Recommendation system to Academic Congresses: a text mining application

Mateus Maia*, Anderson Ara, Francisco Louzada Neto
*mateusmaia11@gmail.com

1ST CONFERENCE ON STATISTICS AND DATA SCIENCE
Salvador, Brazil, November 12-14, 2018

Abstract

Text Mining is one of the fields of research that has been growing along with the use of machine learning techniques. The work was to develop a framework for the creation of a recommendation system using the measure of similarity and the distance between the documents of a congress in mathematics and statistics. The work can be divided into stages of preprocessing, the weighting of more important terms, reduction of the dimensionality of text terms and similarity modeling. In this work, the recommendation system was implemented for a real base, which presented 288 scientific papers over the five days and that presented the best results for the indication of a more interesting work for each participant.

Introduction and Metodology

1 Introduction

In the area of Text Mining the recommendation system has been one of the main products produced over the last few years in order to be able to automatically identify a user's preferences from their data (Pazzani and Billsus, 2007). Large academic conferences have a wide variety of works that occur simultaneously, thus, the automatic and personalized recommendation of the programming for a user, from the analysis of their submitted works can provide a better experience to the person. To build the recommendation system, each activity of congress was considered as an observation, and the title + keywords + abstract was merged as corpus. Also, each participant that submitted a work had it a corpus composed by title + keywords + abstract too.

2 Methodology

In text mining, every Corpus object needs to go through a pre-processsing before starts the analysis. Each object has a specific treatment but a general workframe can be summarized

- **Stemming:** The reduction of the words to their root form. This can avoid plurals or verbs conjugations be considered as different expressions.
- **Stop Words**: Stop Words can be considered as the most commons words in a language, which means, that they add no value to context.
- **Formatting text**: this is general, and involves removing numbers and punctuation, or lowercase all words.

Other important step it's the weighting each term from the corpus. The main reason is to give more importance to terms that are specifics or central to each document. The parameter for weighting is given by the *tf.idf*, which the multiplication of the Term Frequency by the Inverse Document Frequency, that last is given by the equation:

$$idf(W) = log\frac{\#(documents)}{\#documentscontaingtheword(W)}$$

After that, the main object to input the modelling it's the matrix called **Bag of Words**. The representation it's given by Figure 1, where it's easy to see that the rows represents the documents , the columns the words, and each element is the *tf.id* value.

```
> words_dtm_matrix[1:3,1:8]
          Terms
Docs        consumpt variabilities   vitamin intraperson individu  accommod
  1001.txt 0.5868214     0.3420488 0.3420488   0.2280325 0.214588 0.1978479
  1003.txt 0.0000000     0.0000000 0.0000000   0.0000000 0.000000 0.0000000
  1005.txt 0.0000000     0.0000000 0.0000000   0.0000000 0.000000 0.0000000
```

Figure 1: Representation of the matrix Bag of Words

Next, the personal recommendation is built by computing the similarity between the document vector and the bag of words matrix.

Results

2.1 Similarity

In the Text Mining area there a lot of metrics to measure the similarity documents vectors.The standard comparison metric is cosine similarity, which is equivalent to dot product if the vectors are normalized (Huang, 2008). The similarity ranking was calculated by the equation below where the $\mathbf{d_u}$ corresponds to the document vector of the user, and $\mathbf{B}$ is the bag of words.

$$Similarity(u) = \mathbf{d_u}^T\mathbf{B}$$

3 Results

The final product it's the rank list of posters (Figure 2) that would be more interesting to the participant each day. Also a custom schedule of oral presentations (Figure 3) by day it's created, where the greener color represents the presentations that could be more similar to user.

Title	Similarity
K-MEANS E KERNEL K-MEANS: ALGORITMOS PARA CLUSTERIZAÇAO	1
AGRUPAMENTO HARD BASEADO EM KERNEL COM PONDERAÇAO AUTOMATICA DAS VARIAVEIS VIA DISTANCIAS ADAPTATIVAS PARA DADOS SIMBOLICOS DO TIPO INTERVALO	0.355485
CLUSTERIZAÇÃO APLICADA À SEGMENTAÇÃO EM MERCADOS DE PROPAGANDA	0.098408
SUPPORT VECTOR MACHINE APLICADO NO ESTUDO DE SÉRIES TEMPORAIS FINANCEIRAS	0.080409
MODELOS ESPACIAIS PARA OCORRENCIA DE DENGUE, ABORDAGEM CLASSICA VS BAYESIANA	0.078845
OTIMIZAÇAO DE DELINEAMENTOS EM BLOCOS	0.065613
ANALISE ESTATISTICA MULTIVARIADA DA PRECIPITAÇAO DOS MUNICIPIOS DO ESTADO DE SERGIPE ATRAVES DOS FATORES E AGRUPAMENTOS	0.060646
ANALISE DE DADOS LONGITUDINAIS UTILIZANDO A BIBLIOTECA LME4 DO SOFTWARE R	0.055156
LEILAO DO MENOR LANCE: ENSINO DE ESTATISTICA POR MEIO DE JOGOS	0.049839

Figure 2: Example for rank posters personalized.

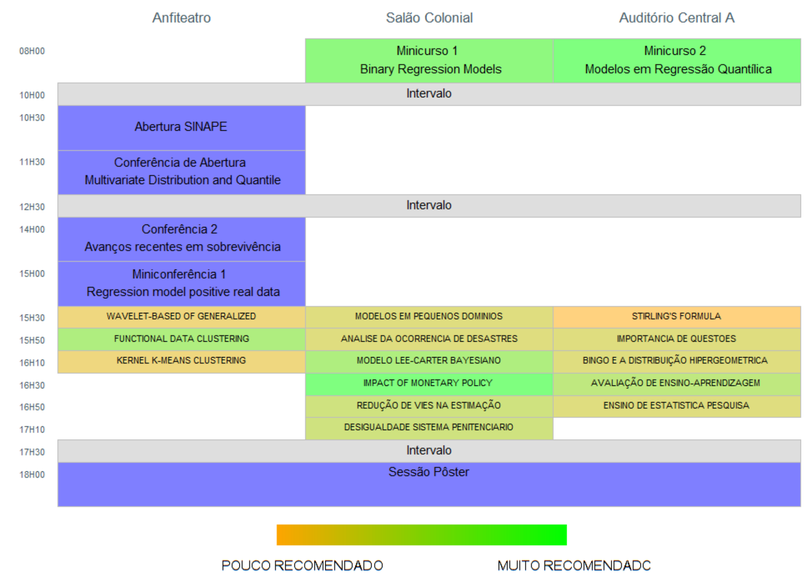

Figure 3: Example of personalized schedule for a user in Monday.

4 Conclusion

The recommendation system showed as useful to improve the participant's experience. The Corpus document could be improved by using the complete article and not just the abstract and title. Embedding strategies could be useful for more complexes analysis in order to improve the method's accuracy.

References

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56.

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer.