Data Science Toolbox Question Sheet

07.1 Topic Models

Daniel Lawson

Block 7

1. In a topic model, describe the **corpus**, the **document**, and the **dictionary**.
2. Interpret the tf-idf. In what sense is it principled?
3. What is an N-gram? What are the advantages and disadvantages of using N-grams?
4. What is the difference between a bag-of-words model and Latent Dirichlet Allocation?
5. Define the False Discovery Rate and accuracy. Let Show how Bayes Theorem useful for understanding these.
6. Why is it important for tractability that we use a method to integrate out $p(x)$ when trying to compute posterior probabilities $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$?
7. Variational methods allow the closest conjugate model to the desired model to be used to compute $p(x)$. Why is this useful?
8. Given verbal explanations of intrinsic and extrinsic coherence. What advantages and disadvantages do they have?