An Information Theoretic Approach to Measuring Text Interestingness

Anonymous Author(s)
Affiliation
Address
email

Abstract

The abstract paragraph should be indented 1/2 inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

2 Our Approach

As a starting point of our model, we assume that we can describe any word with a probability distribution over a fixed set of topics - e.g. learned from a topic model. Now, given a small set of words, we in effect have a set of topic distributions. Let us denote the set of words by $S = \{w_1, ..., w_k\}$, and the distribution of a word w by P_w. We now ask what is the diversity of the set of probability distributions $\mathcal{P}_S = \{P_{w_1}, ..., P_{w_k}\}$? To make this task somewhat more concrete, we make some assumptions of what we expect from the measure. First, if the set $\mathcal{P}_S$ only contains one element, then it is not diverse. Moreover, we have a certain fixed *prior* distribution P, which we can think of as the overall distribution of topics. If a distribution P_{w_i} is equal to P, then it does not carry any information, so it should not have any impact on the diversity of $\mathcal{P}_S$. In accordance with those assumptions, we will will introduce some basic notation. Let $D_{KL}(.\|.)$ denote the Kullback-Leibler Divergence.

Definition 1 *Given a distribution P_w, its* **importance** *with respect to a prior distribution P is defined as*

$$D_w = D_{KL}(P_w \| P).$$

Definition 2 *Given a set of distributions $\mathcal{P}_S = \{P_{w_1}, ..., P_{w_k}\}$ and a prior P, we define a mixture distribution P_S as*

$$P_S = \sum_{i=1}^{k} d_{w_i} P_{w_i},$$

where $d_{w_i} = \frac{D_{w_i}}{\sum D_{w_j}}$ are the normalized importances.

Essentially, P_S is the weighted average of the set $\mathcal{P}_S$, where the weights are chosen according to the importances. Now, we can define the diversity measure.

Definition 3 *We define the Jensen-Shannon Information Diversity of a set of distributions $\mathcal{P}_S$ with respect to prior P as*

$$D_S = \sum_{i=1}^{k} d_{w_i} D_{KL}(P_{w_i} \| P_S),$$

where d_{w_i} and P_S are as in the previous definition.

This definition is closely related to the *general Jensen-Shannon Divergence*, defined in [topsoe]. Another interesting theoretical property of Jensen-Shannon Information Diversity is that it can be interpreted as a generalization of Shannon entropy as a population diversity measure, however we will not go into this here any further.

Applying this general model to the natural language requires some additional considerations. First, how do we obtain a topic ditribution for a given word, and what is the prior topic distribution. As the terminology suggests, we want to rely on a topic model. Specifically, suppose we have a set T of documents, sentences or paragraphs of any length, in which we want to find the most diverse examples. Additionally, we have a large set D of documents which has content *similar* to T. We will train a topic model on D, and use it analyzing T. Of course, both sets could be identical, but T may, for example, be a set of sentences or short messages, which not suitable for training a good topic model on. In those cases, as we will see, the fact that T and D do not need to be identical will give our method a distinct advantage over competing approaches. The key component we will need from the topic model is the word-topic matrix M, where M_{ij} is the number of times i-th word was assigned the j-th topics. So, if we probabilistically normalize the i-th row of M, it will be the topic assignment distribution for the i-th word in our dictionary. This is a good candidate for a topic distribution to use in our model. Similarly, we obtain a prior topic distribution by computing the proportions of overall topic assignments. This corresponds to summing up matrix M along its columns, and then normalizing the resulting vector.

The proposed choice of topic distributions for words has some shortcomings. First, consider a word w which occurs only once (or very few times in the entire set D. Then, its topic distribution will be very concentrated. Assuming that the prior distribution is close to uniform, that will give a very high importance to w, just because there is not enough datapoints in our set to determine its *true* topic distribution.

Furthermore, if we interpret a topic distribution as representing the *meaning* of a word, then having a fixed distribution for each word would not be able to reflect the fact that words often have different meanings in different contexts. In effect, all of the meanings would be combined together, and likely the most frequent one would dominate, distorting the results.

3 Experiments

3.1 Data Sets

- **iPhone cases:**
- **NSF:**

3.2 Baselines

Diversity:

- **Shannon Entropy:**
- **Topic Diversity:** [**?**], [**?**]
- **Word frequency as distribution:**

Classification:

- **Bag of words (BOW):**
- **Latent Semantic Indexing (LSI):**
- **Recursive Auto Encoders (RAE):**

3.3 Results

4 Conclusions

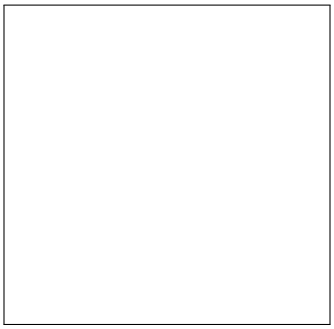

Figure 1: Sample figure caption.

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)