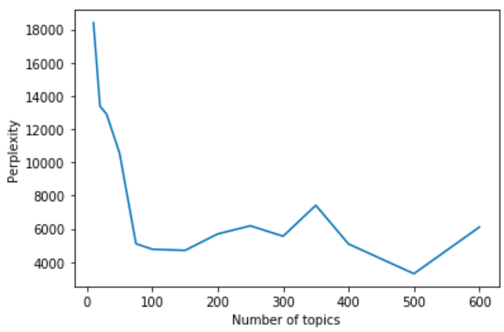
(b) Test perplexity

Setting 17: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*$=$ 500 , *offset* $= 6.5$

t	Training coherence	Test coherence	Test perplexity
10	0.51838	0.43886	18402.19
20	0.51142	0.44934	13374.08
30	0.52964	0.43860	12903.85
50	0.51907	0.40722	10541.61
75	0.53387	0.43944	5109.60
100	0.47781	0.48139	4776.61
150	0.48555	0.55570	4706.65
200	0.55223	0.60337	5700.11
250	0.60484	0.62414	6184.89
300	0.60361	0.61322	5559.59
350	0.58342	0.61336	7419.76
400	0.57833	0.61369	5093.67
500	0.55679	0.58685	3306.94
600	0.48898	0.59105	6106.30



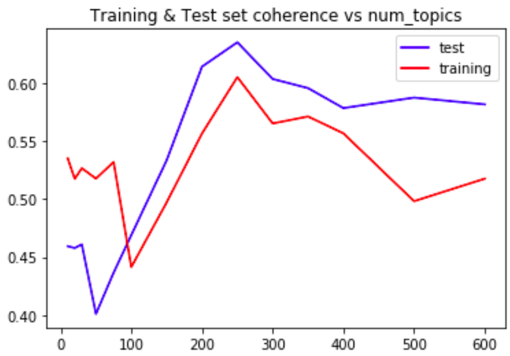
(a) Training and test coherences



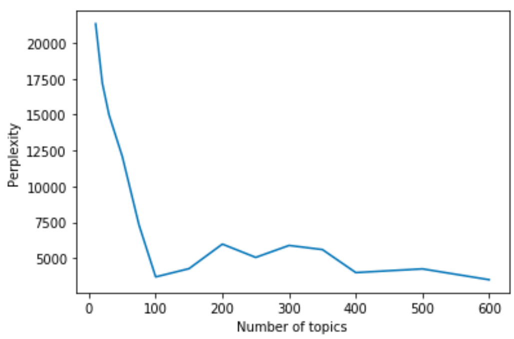
(b) Test perplexity

Setting 18: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, $update\_every = 500$, $offset = 8$

t	Training coherence	Test coherence	Test perplexity
10	0.53545	0.45942	21337.40
20	0.51765	0.45784	17204.84
30	0.52677	0.46099	14999.11
50	0.51779	0.40094	12080.06
75	0.53217	0.43657	7311.73
100	0.44141	0.46899	3688.69
150	0.49723	0.53362	4264.85
200	0.55676	0.61447	5975.21
250	0.60544	0.63550	5048.12
300	0.56540	0.60380	5882.48
350	0.57143	0.59594	5594.28
400	0.55689	0.57869	3990.81
500	0.49829	0.58774	4247.44
600	0.51772	0.58189	3492.45



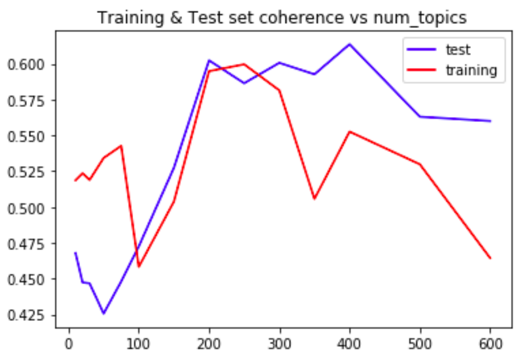
(a) Training and test coherences



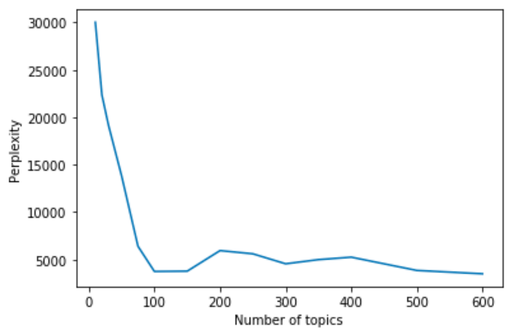
(b) Test perplexity

Setting 19: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 10

t	Training coherence	Test coherence	Test perplexity
10	0.51869	0.46783	30024.60
20	0.52356	0.44735	22359.62
30	0.51889	0.44671	19173.42
50	0.53423	0.42560	13828.12
75	0.54272	0.44805	6393.84
100	0.45822	0.47249	3741.45
150	0.50380	0.52738	3769.79
200	0.59485	0.60239	5940.03
250	0.59964	0.58644	5602.09
300	0.58145	0.60067	4549.25
350	0.50570	0.59265	4987.32
400	0.55263	0.61368	5247.37
500	0.52983	0.56304	3847.68
600	0.46439	0.56005	3493.60



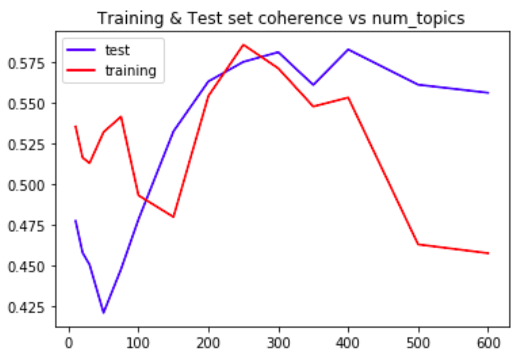
(a) Training and test coherences



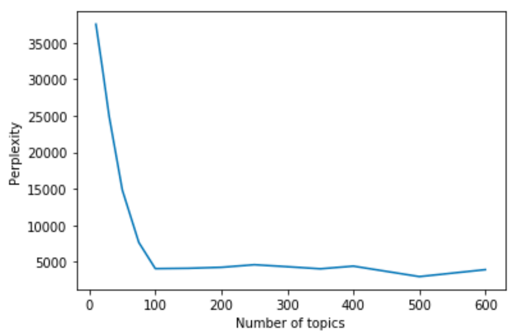
(b) Test perplexity

Setting 20: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 12

t	Training coherence	Test coherence	Test perplexity
10	0.53565	0.47760	37525.41
20	0.51655	0.45803	31380.53
30	0.51314	0.45066	24927.67
50	0.53223	0.42105	14830.47
75	0.54178	0.44787	7653.88
100	0.49330	0.47857	4074.14
150	0.48003	0.53268	4122.00
200	0.55448	0.56345	4255.18
250	0.58602	0.57548	4610.39
300	0.57149	0.58144	4338.63
350	0.54800	0.56128	4050.31
400	0.55350	0.58311	4417.49
500	0.463131	0.56142	2984.69
600	0.45773	0.55646	3934.84



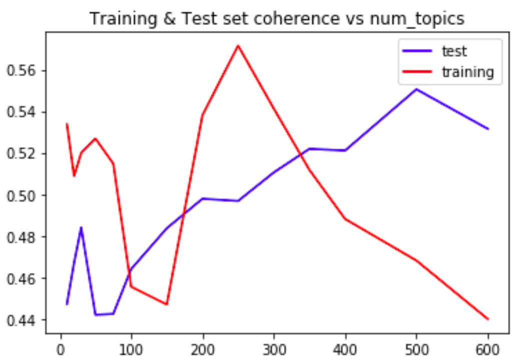
(a) Training and test coherences



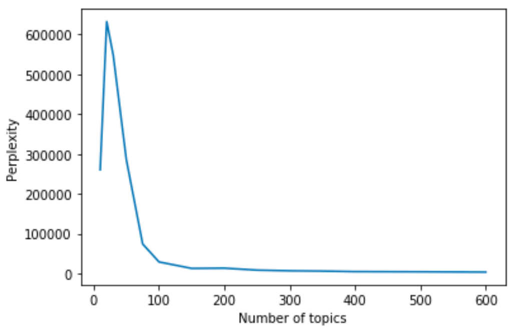
(b) Test perplexity

Setting 21: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 12 , *decay* = 0.6

t	Training coherence	Test coherence	Test perplexity
10	0.53366	0.44716	260343.24
20	0.50874	0.46679	632537.08
30	0.51986	0.48399	547871.09
50	0.52679	0.44191	285997.25
75	0.51472	0.44235	73999.47
100	0.45546	0.46389	28906.25
150	0.44687	0.48357	12643.14
200	0.53803	0.49779	13267.95
250	0.57155	0.49673	8337.14
300	0.54123	0.51046	6643.59
350	0.51169	0.52183	5912.39
400	0.48807	0.52101	4652.17
500	0.46805	0.55050	4176.42
600	0.43989	0.53147	3411.37



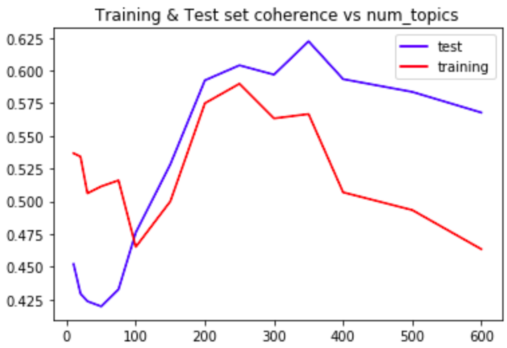
(a) Training and test coherences



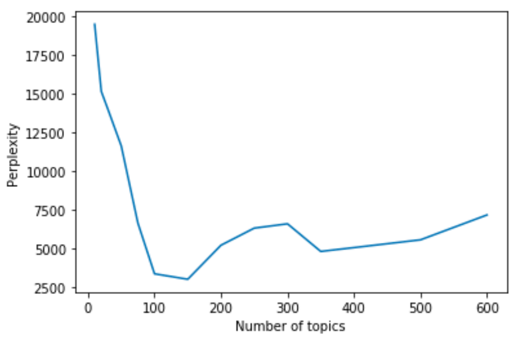
(b) Test perplexity

Setting 22: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 7.3

t	Training coherence	Test coherence	Test perplexity
10	0.53677	0.45210	19513.20
20	0.53425	0.42950	15171.29
30	0.50615	0.42363	14011.44
50	0.51143	0.41967	11652.10
75	0.51603	0.43266	6668.75
100	0.46518	0.47614	3387.97
150	0.50000	0.52833	3035.26
200	0.57499	0.59253	5235.23
250	0.58992	0.60400	6339.48
300	0.56342	0.59687	6618.41
350	0.56675	0.62239	4833.29
400	0.50690	0.59343	5081.07
500	0.49336	0.58368	5583.04
600	0.46343	0.56786	7190.70



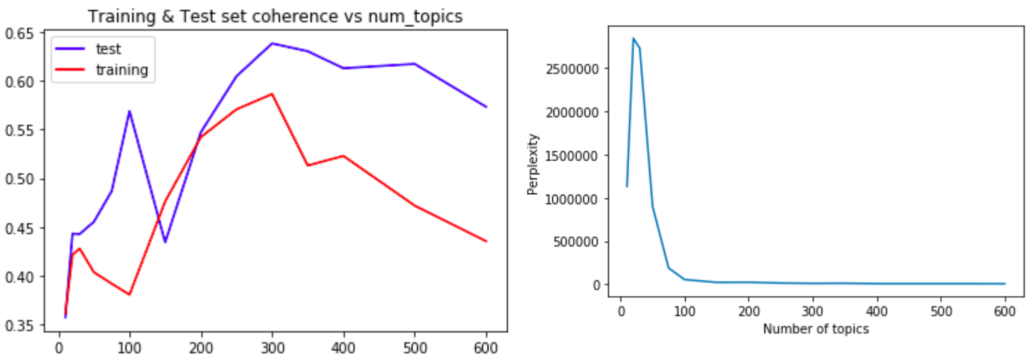
(a) Training and test coherences



(b) Test perplexity

Setting 23: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 12 , *decay* = 0.6

t	Training coherence	Test coherence	Test perplexity
10	0.36170	0.35771	1130820.57
20	0.42155	0.44334	2843078.30
30	0.42806	0.44284	2726412.20
50	0.40380	0.45543	901125.17
75	0.39202	0.48687	188065.69
100	0.38076	0.56884	54298.99
150	0.47629	0.43445	22611.10
200	0.54218	0.54712	23440.74
250	0.57063	0.60445	14243.95
300	0.58627	0.63815	10232.92
350	0.51309	0.63014	11855.07
400	0.52290	0.61272	7274.72
500	0.47200	0.61715	7230.50
600	0.43559	0.57324	6426.51

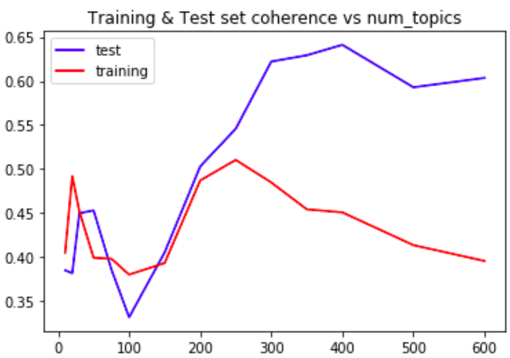


(a) Training and test coherences

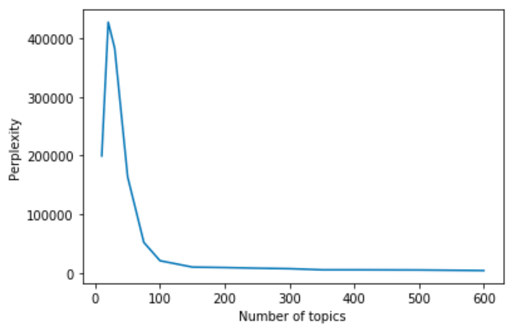
(b) Test perplexity

Setting 24: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 12 , *decay* = 0.6

t	Training coherence	Test coherence	Test perplexity
10	0.40485	0.38466	199260.61
20	0.49184	0.38140	427554.97
30	0.45174	0.44957	382971.55
50	0.39891	0.45284	163562.29
75	0.39787	0.38583	51866.31
100	0.37980	0.33126	20604.97
150	0.39296	0.40506	9762.46
200	0.48687	0.50278	8965.65
250	0.51025	0.54568	7896.48
300	0.48461	0.62214	7041.33
350	0.45403	0.62928	5101.60
400	0.45057	0.64119	5085.89
500	0.41329	0.59287	4816.97
600	0.39530	0.60352	3820.74



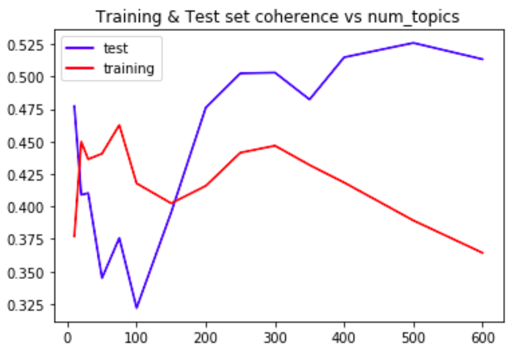
(a) Training and test coherences



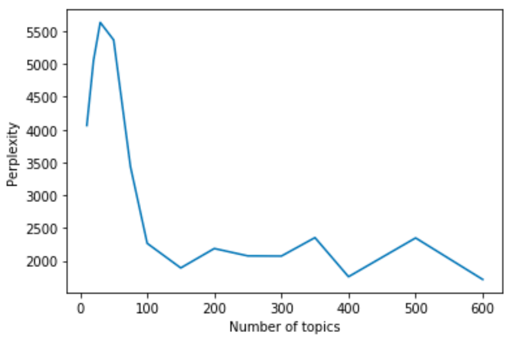
(b) Test perplexity

Setting 25: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*$= 500$, *offset* $= 12$, *decay* $= 0.6$

t	Training coherence	Test coherence	Test perplexity
10	0.37706	0.47727	4062.45
20	0.44995	0.40899	5054.18
30	0.43645	0.41036	5633.44
50	0.44072	0.34490	5366.15
75	0.46278	0.37580	3438.59
100	0.41784	0.32188	2267.61
150	0.40250	0.39554	1893.23
200	0.41589	0.47613	2188.80
250	0.44147	0.50255	2076.50
300	0.44685	0.50314	2073.97
350	0.43198	0.48244	2354.81
400	0.41849	0.51492	1759.26
500	0.38930	0.52601	2350.30
600	0.36435	0.51344	1717.40



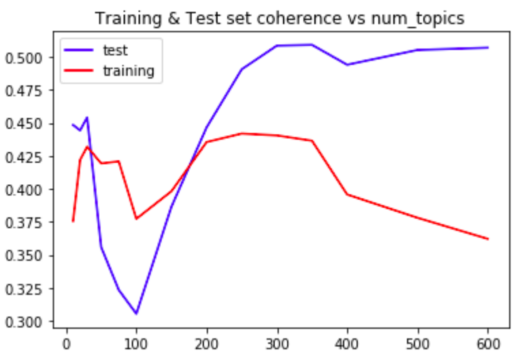
(a) Training and test coherences



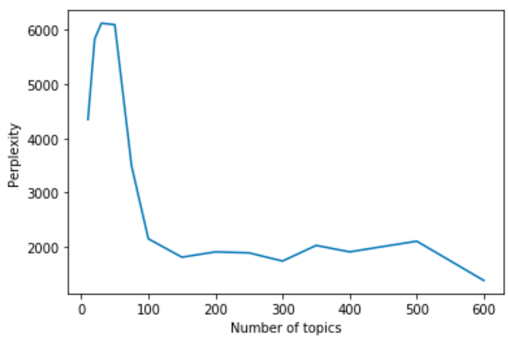
(b) Test perplexity

Setting 26: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 14 , *decay* = 0.6

t	Training coherence	Test coherence	Test perplexity
10	0.37552	0.44844	4349.29
20	0.42197	0.44428	5834.30
30	0.43196	0.45414	6125.79
50	0.41924	0.35576	6099.62
75	0.42079	0.32332	3491.29
100	0.37724	0.30524	2148.00
150	0.39816	0.38647	1808.27
200	0.43544	0.44658	1906.04
250	0.44189	0.49064	1887.92
300	0.44048	0.50857	1736.81
350	0.43645	0.50917	2025.28
400	0.39569	0.49408	1906.90
500	0.37805	0.50527	2103.32
600	0.36206	0.50700	1377.81



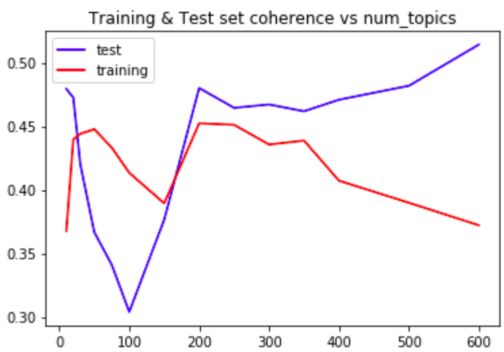
(a) Training and test coherences



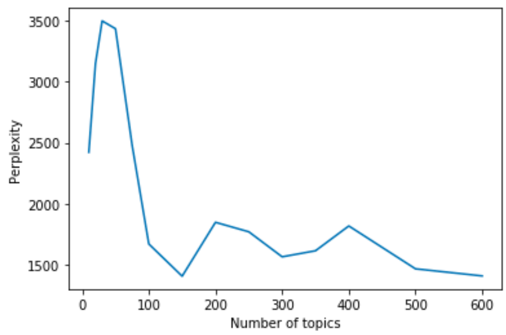
(b) Test perplexity

Setting 27: $\alpha = 80/t$, $\eta = (70 \times t)/\ell$, *update_every*= 500 , *offset* = 14 , *decay* = 0.6

t	Training coherence	Test coherence	Test perplexity
10	0.36770	0.47938	2420.98
20	0.43972	0.47246	3149.50
30	0.44404	0.42010	3497.56
50	0.44765	0.36666	3433.31
75	0.43318	0.34096	2477.51
100	0.41337	0.30392	1669.86
150	0.38936	0.37651	1405.93
200	0.45231	0.48007	1847.45
250	0.45112	0.46446	1769.61
300	0.43562	0.46713	1564.68
350	0.43873	0.46177	1614.25
400	0.40717	0.47085	1817.39
500	0.38981	0.48182	1466.21
600	0.37209	0.51424	1408.21



(a) Training and test coherences

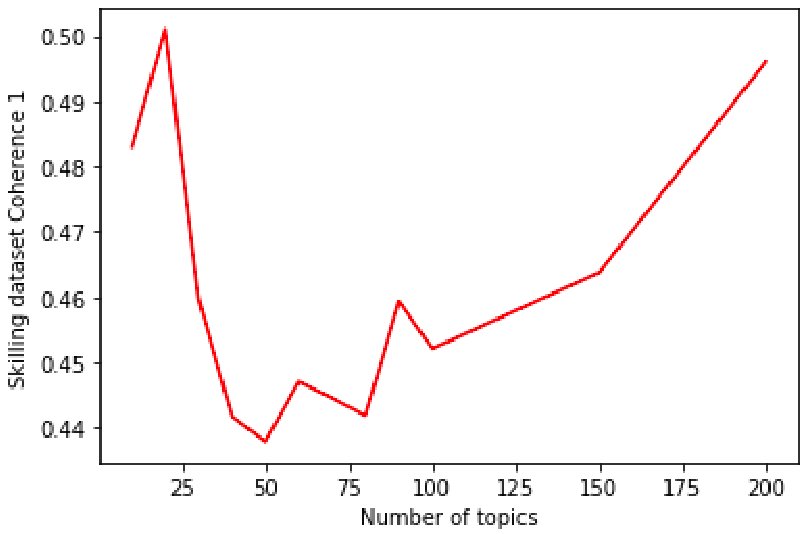


(b) Test perplexity

Coherence of the dataset of Skilling

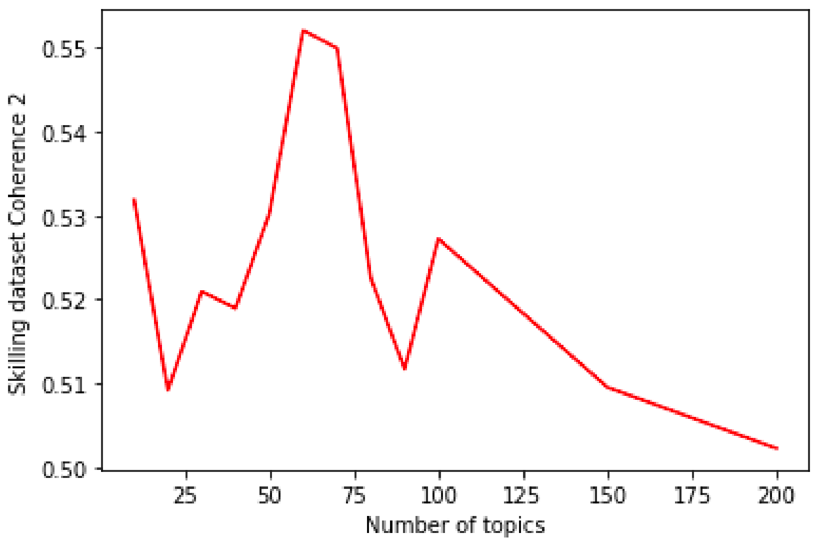
Setting 1: $\alpha = 27/t$, $\eta = (17 \times t)/\ell$, $update\_every = 167$, $offset = 4$

t	Coherence
10	0.48301
20	0.50105
30	0.45979
40	0.44173
50	0.43792
60	0.44710
70	0.44456
80	0.44188
90	0.45938
100	0.45213
150	0.46381
200	0.49607



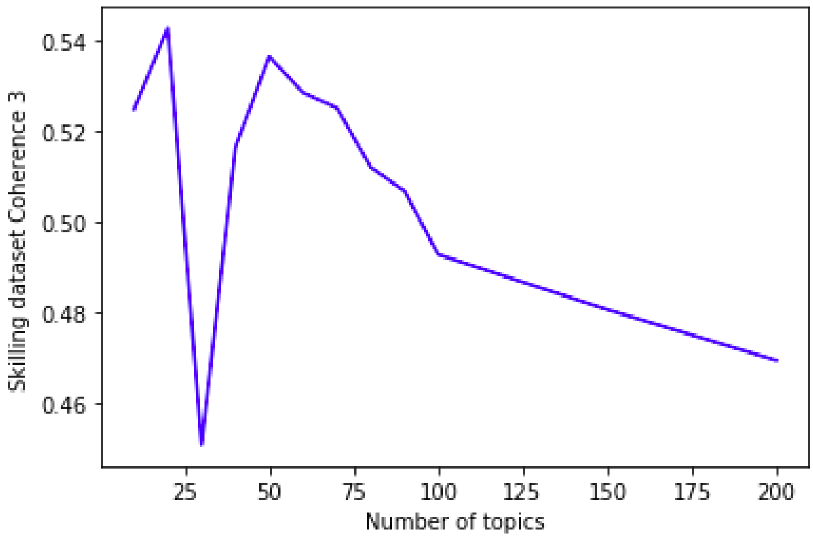
Setting 2: $\alpha = 50/t$, $\eta = (40 \times t)/\ell$, $update\_every = 167$, $offset = 4$

t	Coherence
10	0.53187
20	0.50916
30	0.52094
40	0.51893
50	0.53023
60	0.55205
70	0.54998
80	0.52255
90	0.51168
100	0.52721
150	0.50952
200	0.50223



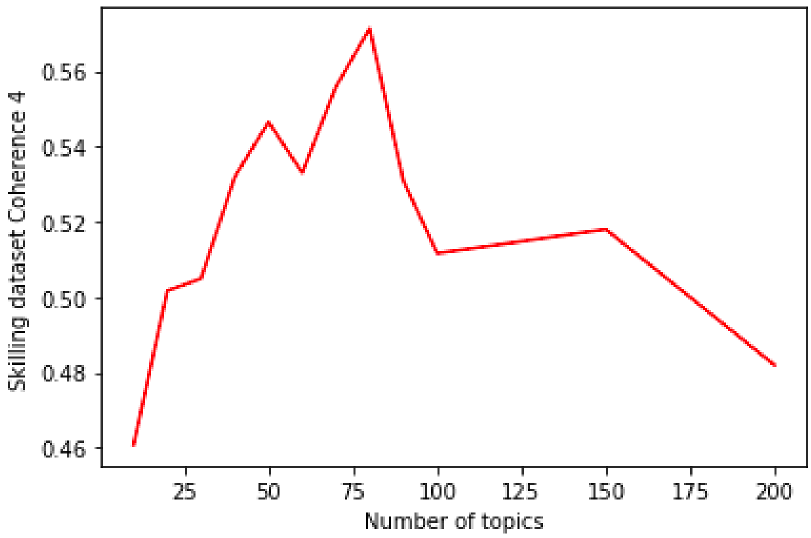
Setting 3: $\alpha = 50/t$, $\eta = (40 \times t)/\ell$, $update\_every = 167$, $offset = 4$

t	Coherence
10	0.52481
20	0.54269
30	0.45097
40	0.51647
50	0.53645
60	0.52847
70	0.52514
80	0.51208
90	0.50684
100	0.49286
150	0.48075
200	0.46959



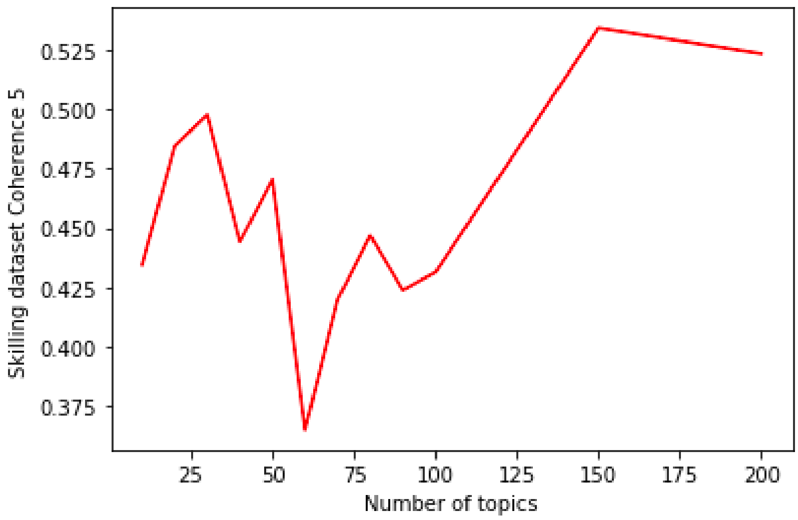
Setting 4: $\alpha = 70/t$, $\eta = (60 \times t)/\ell$, $update\_every = 167$, $offset = 4$

t	Coherence
10	0.46075
20	0.50167
30	0.50492
40	0.53196
50	0.54649
60	0.53303
70	0.55589
80	0.57139
90	0.53087
100	0.51161
150	0.51797
200	0.48193



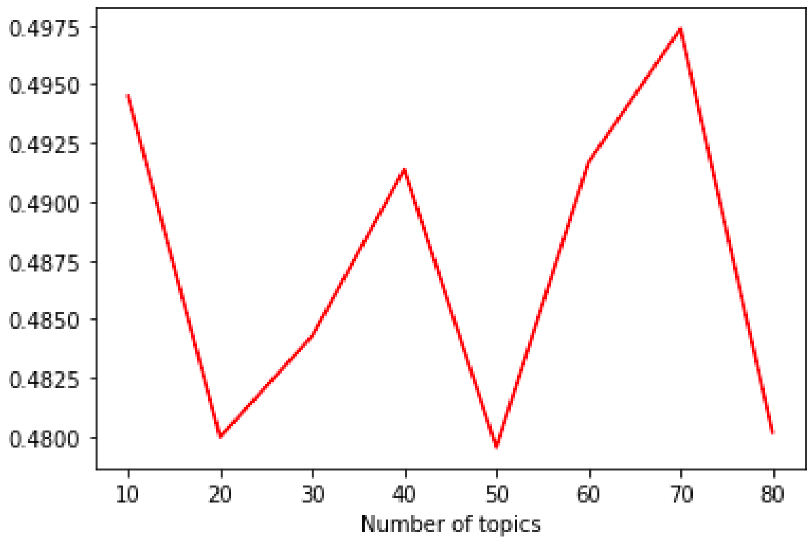
Setting 5: $\alpha = 70/t$, $\eta = (60 \times t)/\ell$, $update\_every = 167$, $offset = 4$

t	Coherence
10	0.43452
20	0.48444
30	0.49769
40	0.44417
50	0.47042
60	0.36513
70	0.41996
80	0.44685
90	0.42389
100	0.43136
150	0.53414
200	0.52336



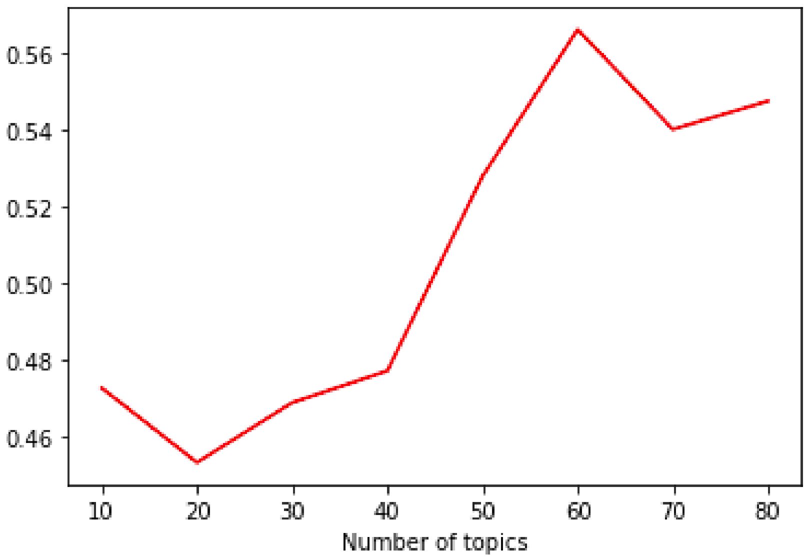
Setting 6: $\alpha = 25/t$, $\eta = (15 \times t)/\ell$, $update\_every = 50$, $offset = 1.1$

t	Coherence
10	0.49449
20	0.47998
30	0.48428
40	0.49138
50	0.47955
60	0.49166
70	0.49736
80	0.48019



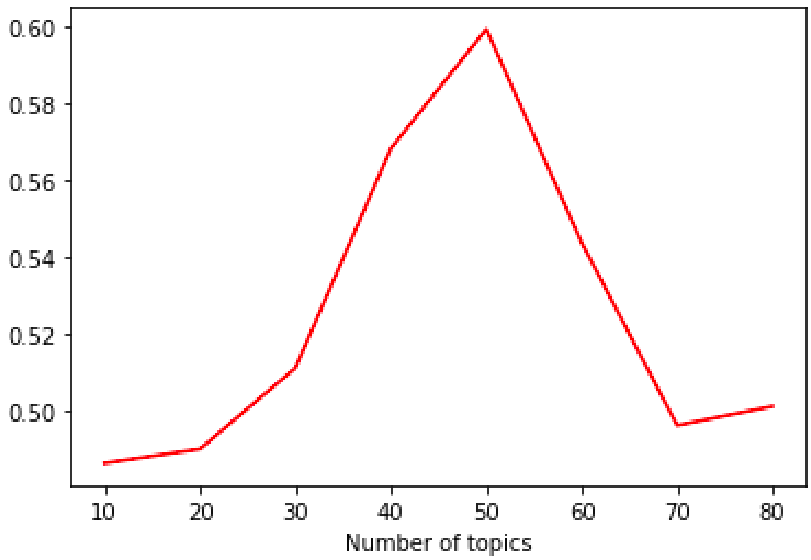
Setting 7: $\alpha = 50/t$, $\eta = (40 \times t)/\ell$, $update\_every = 50$, $offset = 1.1$

t	Coherence
10	0.47270
20	0.45325
30	0.46889
40	0.47719
50	0.52797
60	0.56635
70	0.54026
80	0.54764



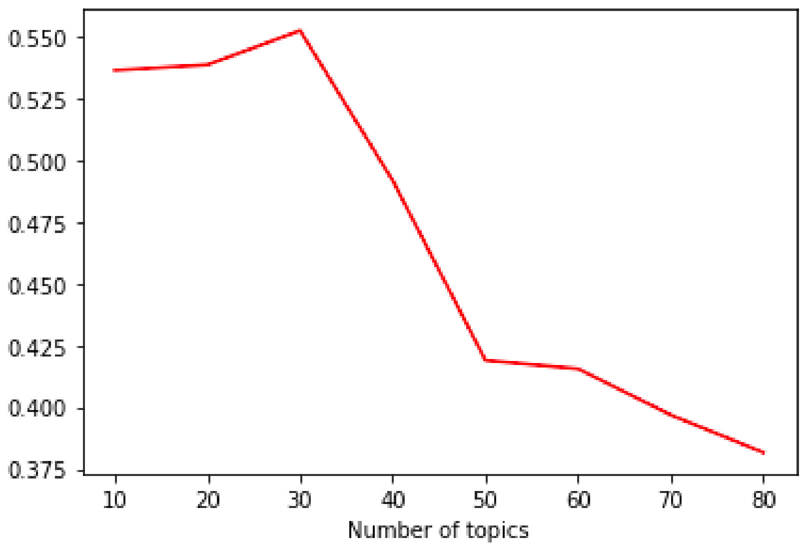
Setting 8: $\alpha = 70/t$, $\eta = (60 \times t)/\ell$, $update\_every = 50$, $offset = 1.1$

t	Coherence
10	0.48645
20	0.49015
30	0.51127
40	0.56841
50	0.59948
60	0.54384
70	0.49627
80	0.50126



Setting 9: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 50$, $offset = 1.1$

t	Coherence
10	0.53625
20	0.53850
30	0.55235
40	0.49163
50	0.41894
60	0.41548
70	0.39681
80	0.38161



Setting 10: $\alpha = 130/t$, $\eta = (120 \times t)/\ell$, $update\_every = 50$, $offset = 1.1$

t	Coherence
10	0.46573
20	0.51195
30	0.45830
40	0.39224
50	0.36215
60	0.35696
70	0.37008
80	0.36320

Setting 11: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 50$, $offset = 2.1$

t	Coherence
10	0.45222
20	0.52127
30	0.51192
40	0.45287
50	0.42070
60	0.39525
70	0.38493
80	0.36041

Setting 12: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 50$, $offset = 3.1$

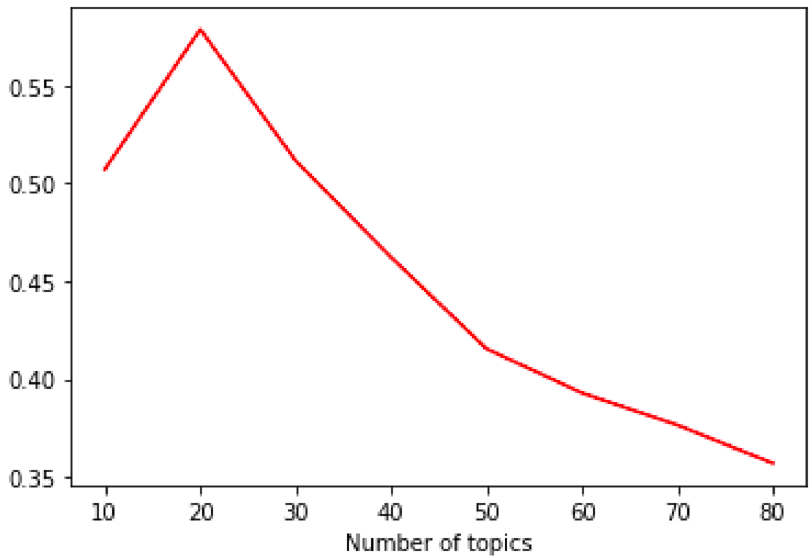
t	Coherence
10	0.52549
20	0.53941
30	0.50957
40	0.43475
50	0.40403
60	0.39284
70	0.36457
80	0.35727

Setting 13: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 50$, $offset = 5.1$

t	Coherence
10	0.44397
20	0.42091
30	0.47393
40	0.40746
50	0.37700
60	0.38104
70	0.35443
80	0.35859

Setting 14: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 100$, $offset = 2.1$

t	Coherence
10	0.50722
20	0.57874
30	0.51153
40	0.46216
50	0.41541
60	0.39283
70	0.37645
80	0.35688



Setting 15: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 30$, $offset = 2.1$

t	Coherence
10	0.44501
20	0.55949
30	0.51854
40	0.46200
50	0.39076
60	0.40405
70	0.38962
80	0.36306

Setting 16: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 130$, $offset = 2.1$

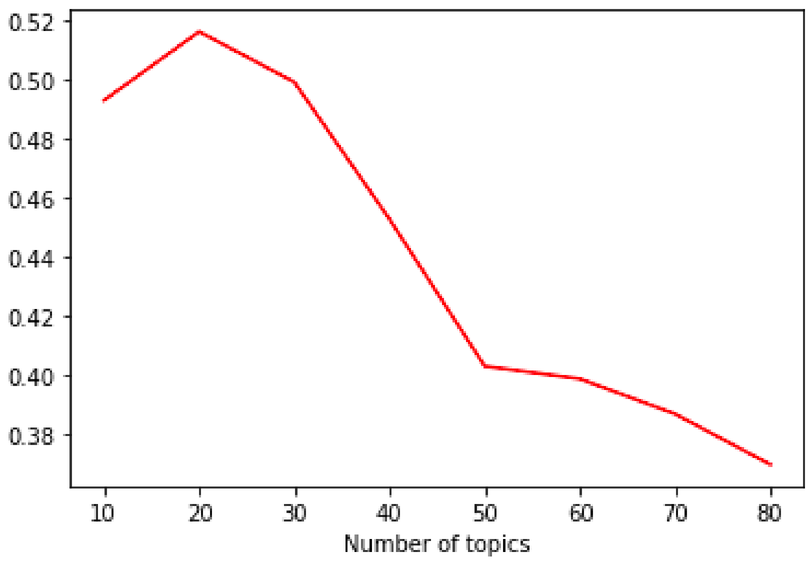
t	Coherence
10	0.42750
20	0.52571
30	0.48581
40	0.44543
50	0.41237
60	0.38752
70	0.39545
80	0.37754

Setting 17: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 160$, $offset = 2.1$

t	Coherence
10	0.50636
20	0.51023
30	0.48845
40	0.45256
50	0.41778
60	0.39388
70	0.38980
80	0.37926

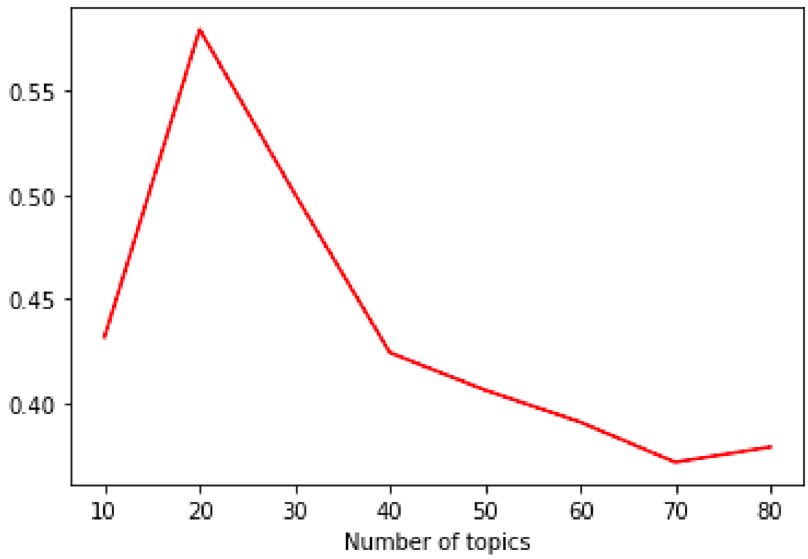
Setting 18: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 200$, $offset = 2.1$

t	Coherence
10	0.49306
20	0.51635
30	0.49923
40	0.45274
50	0.40289
60	0.39859
70	0.38669
80	0.36955



Setting 19: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$

t	Coherence
10	0.43173
20	0.57970
30	0.50061
40	0.42420
50	0.40618
60	0.39082
70	0.37145
80	0.37881



Setting 20: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 400$, $offset = 2.1$

t	Coherence
10	0.48369
20	0.56555
30	0.50152
40	0.41862
50	0.41955
60	0.39967
70	0.38175
80	0.37469

Setting 21: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$, $decay = 0.7$

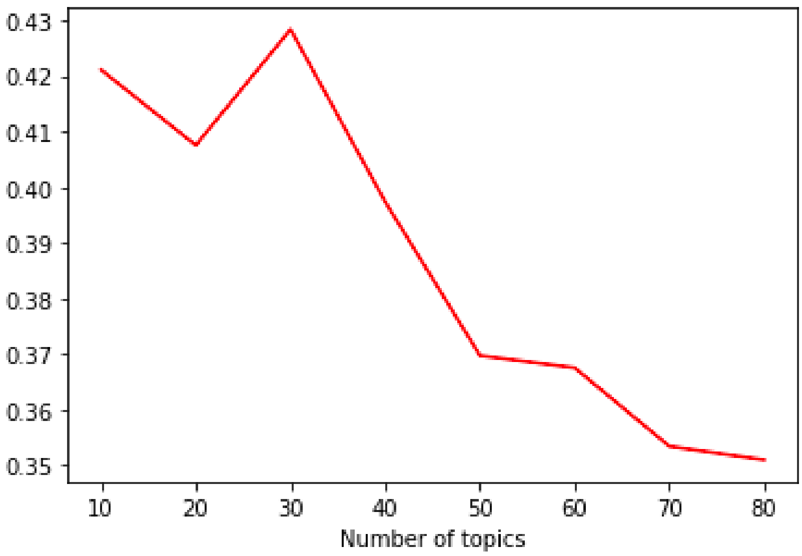
t	Coherence
10	0.44812
20	0.41690
30	0.45799
40	0.41230
50	0.39806
60	0.36257
70	0.37175
80	0.34757

Setting 22: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$, $decay = 0.55$

t	Coherence
10	0.39723
20	0.48772
30	0.50094
40	0.41789
50	0.41551
60	0.38644
70	0.38433
80	0.35503

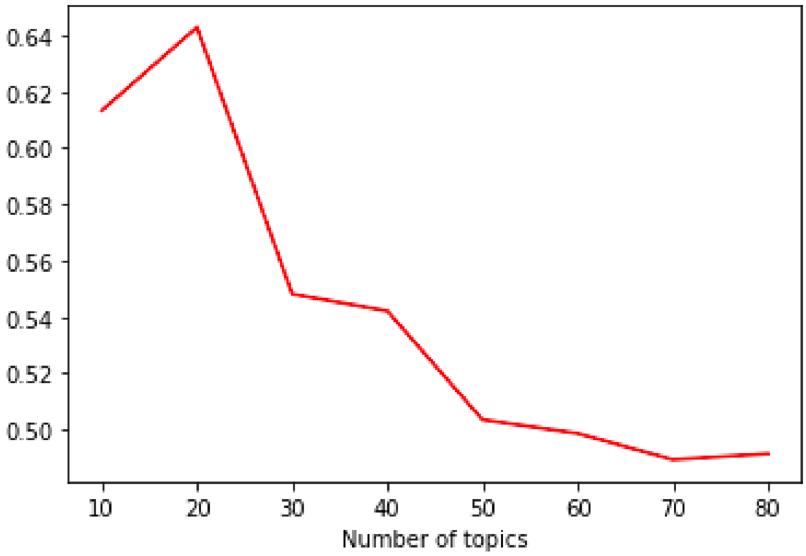
Setting 23: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$, $decay = 0.9$

t	Coherence
10	0.42112
20	0.40753
30	0.42839
40	0.39735
50	0.36969
60	0.36748
70	0.35335
80	0.35092



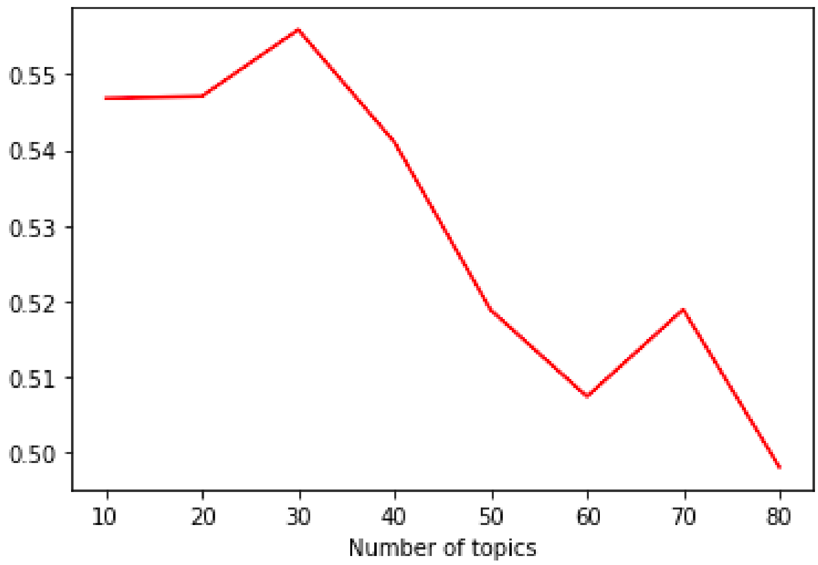
Setting 24: $\alpha = 100/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$

t	Coherence
10	0.61330
20	0.64288
30	0.54818
40	0.54216
50	0.50354
60	0.49870
70	0.48936
80	0.49145



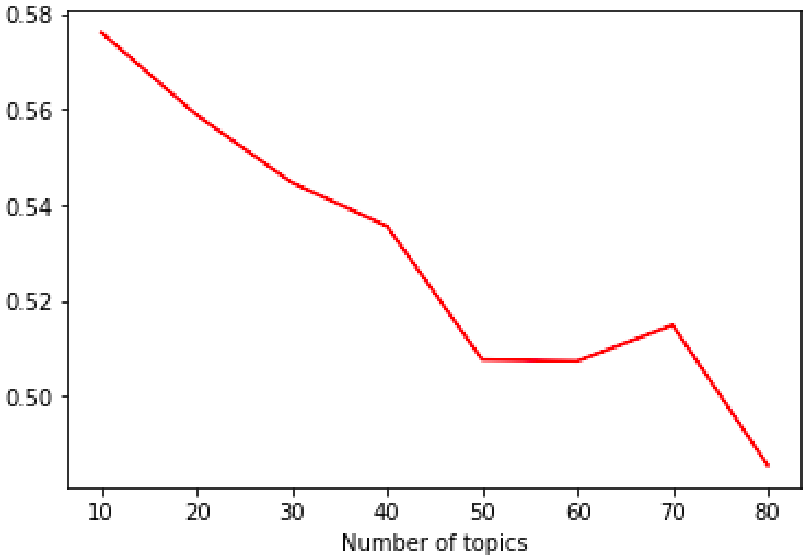
Setting 25: $\alpha = 10/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$

t	Coherence
10	0.54683
20	0.54713
30	0.55593
40	0.54102
50	0.51879
60	0.50733
70	0.51886
80	0.49796



Setting 26: $\alpha = 5/t$, $\eta = (90 \times t)/\ell$, $update\_every = 300$, $offset = 2.1$

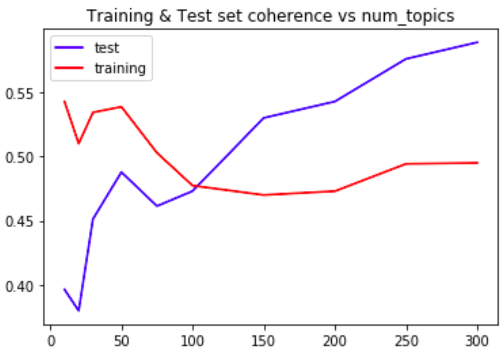
t	Coherence
10	0.57611
20	0.55888
30	0.54473
40	0.53560
50	0.50768
60	0.50739
70	0.51493
80	0.48558



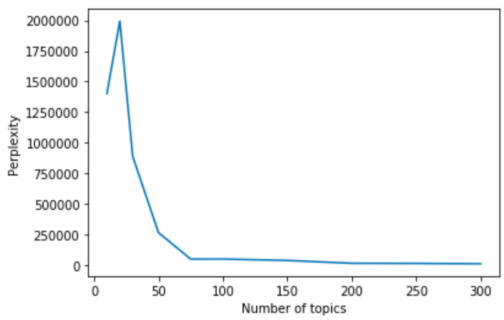
Cross-validation on LSFD

On fold 1:
Setting A: $\alpha = 40/t$, $\eta = (30 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.54302	0.39620	1400596.97
20	0.51019	0.37960	1993926.57
30	0.53453	0.45122	886565.80
50	0.53895	0.48787	262749.46
75	0.50297	0.46137	46388.97
100	0.47735	0.47309	46709.77
150	0.47001	0.53025	35179.79
200	0.47301	0.54307	11949.02
250	0.49437	0.57638	10759.08
300	0.49510	0.58929	7799.31



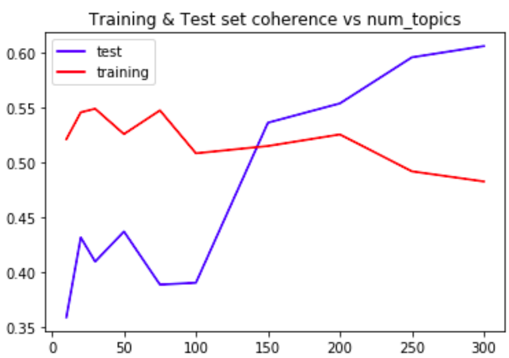
(a) Training and test coherences



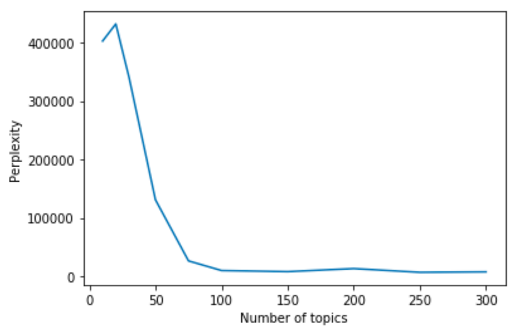
(b) Test perplexity

On fold 1:
Setting B: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.52103	0.35882	402683.04
20	0.54545	0.43147	432035.84
30	0.54870	0.40946	340597.48
50	0.52563	0.43692	131007.39
75	0.54720	0.38854	26509.38
100	0.50812	0.39025	10004.60
150	0.51476	0.53599	8154.20
200	0.52530	0.55356	13330.89
250	0.49171	0.59552	6961.33
300	0.48246	0.60576	7698.23



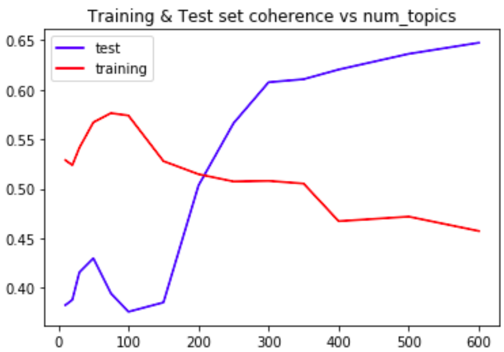
(a) Training and test coherences



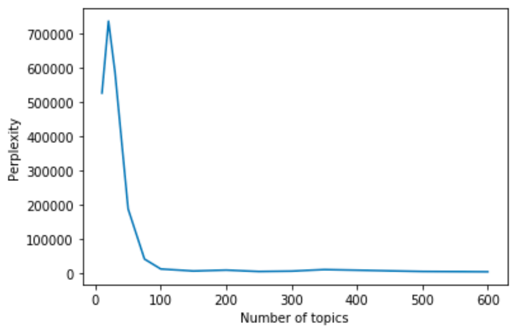
(b) Test perplexity

On fold 1:
Setting C: $\alpha = 80/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.52884	0.38263	527436.70
20	0.52383	0.38790	737927.89
30	0.54127	0.41566	587362.93
50	0.56697	0.42987	188554.73
75	0.57628	0.39456	42022.49
100	0.57385	0.37593	12911.09
150	0.52783	0.38531	7126.16
200	0.51466	0.50329	9756.28
250	0.50732	0.56608	5616.79
300	0.50795	0.60739	6762.91
350	0.50519	0.61046	11315.39
400	0.46726	0.62021	9435.84
500	0.47187	0.63605	5712.01
600	0.45734	0.64724	4802.69



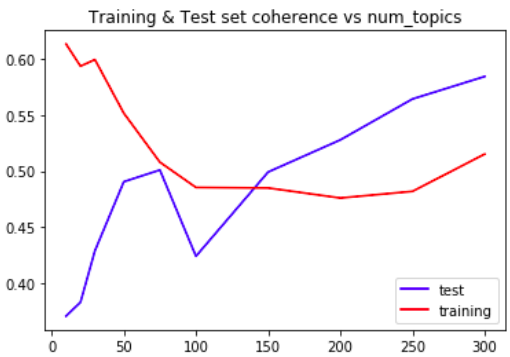
(a) Training and test coherences



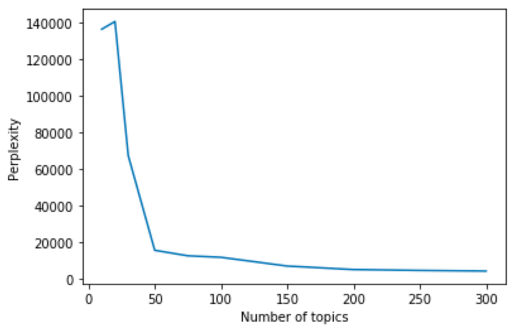
(b) Test perplexity

On fold 1:
Setting a: $\alpha = 40/t$, $\eta = (30 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.61314	0.37073	136442.81
20	0.59340	0.38304	140636.00
30	0.59924	0.42891	67474.87
50	0.55154	0.49041	15803.31
75	0.50786	0.50089	12802.26
100	0.48535	0.42406	11968.42
150	0.48488	0.49910	7201.42
200	0.47595	0.52769	5288.93
250	0.48185	0.56416	4817.22
300	0.51506	0.58419	4466.84



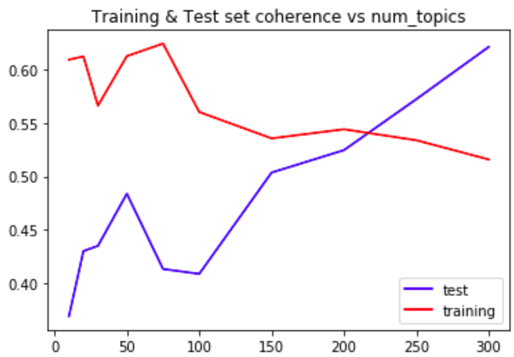
(a) Training and test coherences



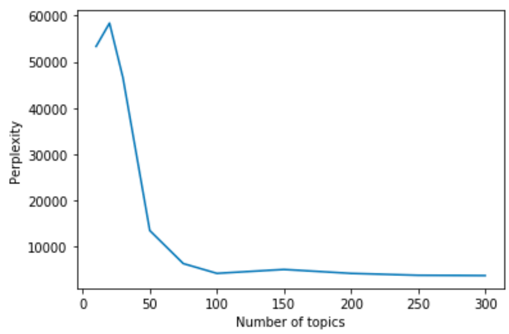
(b) Test perplexity

On fold 1:
Setting b: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.60916	0.36922	53326.13
20	0.61222	0.43009	58386.79
30	0.56613	0.43496	46557.25
50	0.61251	0.48374	13423.37
75	0.62421	0.41332	6262.95
100	0.56018	0.40874	4145.59
150	0.53554	0.50352	4988.59
200	0.54402	0.52451	4141.75
250	0.53376	0.57220	3726.58
300	0.51579	0.62109	3663.68



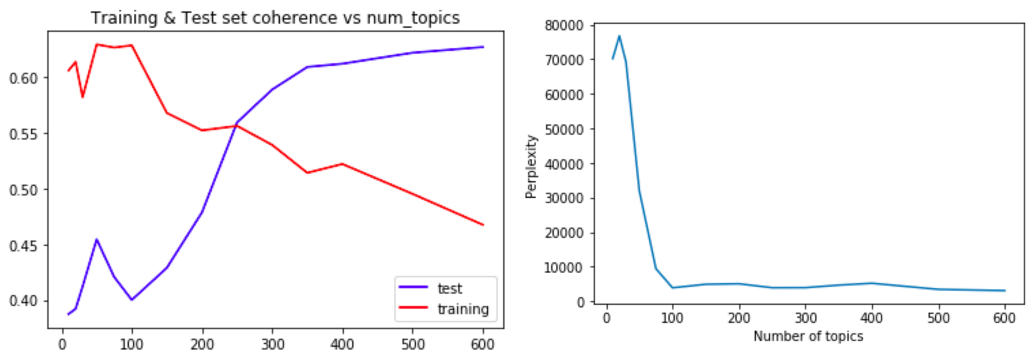
(a) Training and test coherences



(b) Test perplexity

On fold 1:
Setting c: $\alpha = 80/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.60600	0.38742	70174.45
20	0.61378	0.39207	76773.62
30	0.58202	0.41215	69193.20
50	0.62939	0.45447	31957.29
75	0.62664	0.42065	9416.52
100	0.62857	0.40005	3893.86
150	0.56789	0.42888	4910.98
200	0.55225	0.47865	5055.22
250	0.55627	0.55939	3917.09
300	0.53923	0.58893	3935.42
350	0.51416	0.60917	4664.73
400	0.52208	0.61209	5204.69
500	0.49526	0.62194	3440.64
600	0.46755	0.62706	3084.59

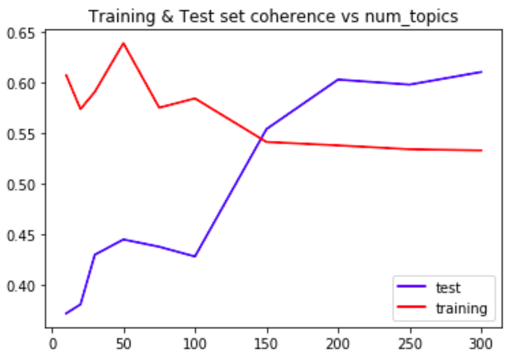


(a) Training and test coherences

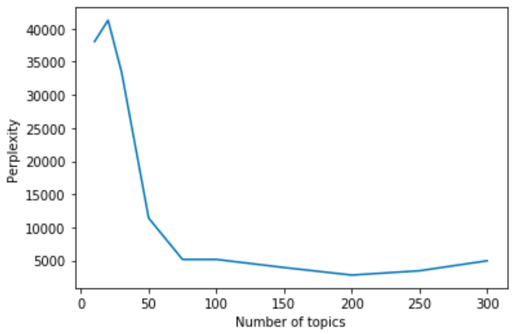
(b) Test perplexity

On fold 2: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.60758	0.37166	38057.67
20	0.57399	0.38057	41276.57
30	0.59103	0.42975	33427.17
50	0.63935	0.44488	11407.13
75	0.57540	0.43767	5167.41
100	0.58456	0.42799	5170.23
150	0.54143	0.55420	3949.99
200	0.53805	0.60321	2808.15
250	0.53422	0.59822	3459.50
300	0.53305	0.61071	4966.94



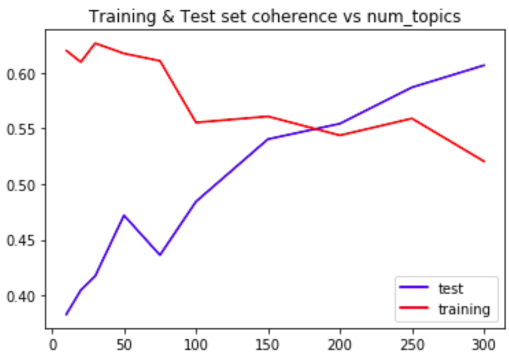
(a) Training and test coherences



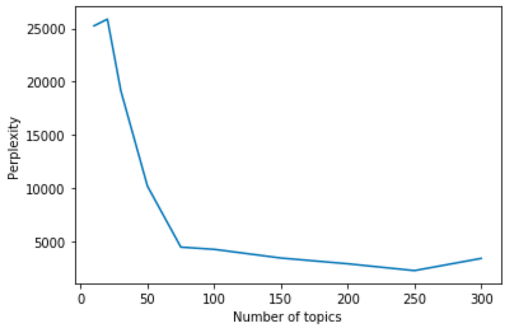
(b) Test perplexity

On fold 3: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.61990	0.38317	25247.03
20	0.60975	0.40482	25869.04
30	0.62661	0.41747	19182.49
50	0.61743	0.47203	10214.65
75	0.61080	0.43630	4499.98
100	0.55542	0.48434	4293.60
150	0.56090	0.54044	3478.30
200	0.54390	0.55436	2939.52
250	0.55904	0.58697	2299.96
300	0.52060	0.60675	3444.64



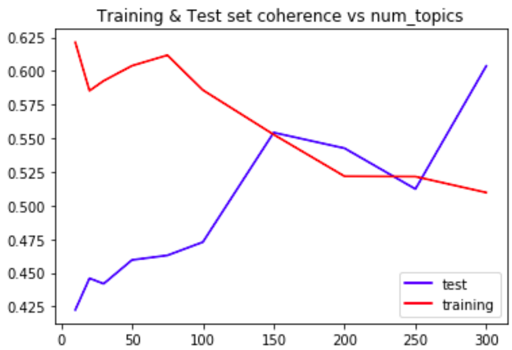
(a) Training and test coherences



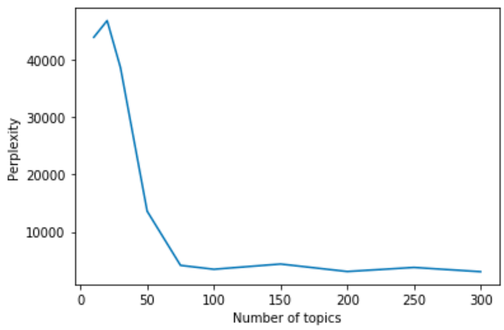
(b) Test perplexity

On fold 4: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.62138	0.42245	43857.23
20	0.58539	0.44595	46751.97
30	0.59277	0.44188	38574.47
50	0.60398	0.45956	13599.61
75	0.61182	0.46298	4179.77
100	0.58589	0.47285	3501.45
150	0.55279	0.55429	4414.10
200	0.52176	0.54262	3111.56
250	0.52156	0.51232	3817.11
300	0.50969	0.60386	3079.52



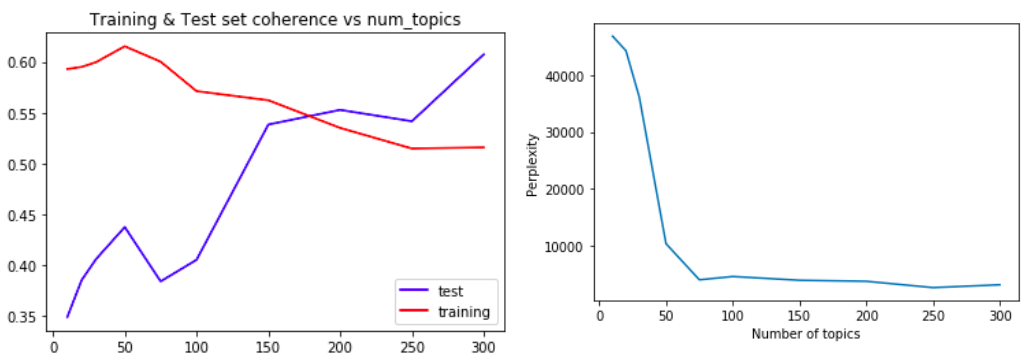
(a) Training and test coherences



(b) Test perplexity

On fold 5: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.59296	0.34854	46925.99
20	0.59506	0.38489	44358.22
30	0.59973	0.40555	44358.22
50	0.61540	0.43716	10304.07
75	0.60019	0.38360	3930.46
100	0.57116	0.40485	4512.00
150	0.56217	0.53824	3850.10
200	0.53493	0.55268	3668.07
250	0.51458	0.54148	2556.34
300	0.51571	0.60740	3052.65

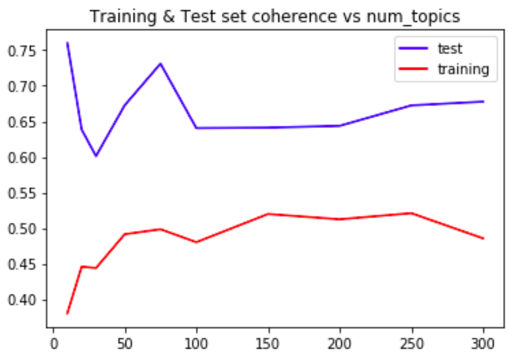


(a) Training and test coherences

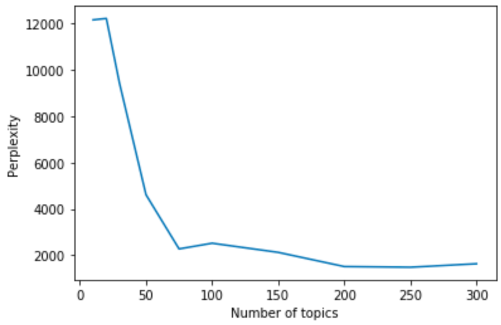
(b) Test perplexity

On fold 6: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.38019	0.76002	12168.54
20	0.44569	0.63842	12228.91
30	0.44378	0.60111	9419.09
50	0.49131	0.67256	4614.67
75	0.49831	0.73110	2275.31
100	0.48012	0.64048	2521.39
150	0.51971	0.64113	2126.33
200	0.51231	0.64367	1513.08
250	0.52087	0.67241	1483.82
300	0.48552	0.67768	1634.74



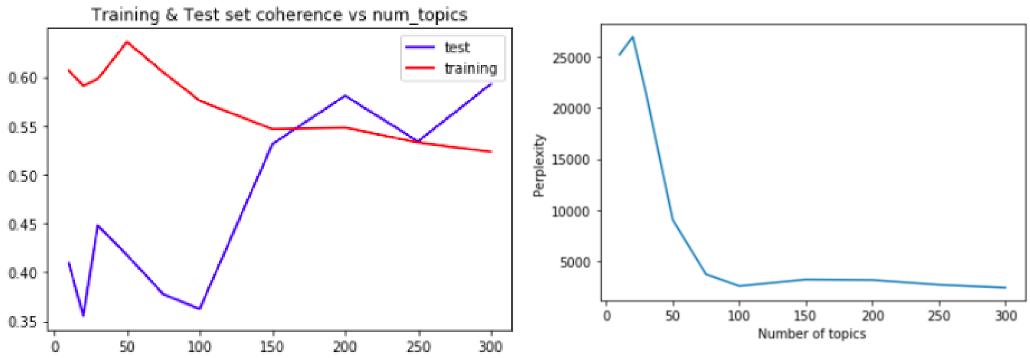
(a) Training and test coherences



(b) Test perplexity

On fold 7: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.60648	0.40910	25198.91
20	0.59094	0.35519	26946.40
30	0.59803	0.44770	21386.21
50	0.63591	0.41767	9072.55
75	0.60473	0.37754	3752.76
100	0.57567	0.36238	2613.26
150	0.54647	0.53129	3239.73
200	0.54799	0.58084	3193.52
250	0.53298	0.53398	2736.33
300	0.52339	0.59256	2452.84

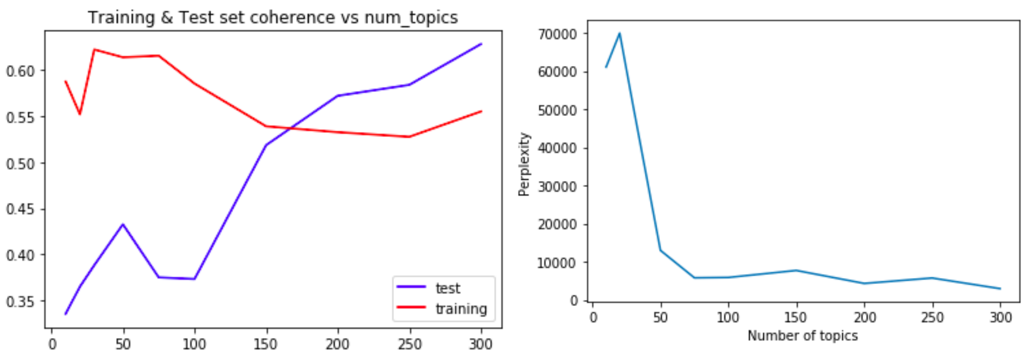


(a) Training and test coherences

(b) Test perplexity

On fold 8: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.58738	0.33553	61059.33
20	0.55174	0.36495	69968.83
30	0.62205	0.38816	50683.26
50	0.61362	0.43252	12983.51
75	0.61534	0.37497	5786.98
100	0.58515	0.37318	5881.85
150	0.53879	0.51837	7715.93
200	0.53240	0.57185	4313.37
250	0.52739	0.58374	5732.37
300	0.55486	0.62794	2949.96

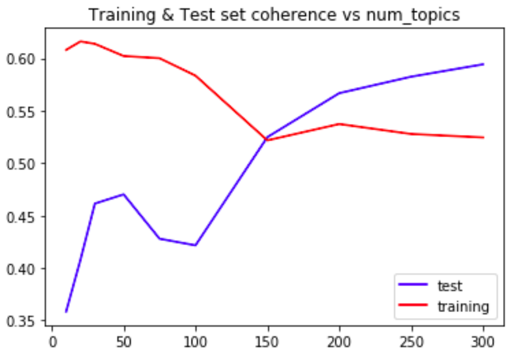


(a) Training and test coherences

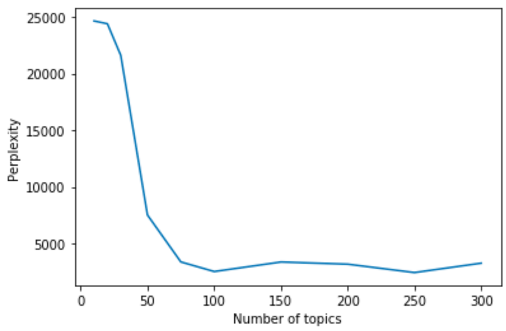
(b) Test perplexity

On fold 9: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.60812	0.35854	24649.26
20	0.61617	0.40806	24400.73
30	0.61387	0.46149	21611.52
50	0.60226	0.47028	7520.61
75	0.60015	0.42798	3381.01
100	0.58352	0.42165	2531.61
150	0.52180	0.52488	3373.58
200	0.53738	0.56681	3181.65
250	0.52790	0.58252	2441.02
300	0.52458	0.59432	3271.19



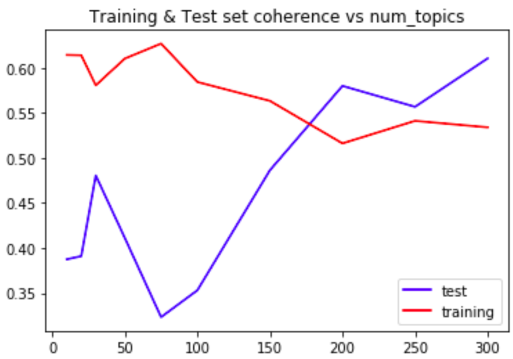
(a) Training and test coherences



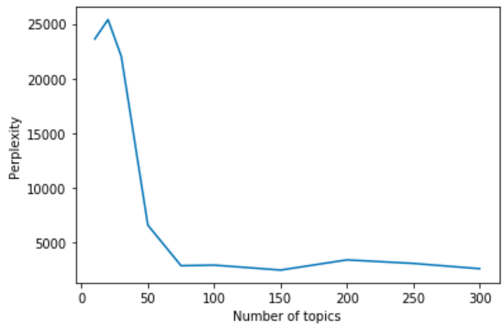
(b) Test perplexity

On fold 10: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.61466	0.38748	23604.96
20	0.61427	0.39097	25387.94
30	0.58088	0.48064	22017.55
50	0.61059	0.41127	6561.69
75	0.62724	0.32322	2848.68
100	0.58459	0.35301	2902.34
150	0.56359	0.48646	2448.02
200	0.51632	0.58023	3380.49
250	0.54138	0.55699	3056.98
300	0.53418	0.61075	2573.61



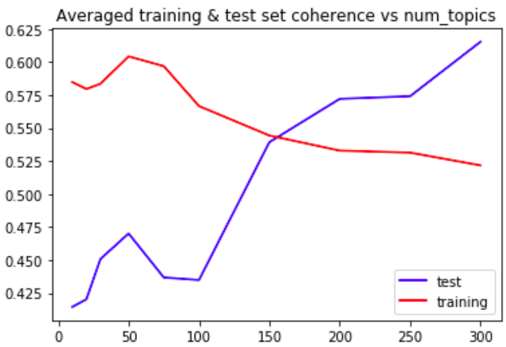
(a) Training and test coherences



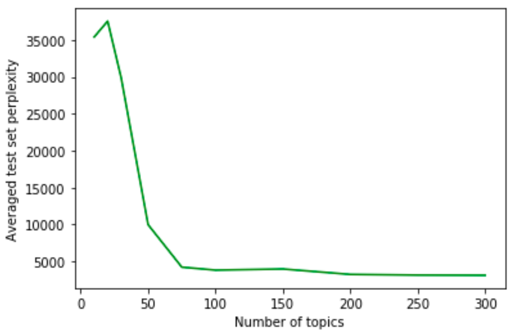
(b) Test perplexity

Setting: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Ave. training coherence	Ave. test coherence	Ave. test perplexity
10	0.58478	0.41457	35409.50
20	0.57952	0.42039	37557.54
30	0.58349	0.45087	29902.25
50	0.60424	0.47017	9970.19
75	0.59682	0.43687	4208.53
100	0.56663	0.43495	3807.33
150	0.54432	0.53928	3958.47
200	0.53291	0.57208	3225.12
250	0.53137	0.57408	3131.00
300	0.52174	0.61531	3108.98



(a) Averaged training and test coherences

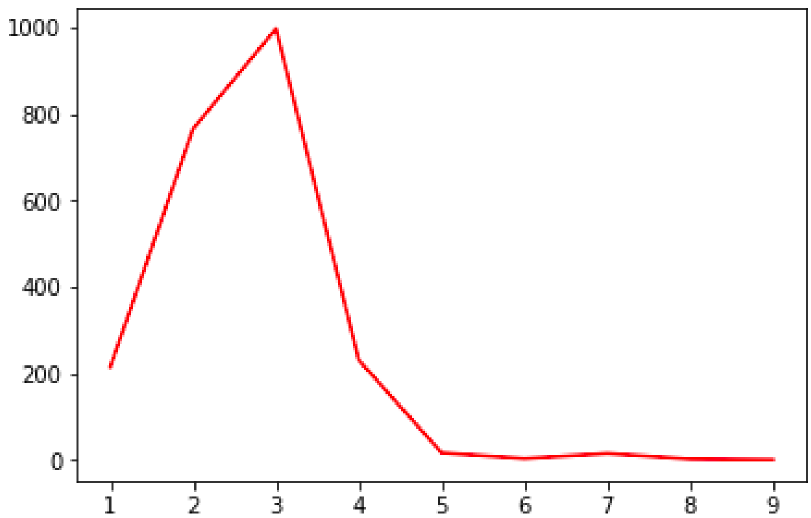


(b) Averaged test perplexity

Rate of test perplexity change

Setting: $\alpha = 60/t$, $\eta = (50 \times t)/\ell$, $update\_every = 200$, $offset = 4$

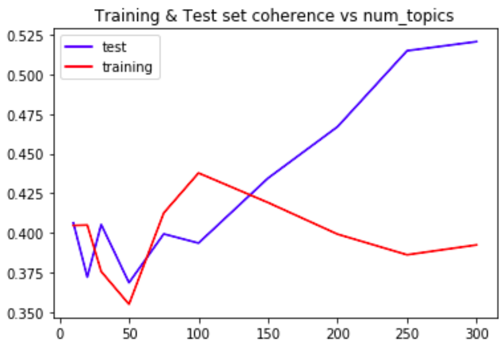
t	rpc
1	214.80
2	765.52
3	996.60
4	230.46
5	16.04
6	3.02
7	14.66
8	1.88
9	0.44



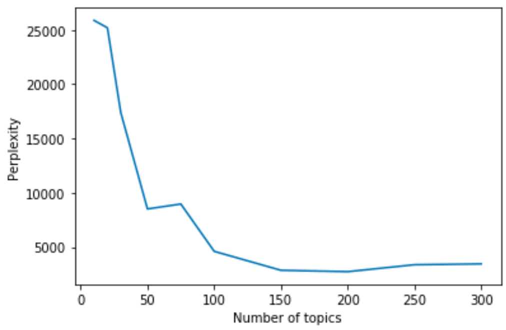
Cross-validation on LSFD based on Benford's law

On fold 1: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.40462	0.40624	25880.46
20	0.40488	0.37200	25199.18
30	0.37553	0.40523	17395.41
50	0.35488	0.36847	8513.07
75	0.41252	0.39927	8966.10
100	0.43777	0.39349	4606.16
150	0.41909	0.43451	2861.06
200	0.39912	0.46700	2735.37
250	0.38610	0.51499	3373.21
300	0.39233	0.52074	3450.17



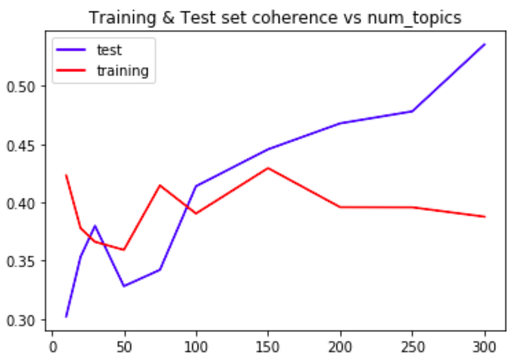
(a) Training and test coherences



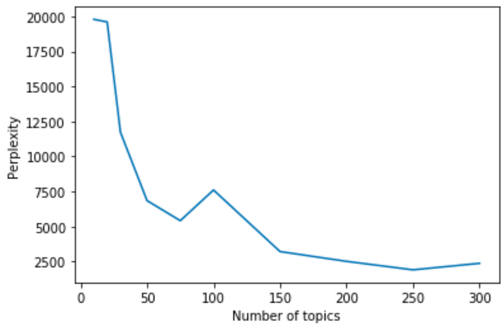
(b) Test perplexity

On fold 2: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.42317	0.30209	19789.82
20	0.37776	0.35334	19603.95
30	0.36613	0.37983	11729.02
50	0.35916	0.32802	6847.71
75	0.41456	0.34202	5415.97
100	0.39043	0.41388	7605.52
150	0.42944	0.44558	3211.94
200	0.39588	0.46787	2512.28
250	0.39576	0.47817	1906.68
300	0.38770	0.53563	2371.30



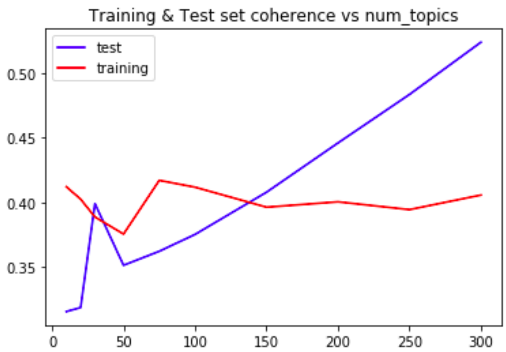
(a) Training and test coherences



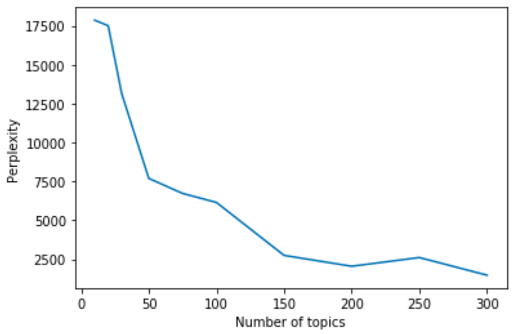
(b) Test perplexity

On fold 3: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.41197	0.31557	17865.04
20	0.40236	0.31868	17509.49
30	0.38862	0.39892	13149.26
50	0.37543	0.35131	7702.37
75	0.41691	0.36214	6734.32
100	0.41164	0.37514	6156.51
150	0.39619	0.40774	2754.12
200	0.40033	0.44581	2055.90
250	0.39430	0.48325	2616.03
300	0.40556	0.52361	1482.53



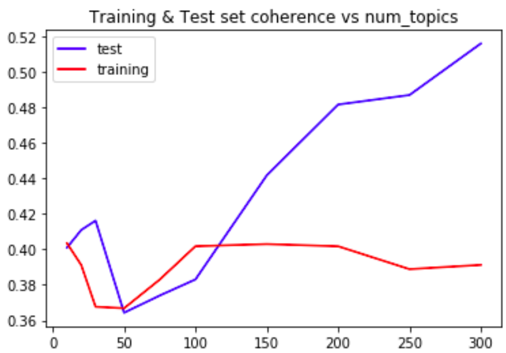
(a) Training and test coherences



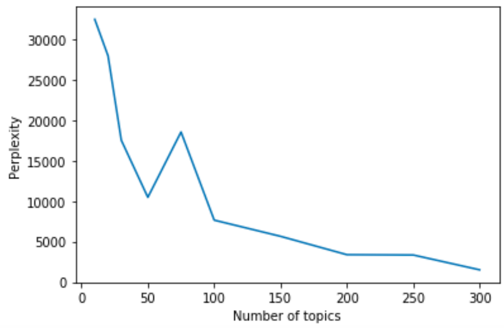
(b) Test perplexity

On fold 4: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.40347	0.40103	32485.10
20	0.39116	0.41100	27973.81
30	0.36767	0.41626	17531.04
50	0.36681	0.36430	10478.60
75	0.38302	0.37400	18554.48
100	0.40182	0.38307	7654.87
150	0.40302	0.44181	5653.04
200	0.40179	0.48177	3379.43
250	0.38888	0.48712	3343.90
300	0.39126	0.51618	1509.02



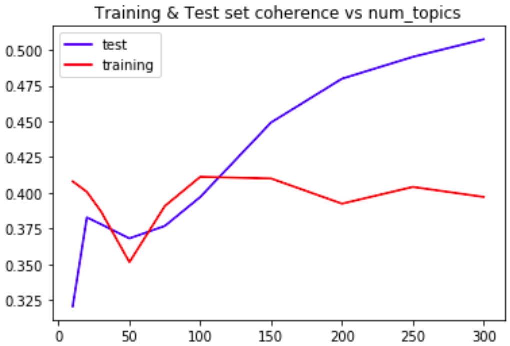
(a) Training and test coherences



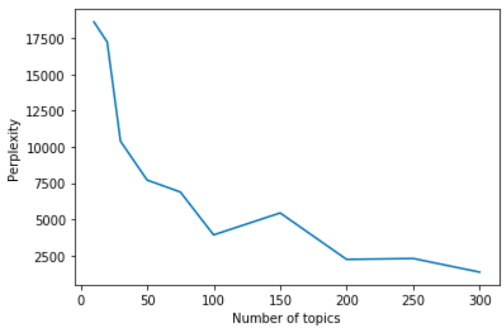
(b) Test perplexity

On fold 5: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.40804	0.32054	18629.32
20	0.40061	0.38279	17232.20
30	0.38686	0.37800	10386.07
50	0.35149	0.36811	7705.32
75	0.39075	0.37680	6877.75
100	0.41120	0.39707	3917.79
150	0.41002	0.44925	5432.70
200	0.39240	0.47985	2223.04
250	0.40413	0.49513	2289.67
300	0.39712	0.50743	1345.88



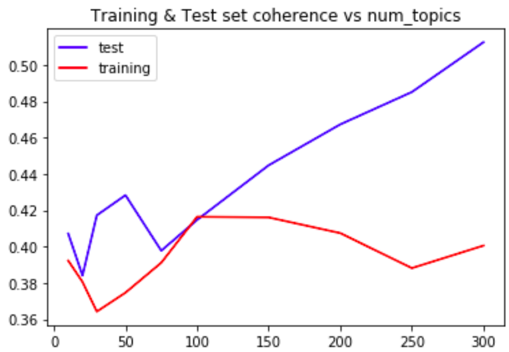
(a) Training and test coherences



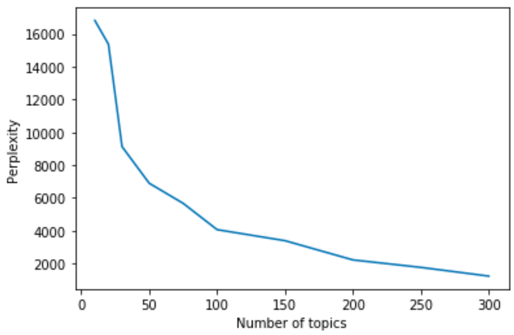
(b) Test perplexity

On fold 6: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.39232	0.40721	16813.05
20	0.38058	0.38413	15367.28
30	0.36430	0.41724	9125.73
50	0.37468	0.42829	6893.23
75	0.39116	0.39777	5668.66
100	0.41641	0.41469	4064.63
150	0.41608	0.44474	3391.94
200	0.40744	0.46723	2219.87
250	0.38818	0.48502	1764.26
300	0.40056	0.51242	1232.68



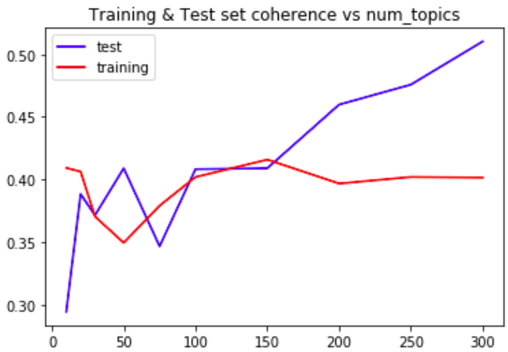
(a) Training and test coherences



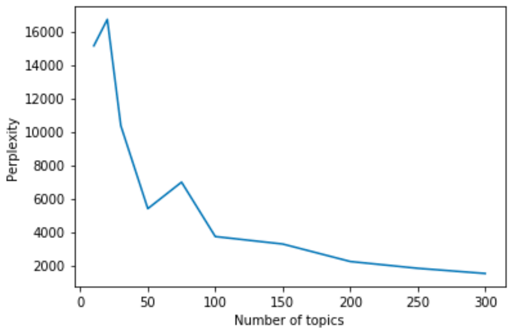
(b) Test perplexity

On fold 7: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.40918	0.29428	15140.96
20	0.40623	0.38830	16722.89
30	0.37035	0.37137	10359.11
50	0.34934	0.40891	5394.14
75	0.37902	0.34658	6979.79
100	0.40191	0.40818	3723.79
150	0.41583	0.40894	3276.15
200	0.39671	0.45981	2230.41
250	0.40195	0.47579	1823.20
300	0.40144	0.51031	1513.81



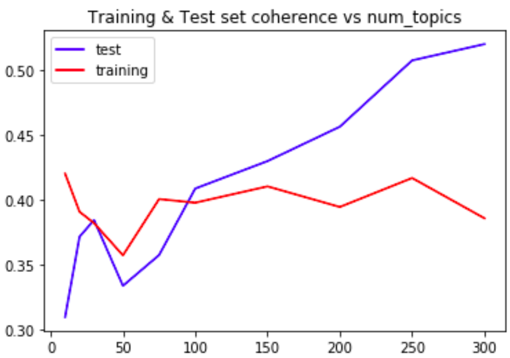
(a) Training and test coherences



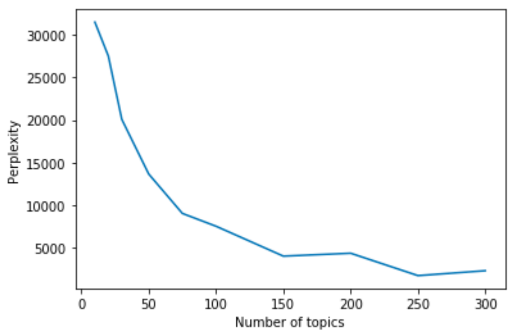
(b) Test perplexity

On fold 8: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.41989	0.30948	31506.36
20	0.39050	0.37134	27520.61
30	0.38179	0.38414	20079.58
50	0.35697	0.33351	13654.92
75	0.40032	0.35717	9026.45
100	0.39751	0.40842	7519.43
150	0.41009	0.42958	4007.17
200	0.39423	0.45607	4347.26
250	0.41654	0.50715	1717.18
300	0.38553	0.51971	2293.42



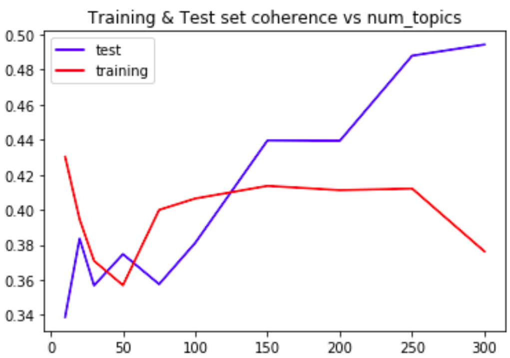
(a) Training and test coherences



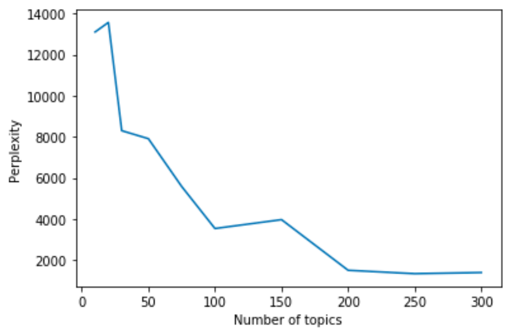
(b) Test perplexity

On fold 9: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.43031	0.33882	13114.98
20	0.39458	0.38358	13581.18
30	0.37087	0.35680	8306.94
50	0.35694	0.37473	7917.14
75	0.40000	0.35756	5588.50
100	0.40647	0.38116	3539.95
150	0.41369	0.43967	3972.15
200	0.41123	0.43940	1504.69
250	0.41215	0.48800	1339.43
300	0.37627	0.49435	1396.61



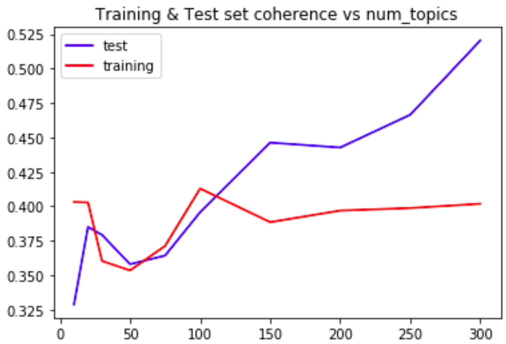
(a) Training and test coherences



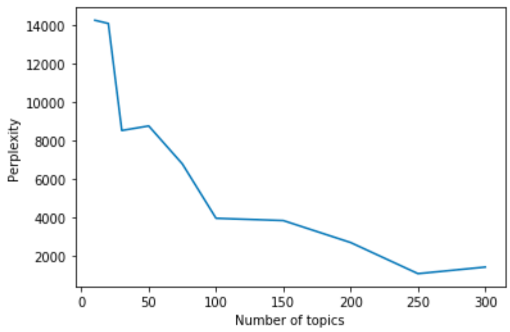
(b) Test perplexity

On fold 10: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Training coherence	Test coherence	Test perplexity
10	0.40324	0.32877	14236.75
20	0.40286	0.38502	14069.57
30	0.36036	0.37932	8500.33
50	0.35351	0.35802	8737.22
75	0.37129	0.36429	6759.64
100	0.41296	0.39577	3931.77
150	0.38856	0.44631	3812.27
200	0.39697	0.44274	2669.14
250	0.39880	0.46657	1050.81
300	0.40192	0.52054	1393.01



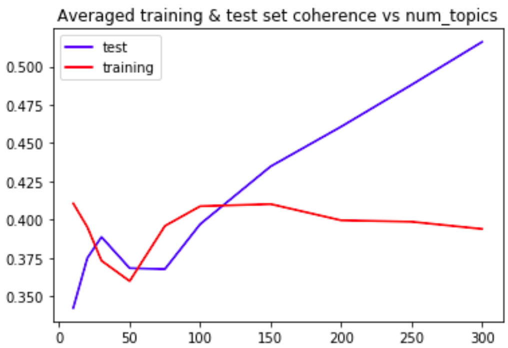
(a) Training and test coherences



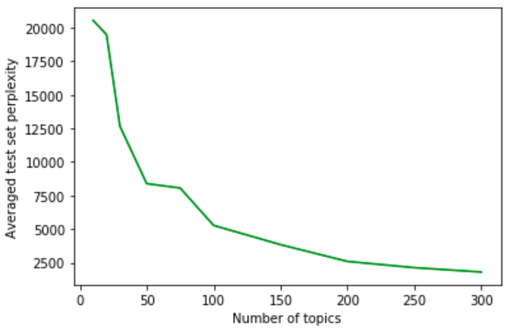
(b) Test perplexity

Setting: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	Ave. training coherence	Ave. test coherence	Ave. test perplexity
10	0.41062	0.34240	20546.18
20	0.39515	0.37502	19478.02
30	0.37325	0.38871	12656.25
50	0.35992	0.36837	8384.37
75	0.39595	0.36776	8057.17
100	0.40881	0.39709	5272.04
150	0.41020	0.43481	3837.26
200	0.39961	0.46075	2587.74
250	0.39868	0.48812	2122.44
300	0.39397	0.51609	1798.84



(a) Averaged training and test coherences



(b) Averaged test perplexity

Rate of test perplexity change

Setting: $\alpha = 20/t$, $\eta = (10 \times t)/\ell$, $update\_every = 200$, $offset = 4$

t	rpc
1	106.81
2	682.17
3	213.59
4	13.08
5	111.40
6	28.69
7	24.99
8	9.30
9	6.47