Interestingness from Text Diversity

Michal Derezinski

September 2, 2014

Abstract

1 Introduction

Measuring diversity in text has been previously achieved using topic modeling on documents. In this approach, we are given a collection of documents, from which we want to select the ones that are most diverse. This can be achieved with good results by performing topic modeling on those documents. Having a topic distribution for a document, we can use a measure of diversity on that distribution, e.g. Rao Diversity proposed in []. We consider a somewhat different problem, where instead of documents, we have titles or sentences - short sequences of usually 5 to 15 words. Now, performing topic modeling directly on such data is not effective, because a single item does not have sufficiently many words. One solution to this problem is to group the items using some relevant meta-data information, obtaining larger collections of words, that can be used for training a topic model. However, if we have to train the topics on the same data, that needs to be classified, then this approach will not support online prediction, where new items come in one by one, which is desirable in many practical applications. We propose a model which addresses all of those issues, while also providing new insight into human cognitive process.

2 Information Diversity

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. using a word-topic matrix of some 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]. Its information-theoretic interpretation will help us understand why it can be an appropriate measure of diversity. Let us consider an information source which selects one topic from a fixed set, each selection being independent from the previous ones. The selection is a two-step process:

1. first, one of the probability distributions is chosen from the set $\mathcal{P}_S$, where P_{w_i} is chosen with probability d_{w_i};
2. then, the output topic is drawn from that distribution.

Suppose, that we want to design an encoder A_1 for the topics, that is optimal for this source (in terms of shortest average code length), assuming that we know the mechanism of the source (including the P_{w_i}'s and d_{w_i}'s), such that the encoder sees the source only as a black box outputting the topics. Denote the shortest average code length achievable by A_1 as L_1. Now, consider a second encoder A_2 with the same goal, but this time it can look *into* the source and adjust its encoding based on which distribution P_{w_i} was selected by the source (and we assume that a hypothetical decoder can also *look up* the word w_i). Denote the shortest average code length in this instance as L_2. As discussed in [topsoe], the general Jensen-Shannon Divergence D_S is equal to $L_1 - L_2$. In other words, it describes how much information is contained in the selection of distribution P_{w_i} within the information source. For example, in the boundary case where all of the distributions in $\mathcal{P}_S$ are identical, then $L_1 = L_2$ because the initial selection is irrelevant and the set $\mathcal{P}_S$ is not diverse.

Building on this interpretation, we will discuss a more detailed thought experiment, which will argue for the choice of weights d_{w_i}. Consider a set of words as a sentence that is supposed to convey certain information. A word w_i is described by the topic distribution P_{w_i}, so the amount of information needed to describe it is the divergence of P_{w_i} from the Observer's prior distribution P, which is precisely D_{w_i}. We can think of P_{w_i} as the distribution of Observer's thoughts, given that they just saw the word w_i. So, if for example this distribution happens to be equal to the prior, this means that the Observer's reaction is independent from whether or not w_i was read, in which case we conclude, that there is no information there.

To encode the information contained in the sentence, we can describe each word using D_{w_i} bits, and concatenate them together. We may now ask what is the topic distribution for the Observer reading this code bit by bit (or, say, randomly sampling bits from it), assuming that at each step they use the distribution corresponding to the word from which that bit comes from. Clearly, it will be the mixture distribution P_S from Definition 2 and, moreover, what we have just described is essentially an instance of the information source model presented earlier.

The diversity measure we are describing works in a different setting than is typically assumed. The standard model for computing diversity is derived from analysing a population of living organisms, each being assigned a label from a fixed set of species. In our case, the species are represented by topics and the organisms are represented by words. The key difference, however, is that we do not force a fixed assignment, but instead give a topic distribution for each word. This can be reduced to a regular population model in the case where every topic distribution has singleton support (set of topics with non-zero probabilities). It is reasonable to ask, then, how does our measure compare to other population diversity measures. As it turns out, it is simply a generalization of the Shannon entropy, which is one of the canonical approaches to population diversity.

Proposition 4 *Let $S = \{P_{w_1}, ..., P_{w_k}\}$ be a set of distributions over topics, such that each has a singleton support. Assume that the prior topic distribution P corresponding to S is a uniform distribution. Then,*

$$D_S = H(P_S),$$

where P_S corresponds to the overall topic distribution in the set S, and H denotes Shannon entropy.

3 Population Diversity

The observation in Proposition 4 motivates us to look more closely into what type of properties do we expect from a population diversity measure in the case where an organism is assigned a probability distribution over species. In the measure we propose above, if a distribution is identical with the observer's prior distribution, then it gets zero weight, and so has no impact on the diversity. This makes sense for text, because we evaluate only the information it conveys. Think of words, that play a mainly syntactic role (instead of semantic) in a sentence. Our estimation of sentence diversity intuitively should not depend on whether a given language has more syntactic words than other. However, in the case of a population of organisms, for example, we can think of the distributions not as probabilities, but as proportions. In this case, let us imagine a diverse population of 10 organisms, measured against a uniform prior over species. If we were to add 90 organisms with uniform distribution assigned to each, how should this change the perceived diversity of the population? One way of looking at this is that now 90% of the organisms have identical characteristics, which ought to indicate low diversity. We will now present an alteration of our diversity measure that accomodates this intuition. For consistency, let us keep the notation from the previous definitions, although the word analogy may be less appropriate here. We will once again try to encode the set $S = \{P_{w_1}, ..., P_{w_k}\}$, with respect to a prior P. For simplicity, suppose that $\{w_1, ..., w_k\}$ is a binary code. Once again, we refer to the observation that encoding the information in P_{w_i} requires D_{w_i}, which means the hypothetical corresponding code would contain $2^{D_{w_i}}$ words (ignoring the non-integrality of D_{w_i}). Denote the set of those words as $V_{ij} = \{v_{i1}, ..., v_{il_i}\}$. As the complete code, then, we will use $C = \{w_i v_{ij} |\ 1 \leq i \leq k,\ 1 \leq j \leq l_i\}$. Like before, we can now describe an instantiation of the information source model for the general Jensen-Shannon divergence. It will simply draw a random word u from C uniformly, and then draw a topic from the distribution P_{w_i}, where w_i is a prefix of u. This corresponds to using diversity as in Definition 3, but with the weights defined as

$$d'_{w_i} = \frac{2^{D_{w_i}}}{\sum_{j=1}^{k} 2^{D_{w_j}}}.$$

We will call this the Jensen-Shannon Population Diversity. Notice, that in this case all of the elements in the population will have non-zero weights, no matter what their distribution is. This definition also has some other nice properties when applied to standard fixed-assignment populations.

To present them, we will first discuss a generative population model that incorporates the concept of a prior species distribution. Suppose we want to measure the diversity of a population S within a larger universe U of organisms assigned species from the set T. We will assume that S was generated from U as follows: We draw an element u uniformly from U. Let $t \in T$ be the species assignment for u. Next, we choose to add u to S with probability $p_{S,t}$ (depending only on the species assignment of u). For the diversity analysis, we will assume that size of the population is large enough that it's species distribution is has converged to the limit, and that the universe is large enough (compared to the population) that sampling does not noticeably affect its species distribution. We will now postulate the following general properties that a diversity measure should have in this model:

1. Given a fixed set of probabilities $p_{S,t}$ for the generative process, the population diversity does not depend on the species distribution of the universe.

2. If the species distribution of the universe is uniform, then we can revert to a standard model of population diversity. In our case, we will use Shannon entropy.

Remark 5 *A population that is uniformly sampled from the universe has highest diversity.*

This is a simple consequence of the two properties. In fact, we can say something much more precise.

Proposition 6 *Let S be a population sampled within universe U, where P_S and P_U are the species distributions of S and U, respectively. Denote T as the set of species. Let Q be a species distribution such that*

$$Q(t) = \frac{P_S(t)(P_U(t))^{-1}}{\sum_{v \in T} P_S(v)(P_U(v))^{-1}}.$$

Then, the diversity of S within U is equal to $H(Q)$.

We can now come back to how this relates to the Jensen-Shannon Population Diversity. Interestingly, it turns out that it is a direct generalization of the generative population diversity model.

Proposition 7 *Let $S = \{P_{w_1}, ..., P_{w_k}\}$ be a set of distributions over species, such that each has a singleton support. Let P be the prior species distribution corresponding to S. Then, the Jensen-Shannon Population Diversity is equal to the diversity of S as a population within a universe with species distribution P.*

4 Experiments

Measuring diversity in text has been previously achieved using topic modeling on documents. In this approach, we are given a collection of documents, from which we want to select the ones that are most diverse. This can be achieved with good results by performing topic modeling on those documents. Having a topic distribution for a document, we can use a measure of diversity on that distribution, e.g. Rao Diversity proposed in []. We consider a somewhat different problem, where instead of documents, we have titles or sentences - short sequences of usually 5 to 15 words. Now, performing topic modeling directly on such data is not effective, because a single item does not have sufficiently many words. One solution to this problem is to group the items using some relevant meta-data information, obtaining larger collections of words, that can be used for training a topic model. However, if we have to train the topics on the same data, that needs to be classified, then this approach will not support online prediction, where new items come in one by one, which is desirable in many practical applications. We propose a model which addresses all of those issues, while also providing new insight into human cognitive process.

5 Concluding Remarks

Measuring diversity in text has been previously achieved using topic modeling on documents. In this approach, we are given a collection of documents, from which we want to select the ones that are most diverse. This can be achieved with good results by performing topic modeling on those documents. Having a topic distribution for a document, we can use a measure of diversity on that distribution, e.g. Rao Diversity proposed in []. We consider a somewhat different problem, where instead of documents, we have titles or sentences - short sequences of usually 5 to 15 words. Now, performing topic modeling directly on such data is not effective, because a single item does not have sufficiently many words. One solution to this problem is to group the items using some relevant meta-data information, obtaining larger collections of words, that can be used for training a topic model. However, if we have to train the topics on the same data, that needs to be classified, then this approach will not support online prediction, where new items come in one by one, which is desirable in many practical applications. We propose a model which addresses all of those issues, while also providing new insight into human cognitive process.

Appendix A.