**TOPIC MODELLING**

**Introduction:**

1. [**Prof. David Blei - Probabilistic Topic Models and User Behavior**](https://www.youtube.com/watch?v=FkckgwMHP2s&t=139s)
2. [**Probabilistic Topic Modeling — Pyro Tutorials 1.7.0 documentation**](https://pyro.ai/examples/prodlda.html)
3. [**MLSS 2019 David Blei: Variational Inference: Foundations and Innovations (Part 1)**](https://www.youtube.com/watch?v=DaqNNLidswA&t=3809s)
4. [**David Blei Variational Inference Foundations and Innovations Part 2**](https://www.youtube.com/watch?v=Wd7R_YX4PcQ)

**Papers:**

**[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003).** [**Latent Dirichlet allocation.**](https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)

**[2] Akash Srivastava, & Charles Sutton. (2017).** [**Autoencoding Variational Inference For Topic Models.**](https://arxiv.org/abs/1703.01488)

**[3]** [**TOP2VEC: DISTRIBUTED REPRESENTATIONS OF Topics**](https://arxiv.org/pdf/2008.09470.pdf)

**Problem 1: what is the problem the three papers aim to solve, and why is this problem important or interesting?**

**Problem 2: 1) summarize the three methods, including high-level ideas as well as technical details: the relevant details that are important to focus on (e.g., if there’s a model, define it; if there is a theorem, state it and explain why it’s important, etc) 2) what are the major differences of the three methods?**

**Problem 3: implement the ProdLDA [2] topic model and test it on the 20 newsgroups text dataset (report the 20 main topics you discovered from the data and visualize their word clouds)**

**Reference code:** [**https://pyro.ai/examples/prodlda.html**](https://pyro.ai/examples/prodlda.html)

**Problem 4: implement Latent Dirichlet Allocation (LDA) [1] and test it on the 20 newsgroups text dataset (report the 20 main topics you discovered from the data and visualize their word clouds)**

**Reference code:** [**https://pyro.ai/examples/lda.html#**](https://pyro.ai/examples/lda.html)

**Problem 5: implement TOP2VEC [3] and test it on the 20 newsgroups text dataset (report the 20 main topics you discovered from the data and visualize their word clouds)**

**Reference code:** [**https://github.com/ddangelov/Top2Vec**](https://github.com/ddangelov/Top2Vec)

**Problem 6: compare the experimental results of the three methods with the ground truth topics of the 20 newsgroups. Which method do you think is better? Explain why.**

**Problem 7: fetch real-time tweets and discover real-time topics on them with the three methods, respectively. You need to report the top 5 topics for the last 5 seconds in real-time and the top 5 topics for each hour in the most recent 24 hours.**

**Problem 8: fetch real-time news and discover real-time topics on them with the three methods, respectively. You need to report the top 5 topics for each day in the most recent week.**

**Problem 9 (bonus): analyze the pitfalls of the existing topic modeling methods above and come up with one way to address the pitfalls. Can you realize your ideas?**