Indicators to Assess the Statistical Fit of Topic Modeling

1 Arun et al. 2010

Arun et al. (2010) obtained their measure leveraging on the Topic-Word matrix (M_1), and the Document-Topic matrix (M_2). In particular, employing the Symmetric Kullback-Leibler (KL) divergence measure, the optimal number of topics is obtained when the following equation reaches its minimum:

$$Arun2010(M_1, M_2) = KL(C_{M_1} \parallel C_{M_2}) + KL(C_{M_2} \parallel C_{M_1}) \tag{1}$$

Where, C_{M_1} is the distribution of singular values obtained applying Singular Value Decomposition (SVD) to the matrix M_1, and C_{M_2} is the distribution obtained normailizing the vector $L \times M_2$ (L is a vector of documents lenghts)(Arun et al. 2010).

2 Cao et al. 2009

Cao et al. (2009) measure computes the average cosine distance among topics, to measure the topic structure stability:

$$ave\_dis(structure) = \frac{\sum_{i=0}^{K} \sum_{j=i+1}^{K} corre(T_i, T_j)}{K \times (K-1)/2} \tag{2}$$

Where K is the number of topics, and T_i and T_j represent two topics. The correlation is measured as:

$$corre(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \sqrt{\sum_{v=0}^{V} (T_{jv})^2}} \tag{3}$$

A lower distance corresponds to a better structure, therefore the optimal topic number is the one with minimum $ave\_dis(structure)$(Cao et al. 2009).

3 Deveaud et al. 2014

Deveaud et al. (2014) method is based on the following:

$$\hat{K} = arg\max_{K} \frac{1}{K(K-1)} \sum_{k,k' \in T_k} D(k \parallel k') \tag{4}$$

Where K is the number of topics given as parameters, T_K is the set of K topics modeled, and $D(k \parallel k')$ is the Jensen-Shannon divergence between pairs of topics(Deveaud, SanJuan, and Bellot 2014):

$$D(k \parallel k') = \frac{1}{2} \sum_{w \in W_k \cap W_{k'}} P_{TM}(w|k) \log \frac{P_{TM}(w|k)}{P_{TM}(w|k')} + \frac{1}{2} \sum_{w \in W_k \cap W_{k'}} P_{TM}(w|k') \log \frac{P_{TM}(w|k')}{P_{TM}(w|k)} \tag{5}$$

Therefore, $\hat{K}$ is the number for which the model produces the best topics (or most scattered).

4 Dispersion of Residuals

This method considers the linkage between number of topics and model fit(Taddy 2012). In particular, since the theoretical multinomial dispersion of σ^2 should be equal to one, this method consists testing for overdispersion of the variance. Therefore, if the model σ^2 is higher than 1, the true K is larger than what estimated.

5 Document-completion Held-out Likelihood

The model predictive performance can be assessed estimating the probability of a slice of the document (words or a half) on the base of another slice of the same document(Wallach et al. 2009; Roberts, Stewart, Tingley, et al. 2014). In particular, Wallach et al. (2009) formalized this estimation as follows:

$$P(w^{(2)} \mid w^{(1)}, \Phi, \alpha m) = \frac{P(w^{(2)}, w^{(1)} \mid \Phi, \alpha m)}{P(w^{(1)} \mid \Phi, \alpha m)} \tag{6}$$

Where $w^{(1)}$ is the first half and $w^{(2)}$ is the second half of the document w, $\Phi = \phi_1, ...\phi_T$ and ϕ_t is a probability vector for topic t over words, α is a concentration parameter, and m is

a base measure.

6 Frequency and Exclusivity – FREX

Airoldi and Bischof (2016) built a composite measure that consider both words to topic frequency and words to topic exclusivity, trying to avoid compensation effects among the two. In particular, the following measure is an harmonic mean of both(Airoldi and Bischof 2016):

$$FREX_{fk} = \left(\frac{\omega}{ECDF_{\phi_{.,k}}(\phi_{f,k})} + \frac{1-\omega}{ECDF_{\mu_{.,k}}(\mu_{f,k})} \right)^{-1} \tag{7}$$

Where ω is a weight to favour exclusivity over frequency (or vice-versa), $ECDF_{x,k}$ is the empirical comulative distribution function for x, $\phi_{f,k} = \frac{\beta_{f,k}}{\sum_{j=1}^{K} \beta_{j,v}}$ represents the exclusivity, and $\mu_{f,k} \equiv \beta_{f,k}$ represents the frequency (where $\beta_{f,k}$ is the rate of occurrence for word f in topic k).

7 Griffiths and Steyvers 2004

The Griffiths and Steyvers (2004) method consist in estimating $P(w \mid T)$ (where w are words in the corpus and T the number of topics) for different numbers of topics. In particular, authors suggest employing samples of the posterior distribution obtained throug Gibbs sampling(Griffiths and Steyvers 2004). In their example, for almost all T values, eight Markov chains were run (discarding the first 1,000 iterations) and 10 samples were taken from each chain (with a step of 100). Therefore, the best T correspond to the maximum value of $P(w \mid T)$.

8 Perplexity

Perplexity score is computed as follow(Blei, Ng, and Jordan 2003):

$$perplexity(D_{test}) = \exp \left[- \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right] \tag{8}$$

Where D_{test} is a test set composed by M documents d, w is a sequence of words for document d and N_d is the number of words in document d. Perplexity decreases monotonically.

9 Semantic Coherence

This metric has been introduced by Mimno et al. (2011) and it is maximized when words with higher probability in a topic tend to frequently co-occur together(Roberts, Stewart, Tingley,

et al. 2014). Topic coherence is defined as(Mimno et al. 2011):

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{9}$$

Where $D(v)$ is the number of documents with the word v, $D(v, v')$ is the number of documents containing one or more v and at least one v', and $V^{(t)}$ is a list of the M most probable words per topic t(Mimno et al. 2011).

10 Silhoutte Coefficient Metric

This metric aims to assess the quality of clusters (topics) produced by LDA, looking at similarity and dissimilarity between them(Panichella et al. 2013). The coefficient for a document d_i is:

$$s(d_i) = \frac{b(d_i) - a(d_i)}{max(a(d_i), b(d_i))} \tag{10}$$

Where $a(d_i)$ is the maximum distance of d_i from other documents in the same cluster, and $b(d_i)$ is the minimum distance from the centroids ($Centroid(C) = \sum_{d_i \in C} d_i/|C|$) of other clusters C. This metric ranges between -1 (bad clustering) and +1 (optimal clustering). The mean Silhouete coefficient can also be computed as:

$$s(C) = \frac{1}{n} \sum_{i=1}^{n} s(d_i) \tag{11}$$

11 Word and Topic Intrusion

In the word intrusion task, a human evaluator is provided with a set of high probability words (e.g. 5) for a topic(Chang et al. 2009). Beside these words, an intruder is randomly included from a set of words with low probability with respect to that topic. For example

Set of high probability words: dog, cat, horse, pig, cow

Inclusion of the intruder word: dog, cat, horse, apple, pig, cow

As the coherence of this illustrative topic example is high, apple is easily identified as the intruder. However, there are cases where coherence is not so evident, e.g.: car, teacher,

platypus, agile, blue, Zaire. Therefore, the model precision (MP) is evaluated as:

$$MP_k^m = \frac{\sum_s 1(i_{k,s}^m = \omega_k^m)}{S} \tag{12}$$

Where ω_k^m is the index of the intruder word for the k_{th} topic and model m, $i_{k,s}^m$ represent the intruder selected by individual s among the words generated for topic k_{th}, and S is the sum of individuals.

Similarly, in the topic intrusion task, a human evaluator is provided with a set of high probability topics for a document, then an intruder topic is randomly added(Chang et al. 2009). The coherence evaluation task follows the structure of word intrusion one. The results of this task is then employed to generate the topic log odds (TLO):

$$TLO_d^m = \frac{\sum_s \log \hat{\theta}_{d,j_{d,\star}^m}^m - \log \hat{\theta}_{d,j_{d,s}^m}^m}{S} \tag{13}$$

Where j_d^m is the true intruder for document d in model m, $j_{d,s}^m$ is the intruder selected by individual s, and θ is the probability assigned.

References

Airoldi, Edoardo M. and Jonathan M. Bischof. "Improving and Evaluating Topic Models and Other Models of Text". In: *Journal of the American Statistical Association* 111.516 (Oct. 2016), pp. 1381–1403. ISSN: 1537274X. DOI: 10.1080/01621459.2015.1051182.

Arun, R et al. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations". In: *Pacific-Asia conference on knowledge discovery and data mining*. 2010, pp. 391–402.

Blei, David M, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

Cao, Juan et al. "A density-based method for adaptive LDA model selection". In: *Neurocomputing* 72.7-9 (Mar. 2009), pp. 1775–1781. ISSN: 09252312. DOI: 10.1016/j.neucom.2008.06.011.

Chang, Jonathan et al. "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Advances in neural information processing systems*. 2009, pp. 288–296. URL: http://rexa.info.

Deveaud, Romain, Eric SanJuan, and Patrice Bellot. "Accurate and effective latent concept modeling for ad hoc information retrieval". In: *Document numérique* 17.1 (2014), pp. 61–84.

Griffiths, Thomas L and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

Mimno, David et al. "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

Panichella, Annibale et al. "How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms". In: *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press. 2013, pp. 522–531.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, et al. "stm: R package for structural topic models". In: *Journal of Statistical Software* 10.2 (2014), pp. 1–40.

Taddy, Matt. "On estimation and selection for topic models". In: *Artificial Intelligence and Statistics*. 2012, pp. 1184–1193.

Wallach, Hanna M et al. "Evaluation Methods for Topic Models". In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 1105–1112. DOI: 10.1145/1553374.1553515. URL: https://doi.org/10.1145/1553374.1553515.