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 (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.

Hyperparameters tuning

- **ALPHA**: set the prior on the per-document topic distribution.
  - Do people talk about many topics when commenting?

LOW ALPHA:	HIGH ALPHA:
Each comment covers only few topics (higher impact on topic sparsity)	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 BETA:	HIGH BETA:
Each topic consists of few words. Result in more topics and more specific. (Higher impact on word sparsity)	Each topic consists of many words. Few topics, more general