TUTORIAL DESCRIPTION - PROPOR 2018

International Conference on the Computational Processing of Portuguese

September, 24-26, 2018, Canela, Brazil

1. Title

Mining the Opinion of Twitter Users in Portuguese with Python

2. Description

The growth of user-generated text on micro-blogs, social media, and e-commerce websites provides a massive quantity of data that allows discovering the experiences, opinions, and feelings of electors, fans, customers, and others [1]. These electronic *Word of Mouth* statements expressed on the web are prevalent in the business and service industries to enable customers to share their point of view [2]. In order to enhance the acquisition of a product or service and to improve the user satisfaction, most websites provide the opportunity for users to write reviews. On the other hand, customers identify online reviews as having a significant influence on their purchase in various economic sectors: 87% of customers consider those reviews on their purchase in the hotel sector, 84% in the travel sector, 79% for restaurants, 79% in the legal sector, 78% in the automotive sector, 76% in the medical sector, and 73% for home purchasing [3].

Since it is a rich source of real-time information, there has been an increasing interest in the scientific community to create systems capable of extracting information from this kind of data [4]. According to [5], opinion mining (OM), also known as sentiment analysis, is the field of study that analyzes people's sentiments, evaluations, opinions, reviews, attitudes, and emotions about different entities expressed in textual data. This is accomplished through the opinion classification of a document, sentence, or aspect into categories, such as: *positive*, *negative*, or *neutral*, using techniques taken from Natural Language Processing (NLP) and Data Mining (DM).

The OM techniques can be divided into machine learning (ML), lexicon-based, and hybrid approaches. The last one makes use of both ML and lexicon approaches [2], [6], [7]. The supervised ML applies classification algorithms to learn underlying patterns from example data to later attempt to classify new unlabeled data [4]. The lexicon-based approach, also known as semantic-based or symbolic-based, makes use of positive opinion words, used to express some desired states, and negative opinion words, used to express some undesired states. There are also opinion phrases and idioms which together are called *opinion lexicon*[6].

In this tutorial, the participants will learn and perform the four main steps to build an Opinion Mining application using the Python programming language. We will start by building a corpus containing opinions of consumers, written in the Portuguese language, about products and services on Twitter. After that, we will try out different pre-processing techniques available on Natural Language Toolkit (NLTK) library. In the sequence, the participants will experiment three supervised machine learning algorithms and evaluate the results using the resources from the Scikit-learn library.

3. Structure for six-hour slot

Part 1: Introduction to Opinion Mining and Sentiment Analysis – 1 hour

Part 2: Data collecting – 2 hours

Part 3: Text pre-processing – 1 hour

Part 4: Processing – 1 hour

Part 5: Evaluation – 1 hour

4. Instructors biography

Ellen Souza is a professor at the Information Systems undergraduate course at the Federal Rural University of Pernambuco (UFRPE) since 2009, teaching classes of Decision Support Systems and Advanced Topics in Artificial Intelligence, and leading the MiningBR research group (miningbrgroup.com.br) that investigates text mining applications, tools, languages and corpora for the Brazilian Portuguese language. In her PhD, Ellen also researched Swarm Optimization Clustering methods for Opinion Mining.

Lattes: http://lattes.cnpq.br/6593918610781356

Douglas Vitório concluded his bachelor's degree in Information Systems at the Federal Rural University of Pernambuco (UFRPE), building an Opinion Mining application to analyze the sentiments of Brazilian and Portuguese Twitter users as his undergraduate thesis. He is doing his master degree at the Federal University of Pernambuco (UFPE). He is also a member of the MiningBR research group(miningbrgroup.com.br) and has six published papers in the field of Text Mining with the Portuguese language, mainly in Opinion Mining or Sentiment Analysis.

Lattes: http://lattes.cnpq.br/2138402381175111

5. Technical requirements

Internet access and video projector.

The participants should bring their computers with Python 3, Anaconda, and the following Python libraries installed: Tweepy, NLTK* and Scikit-learn. They also must have a Twitter account.

*The NLTK has to be installed and downloaded.

6. References

[1] E. Marine-Roig and S. Anton Clavé, "Tourism analytics with massive user-generated content: A case study of Barcelona," *J. Destin. Mark. Manag.*, pp. 1–11, 2015.

[2] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Syst.*, vol. 89, p. , 2015.

[3] ComScore, "ComScore: cross-platform measurement company," 2016. [Online]. Available: http://www.comscore.com/.

[4] J. A. Balazs and J. D. Velásquez, "Opinion Mining and Information Fusion: A survey," *Inf. Fusion*, vol. 27, pp. 95–110, 2016.

[5] G. Li and F. Liu, "Application of a clustering method on sentiment analysis," *J. Inf. Sci.*, vol. 38, no. 2, pp. 127–139, 2012.

[6] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," vol. 2, no. 1, 2008.