**LDA MODELS AND PARAMETERS TUNING**

* Training size and number of topics picked from the literature

|  |  |  |
| --- | --- | --- |
| **TRAINING SIZE** | **NUMBER OF TOPICS** | **LITERATURE** |
| 69 Interviews with avg 8234 words length | 100 topics through triangulation (*ldatuning* in R)   * Griffiths2004 * CaoJuan2009 * Arun2010 * Deveaud2014 | <https://doi.org/10.1002/smj.3067>  (Choudhury et al, 2019) – Supporting Information |
| 317000 worker attribute sentences | 140 topics base on topic distances (Cao, Xia,  Li, Zhang, and Tang (2009)) and matrix factorization (Arun, Suresh, Madhavan, & Murthy,2010) | http://journals.sagepub.com/doi/suppl/10.1177/  1094428117722619  (Kobayashi et al., 2017) - Supplemental Material |
| 2,826 fullerene and nanotube  patents | 100 Topics. Three field experts separately reviewed each of the 100 topics top 20 words.   * Krippendorff | <https://doi.org/10.1002/smj.2294>  (Kaplan et al., 2014) – Supporting Information |
| 1156 research articles abstracts | 10 Topics. Internal validity (DiMaggio, Nag, Bei, 2013) and coherence score (Mimno, Wallach, Talley et al., 2011)  LDA Mallet | <https://doi.org/10.1016/j.leaqua.2019.101338>  (Sieweke et al., 2019) – Supporting Information |
| 1992758 twitter messages | 100 Topics. Based on classification performance. | [https://snap.stanford.edu/soma2010/papers/ soma2010\_12.pdf](https://snap.stanford.edu/soma2010/papers/%20soma2010_12.pdf)  (Hong et al., 2010) |

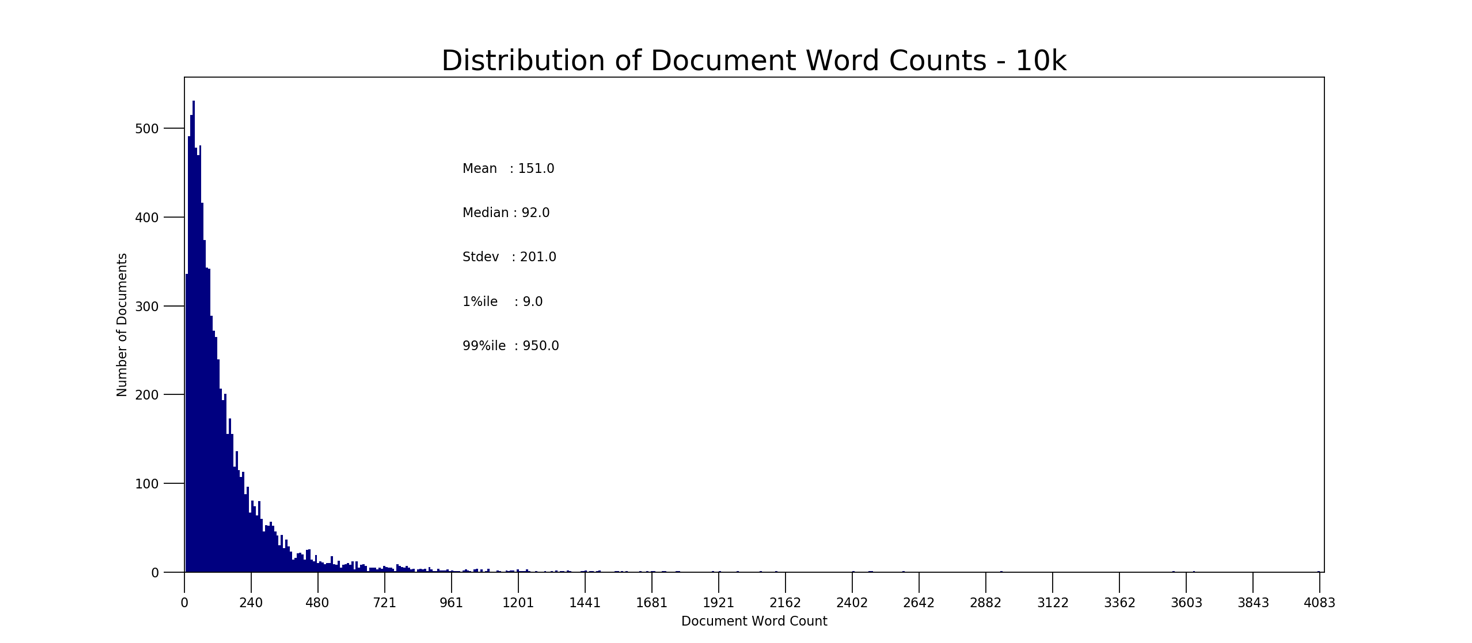
**NOTE:** training size differs a lot across the studies. Triangulation seems to be a good means to identify the number of topics, this will be used in our modelling phase.

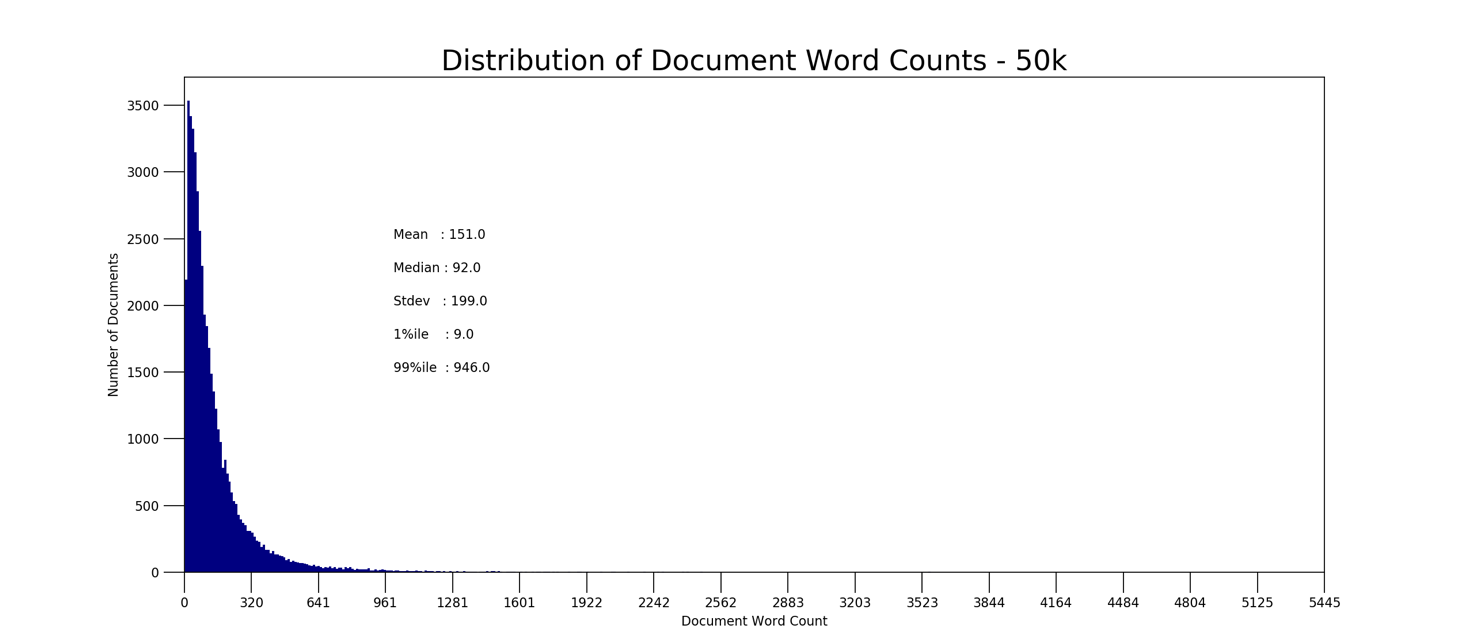
**OUR SPECIFIC CASE ASSUMPTIONS:**

* We are dealing with sparsity (the same word is rarely used. The most used word is in 22% of the comments).
  + From the dictionary:

|  |  |  |
| --- | --- | --- |
| Comments | Unique words (dictionary length) | Descriptive Statistics: word used in how many comments? |
| 9308 | 11108  (6001 after filtering out words that occur less than in 2 comments)  See attached excel file: Dictionaries | mean 16.369553 (0.18%)  std 71.812963  min 1.000000  25% 1.000000  50% 2.000000  75% 6.000000  max 2075.000000 (22.29%) |
| 46609 | 23529  (12775 after filtering out words that occur less than in 2 comments)  See attached excel file: Dictionaries | mean 38.875473 (0.08%)  std 252.560218  min 1.000000  25% 1.000000  50% 2.000000  75% 7.000000  max 10429.000000 (22.38%) |

* Pragmatical: how people behave on reddit.
  + People write relatively short messages:



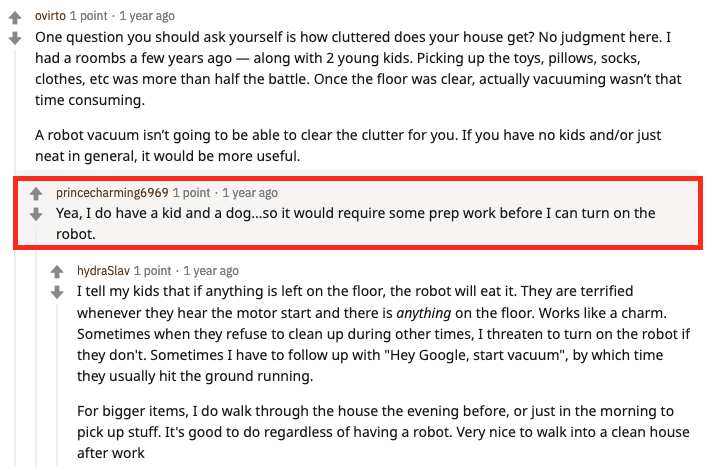


* + Messages are highly susceptible to intertwined conversations and incoherence. Interesting paper about how to deal with it: <https://doi.org/10.25300/MISQ/2018/13239> (Abbasi et al., 2018)

**Example:**

Red is the focal message. Its alone is hard to understand the topics the author is talking about.

If you read all the parents’ comments, thus the comments within the green box, it gets clear that they are talking about ROBOT VACUUM AND THEIR USEFULNESS WITH YOUNG KIDS AROUND.



**OUR SPECIFIC GOAL:**

Within a specific area, smart home, analyze consumers behavior.

**HYPERPARAMETERS TUNING:**

**ALPHA**: set the prior on the per-document topic distribution.

Do people talk about many topics when commenting?

* LOW: each comment covers only few topics (higher impact on topic sparsity)
* HIGH: each comment covers many topics

**BETA**: set the prior on the per-topic word distribution

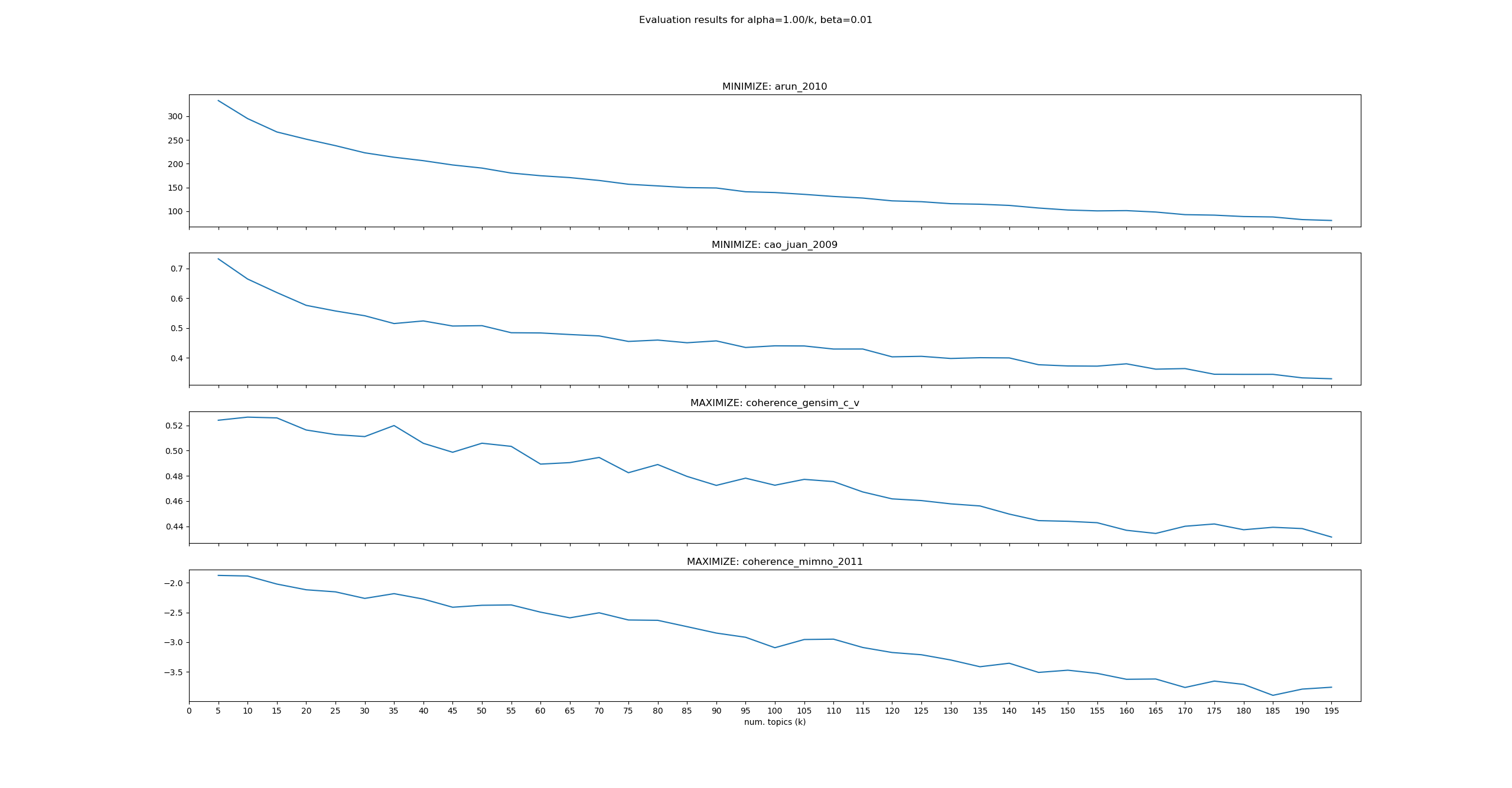
Are topics interrelated? Same word used in different contexts.

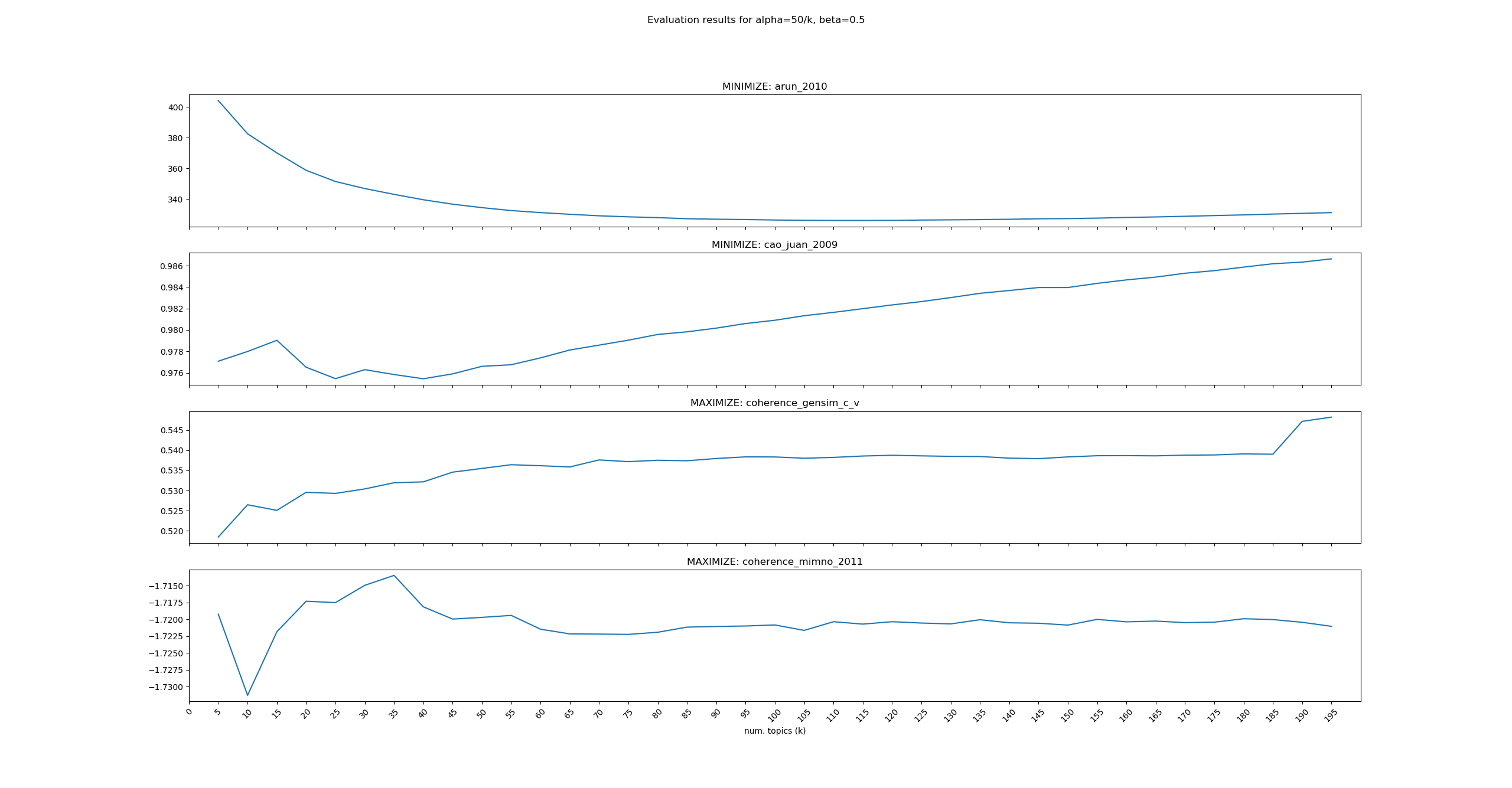
* LOW: each topic consists of few words (specific topics) > results in more topics, more specific (higher impact on word sparsity)
* HIGH: each topic consists of many words > fewer topics, more general

**NOTE:** green columns models have the graph depict below and an excel file with the topics found based on coherence genism c\_v

* Parameters tuning: 9308 comments > **NO CONVERGENCE**

|  |  |  |  |
| --- | --- | --- | --- |
| **EVALUATION METRICS** | **ALPHA=1/k**  **BETA=0.01**  **TOPICS=(5, 200, 5)** | **ALPHA=10.00/k**  **BETA=0.1**  **TOPICS=(5,200,5)** | **ALPHA=50.00/k**  **BETA=0.5**  **TOPICS=(5, 200, 5)** |
| Min: Arun 2010 | 80.5942 @topic:195 | 205.7174 @topic:195 | 326.0647 @topic: 115 |
| Min: Cao Juan 2009 | 0.33034 @topic:195 | 0.9357 @topic:195 | 0.9754 @topic: 40 |
| Max: Coherence gensim c\_v | 0.5266 @topic:10 | 0.5483 @topic:190 | 0.5482 @topic: 195 |
| Max: Coherence Mimno 2011 | -1.8766 @topic:5 | -1, 7143 @topic:5 | -1.71347 @topic: 35 |





* Parameters tuning: 46609 comments > **DID NOT HELP MUCH**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EVALUATION METRICS** | **ALPHA=0.1/K**  **BETA=0.001**  **TOPICS=(5,501,5)** | **ALPHA=1/k**  **BETA=0.01**  **TOPICS=(5, 501, 5)** | **ALPHA=10.00/k**  **BETA=0.1**  **TOPICS=(5,501,5)** | **ALPHA=50.00/k**  **BETA=0.5**  **TOPICS=(5, 501, 5)** |
| Min: Arun 2010  <http://doi.org/10.1007/978-3-642-13657-3_43> | 139.338  @topic: 500 | 341.55349 @topic:500 | 498 @topic:500 | 743.495 @topic: 225 |
| Min: Cao Juan 2009  <https://doi.org/10.1016/j.neucom.2008.06.011> | 0.2226  @topic:495 | 0.75011 @topic:5 | 0.9572 @topic:500 | 0.9752 @topic: 10 |
| Max: Coherence gensim c\_v | 0.5569  @topic 30 | 0.5498 @topic:145 | 0.5571 @topic:330 | 0.55616 @topic: 390 |
| Max: Coherence Mimno 2011 | -1.8953  @topic 5 | -1.71868 @topic:180 | -1, 70309 @topic:30 | -1.7014 @topic: 20 |

