

Portfolio Question 7.2.P - Interpreting LDA

Daniel Lawson University of Bristol

Portfolio 7.2.P (v4.0.0)

Portfolio Question 7.2.P - Interpreting LDA

- ▶ If we've looked at a collection of documents and a model like LDA has identified some "topics", how would we interpret them?
 - ▶ You might start with perplexity and coherence.
 - ▶ Do small language models help with interpretability?
 - ▶ What about large language models?

Quantifying solutions

- ▶ There are many ways to quantify how good a particular LDA model is. The most popular are:
- ▶ **Perplexity**: the perplexity is $2^{-H(D)}$ where $H(D) = \sum_{t=1}^T \log(p(t|\theta_d))$
 - ▶ $p(t|\theta_d) = \sum_{v=1}^V \theta_d(v)p(t|v)$ uses the model-learned topics V for the (held out!) document d with topic distribution θ_d .
 - ▶ It is the entropy of term t (normally reported as the average per-word).
 - ▶ Perplexity is low (better) when each word appears in only one topic.
 - ▶ Perplexity is high when words are distributed across topics.
- ▶ **Coherence**: a measure of how often pairs of words appear together. there are two ways to examine this:
 - ▶ **intrinsic coherence**: called u_mass , this compares within a corpus.
 - ▶ **extrinsic coherence**: called c_v , this compares to some standard reference documents.

▶ Neither is particularly consistent with human judgement¹.

¹Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang and David M. Blei. 2009. **Reading Tea Leaves: How Humans Interpret Topic Models.**

Coherence

- ▶ The coherence is based on the score ² (defined next):

$$Coherence(V) = \sum_{(t_i, t_j) \in V} score(t_i, t_j)$$

- ▶ Where V is a topic, and t_i, t_j are word pairs.
- ▶ In both cases we use a **regulariser** ϵ .
 - ▶ $\epsilon = 1$ is natural but not obligatory.

²Stevens, Kegelmeyer, Andrzejewsk and Buttler **Exploring Topic Coherence over many models and many topics**

intrinsic coherence

- ▶ Using the score function:

$$u_mass(v_i, v_j) = \log \left(\frac{p(v_i, v_j, \epsilon)}{p(v_i)p(v_j)} \right)$$

- ▶ i.e. we compare the probability that the words co-occur in a document with their relative frequencies.
- ▶ ϵ assigns non-zero weight to word pairs that do not occur together in a document.

Extrinsic coherence

- ▶ Using the score function:

$$c_v(v_i, v_j) = \log \left(\frac{D(v_i, v_j, \epsilon)}{D(v_j)} \right)$$

- ▶ where D counts documents that contain the word(s);
- ▶ i.e. we compare the frequency in which words co-occur in an external dataset, compared to their external frequency.

Discussion

- ▶ These statistics try to find “things that some but not all documents have in common”.
 - ▶ Is this sensible?
 - ▶ Are there other approaches?
- ▶ Investigate interpretation of small language models like **BERT** (**Bidirectional Encoder Representations from Transformers**)
- ▶ And Large language models. Are they easier to interpret and/or more reliable?