B2W-Reviews01 - A Brazilian Portuguese reviews corpus

Name	B2W-Reviews01
Link	https://github.com/b2wdigital/b2w-reviews01
Title	B2W-Reviews01 - A Brazilian Portuguese reviews corpus
Presented by	Real, L. , Oshiro, M. Mafra, A.
Language	Brazilian Portuguese
Language code	pt-BR
Category	*resource*
Status	*available*
Type	*corpora*
Year	2019

B2W-Reviews01 is an open corpus of product reviews. It contains more than 130k e- commerce customer reviews, collected from the Americanas.com website between January and May, 2018. B2W-Reviews01 offers rich information about the reviewer profile, such as gender, age, and geographical location. The corpus also has two different review rates: the usual 5-point scale rate, represented by stars in most e-commerce websites, and also a 'recommend to a friend' label; a 'yes or no' question representing the willingness of the customer to recommend the product to someone else.

This corpus can be useful for several Natural Language Processing (NLP)/ Computational Linguistics (CL) tasks. The first that comes to mind is probably sentiment analysis. Sentiment analysis is the task of assigning a sentiment (or a position) to the content of a given text. For this task, B2W-Reviews01 offers the two distinct evaluation ratings. Product reviews often have complex information, related not only to the product that was purchased, but also to the online shopping experience, payment methods, or even the product delivery process. Therefore, for real world applications, dealing with topic modeling, user intent identification and feature extraction also become necessary.

Since B2W-Reviews01 offers the exact text written by users, this corpus also offers rich material for those interested on out-of-vocabulary words, slang identification, or spell-checker tasks. For those interested on socio-linguistics analysis, the present corpus offers a rich possibility of crossing reviewer information considering gender, age and geographical location. One can, for example, find easily how negative or positive reviews are distributed among age groups or which product categories receive more reviews from women or men. It is also possible to conduct a study on bias in reviews by joining and aggregating data.

Although B2W-Reviews01 is mainly a product review dataset, we believe that important insights about the current language in use in the web register can be made, since Americanas.com customers are spread throughout Brazil and have different social backgrounds.