Avaliação automática de modelos de tópicos a partir da métrica *Normalized Pointwise Mutual Information* (NPMI)

JOSÉ ROBSON DA SILVA ARAUJO JUNIOR

No contexto de modelagem de tópicos, a escolha de um melhor modelo é, muitas vezes, inviável de ser feita manualmente. Dessa forma, nesta pesquisa, fez-se necessário estabelecer uma métrica que pudesse ajudar a realizar uma avaliação automática. Para isso, recorreu-se à métrica *Normalized Pointwise Mutual Information* (NPMI), já proposta e aplicada anteriormente para a avaliação de modelos de tópicos [2].

O presente documento tem como intuito apresentar, de maneira resumida, as intuições associadas ao NPMI e como o valor dessa medida foi utilizado para atribuir um *score* para cada modelo de tópicos gerado na pesquisa.

1 PMI (*POINTWISE MUTUAL INFORMATION*)

Pointwise Mutual Information (PMI) é uma medida que quantifica a associação entre duas palavras x e y. O seu cálculo utiliza as probabilidades de ocorrência dessas palavras ($P(x)$ e $P(y)$, respectivamente) e, além disso, a probabilidade de ambas ocorrerem em um mesmo documento ($P(x, y)$). A Equação 1 demonstra o cálculo de PMI, utilizando o logaritmo na base 2:

$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{1}$$

É importante ressaltar que, em geral, o cálculo das probabilidades de ocorrências de termos é feito externamente ao conjunto de dados utilizado. Algumas abordagens, por exemplo, utilizam páginas da Wikipedia[1] para determinar esses valores [2]. Para os propósitos desta pesquisa, no entanto, o cálculo das probabilidades foi feito a partir dos documentos do próprio *corpus*.

Além disso, costuma-se preestabelecer uma janela de distância limite entre palavras para se considerar que elas coocorrem. Dessa forma, se duas palavras ocorrem simultaneamente em um documento, mas estão muito distantes entre si, não se considera esse caso uma coocorrência. Como os documentos sob análise neste trabalho são poemas curtos e a ordem das palavras não é considerada pelo algoritmo de modelagem de tópicos empregado, resolveu-se considerar a janela de distância do tamanho de cada documento. Assim, sempre que duas palavras estão presentes em um mesmo texto, considera-se que elas coocorrem.

Aqui, considera-se que a probabilidade de uma palavra ocorrer equivale à razão entre o número de documentos em que ela está presente e o total de documentos. Considerando que uma palavra x ocorre em 25 documentos dentre 100, por exemplo, teríamos $P(x) = 25/100 = 0{,}25$. A probabilidade de duas palavras x e y coocorrerem ($P(x, y)$) é definida analogamente, dividindo-se a quantidade de documentos que contêm ambas as palavras x

[1] http://en.wikipedia.org/

e y pelo total de documentos. Dessa forma, apresentam-se, na Tabela 1, alguns exemplos retirados a partir do conjunto de dados da pesquisa:

Tabela 1. Termos e probabilidades retiradas do *corpus* da pesquisa

Termo(s)	**Probabilidade**
"canção"	7/684
"exílio"	4/684
"canção" e "exílio"	4/684

Consequentemente, o cálculo de PMI("canção", "exílio") é feito como apresentado na Equação 2:

$$\text{PMI}(\text{“canção”}, \text{“exílio”}) = \log_2 \frac{P(\text{“canção”}, \text{“exílio”})}{P(\text{“canção”})P(\text{“exílio”})} = \log_2 \frac{4/684}{(4/684) \cdot (7/684)} \approx 6{,}610 \tag{2}$$

2 NPMI

Como o PMI traz resultados que não variam em uma escala específica, uma versão normalizada foi proposta para torná-los comparáveis entre si: o *Normalized Pointwise Mutual Information* (NPMI) [1], cujo cálculo está apresentado na Equação 3.

$$\text{NPMI}(x, y) = \frac{\text{PMI}(x, y)}{-\log_2 P(x, y)} \tag{3}$$

O resultado do NPMI é sempre um valor entre -1 (as palavras nunca ocorrem simultaneamente) e 1 (as palavras sempre ocorrem simultaneamente). Para o exemplo apresentado anteriormente, tem-se NPMI("canção", "exílio") apresentado na Equação 4:

$$\text{NPMI}(\text{“canção”}, \text{“exílio”}) = \frac{\text{PMI}(\text{“canção”}, \text{“exílio”})}{-\log_2 P(\text{“canção”}, \text{“exílio”})} \approx \frac{6{,}601}{-\log_2 (4/684)} \approx 0{,}891 \tag{4}$$

Esse resultado sugere, portanto, que as palavras "canção" e "exílio" estão fortemente associadas entre si. Isso faz sentido, considerando-se que sempre que "exílio" ocorre, "canção" também ocorre (mas não necessariamente o contrário).

3 *SCORE* BASEADO EM NPMI PARA MODELOS DE TÓPICOS

Com base no NPMI, definiu-se um cálculo de *score* para um dado tópico T: a média dos valores de NPMI para cada par de termos p_i e p_j do tópico. Nesse caso, limitou-se a utilizar as dez palavras mais frequentes do tópico. Esse cálculo está representado mais formalmente na Equação 5:

$$\text{score}(T) = \text{média}\{\text{NPMI}(p_i, p_j); i, j \in 1...10, i < j\} \tag{5}$$

Finalmente, a definição de *score* de um modelo M é a média ponderada dos score(T_i) para cada um dos t tópicos T_i que compõem o modelo. A escolha da média ponderada foi feita tendo em vista compensar as divergências em tamanho dos grupos formados, já que um tópico com poucos representantes, mas com *score* alto, poderia influenciar muito caso se adotasse uma média aritmética, por exemplo. O cálculo de score(M) está representado formalmente na Equação 6, com N sendo o número total de documentos e n_i sendo a quantidade de documentos do i-ésimo tópico.

$$\text{score}(M) = \frac{\sum_{i=1}^{t}(n_i \cdot \text{score}(T_i))}{N} \tag{6}$$

Utilizou-se, portanto, o $score(M)$ na seleção do melhor modelo. Quanto maior o valor dessa métrica para um modelo, mais coerentes parecem ser os tópicos que o compõem e, consequentemente, mais coerente parece ser o modelo. É importante ressaltar que os resultados não indicam necessariamente a qualidade dos modelos, servindo apenas como guia na escolha dos melhores parâmetros.

REFERÊNCIAS

[1] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* (2009), 31–40.

[2] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.