Data Science Toolbox Question Sheet

07.1 Topic Models and Bayes

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. Here we will 'derive' a justification for tf-idf.
   a. Given the tf-idf definition $\text{tf}(t,d) = X_d(t)/\sum_{t=1}^{T} X_d(t)$ and $\text{idf}(t,d) = -log\left(\frac{1+n_d(t)}{D}\right)$, write tf and idf in terms of joint, conditional or marginal probabilities of t and d with justification.
   b. Use Bayes' Theorem to write the joint probability $p(t,d)$ in terms of tf, explaining any assumptions that are needed about $p(d)$.
   c. Use Bayes' Theorem to write the log joint probability $-log(p(t,d))$ in terms of idf and the apriori probability of the terms $p(t)$.
   d. Given the formula for the Mutual Information between a random variable D describing documents, and T describing terms:

$$(T,D) = \sum_t \sum_d p(t,d) log\left(\frac{p(t,d)}{p(t)p(d)}\right),$$

   write this in terms of tf-idf using the results above.
4. What is an N-gram? What are the advantages and disadvantages of using N-grams?
5. What is the difference between a bag-of-words model and Latent Dirichlet Allocation?
6. Define the False Discovery Rate and accuracy. State Bayes Theorem and explain how it useful for understanding these.
7. Why is it important for tractability that we use a method to integrate out $p(x)$ when trying to compute posterior probabilities $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$?
8. Variational methods allow the closest conjugate model to the desired model to be used to compute $p(x)$. Why is this useful?
9. Given verbal explanations of intrinsic and extrinsic coherence. What advantages and disadvantages do they have?