Universidade Federal Fluminense

Instituto de Computação
Departamento de Ciência da Computação
Prof. Aline Paes
Inteligência Artificial

Trabalho 3 (Opção 1) - 2017/1
Análise de sentimentos com Aprendizado de Máquina

A área de Análise de Sentimentos[1,2], também conhecida como Mineração de Opiniões, tem como objetivo usar técnicas de Processamento de Linguagem Natural, e áreas correlatas, para identificar estados afetivos em um texto. Usualmente, um texto é classificado como positivo (quando contém uma informação ou tendência positiva a respeito de um determinado tópico ou produto), negativo (quando contém uma informação ou tendência negativa a respeito de um determinado tópico ou produto), ou neutro, quando não há estado afetivo ou emoção associada. Nesse trabalho, você usará técnicas de Aprendizado de Máquina para criar um classificador binário (positivo/negativo; neutro/afetivo), ternário (positivo/negativo/neutro), quaternário (positivo/negativo/neutro/irrelevante) que consiga detectar o estado afetivo de textos. Neste trabalho, experimentaremos diferentes tipos de classificadores e técnicas de seleção de atributos, usando a abordagem clássica de Bag of words[4].
Para a execução do trabalho, você deverá seguir os seguintes passos:

1 - Escolher a base de dados de sentimentos, que já tenha as classes anotadas. A escolha da base de dados deve ser justificada no relatório. Algumas opções são como segue, mas outras opções também podem ser utilizadas, desde que devidamente justificadas.
http://ai.stanford.edu/~amaas/data/sentiment/
http://www.sananalytics.com/lab/twitter-sentiment/
https://inclass.kaggle.com/c/si650winter11/data
http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip
http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/
https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences
http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/twitter-sanders-apple2.zip
http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/twitter-sanders-apple3.zip
http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/epinions3.zip

2 - Tratar o texto para que ele usado como entrada para um classificador, usando a

abordagem de Bag of words. Para tanto, cada palavra pertencente ao Corpus será um atributo na base de dados, e cada sentença ou texto (de acordo com a base de dados) será um exemplo. Os tutoriais a seguir podem ser úteis para te ajudar nessa fase. A ferramenta NLTK pode te ajudar nessa fase.

http://www.nltk.org/
http://fjavieralba.com/basic-sentiment-analysis-with-python.html
http://andybromberg.com/sentiment-analysis-python/
https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words
http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

3 - Experimentar com a ferramenta de aprendizado de máquina escolhida. Você pode escolher qualquer ferramenta de sua preferência, atentando para a existência das implementações requeridas pelo trabalho. Algumas opções são:
http://scikit-learn.org/stable/
http://www.cs.waikato.ac.nz/ml/weka/
https://orange.biolab.si/

Os experimentos são como seguem:

(a) Comparar abordagens de redução de dimensionalidade (PCA [4]) e seleção de atributos (RFE [5] e Randomized Lasso [6]) . Escolha duas dentre essas opções. O objetivo dessa fase é reduzir o conjunto de atributos, ou os agrupando, ou os removendo.
(b) Seleção de parâmetros com GridSearch [7]. Algoritmos de Aprendizado de Máquina requerem diversos parâmetros a serem usados durante o processo de otimização. O método de GridSearch é útil para essa escolha de parâmetros.
(c) Seleção de classificadores. No mínimo três classificadores devem ser escolhidos, dentre as opções: Naive Bayes, SVM, Decision tree, Random Forest, AdaBoost [8,9]. Observe que as duas últimas são métodos de Ensemble [10] e é interessante que ao menos um deles seja escolhido.

Referências:

[1] Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, *11*(538-541), 164.

[2] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

[3] Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43-52.

[4] Jolliffe, I. (2002). *Principal component analysis*. John Wiley & Sons, Ltd.

[5] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

[6] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

[7] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951-2959).

[8] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

[9] Mitchell, T. M. (1997). Machine learning. WCB.

[10] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer Berlin Heidelberg.

Instruções Adicionais

- O grupo deve ser composto de 2 a 5 componentes.
- O código e um relatório descrevendo o trabalho devem ser enviados por um dos integrantes do grupo na tarefa do Classroom, até o dia 06.07. Todas as instruções para a execução do trabalho, bem como comentários ao longo do código, devem ser incluídos. Por favor, especificar o nome de todos os componentes.
- Cada integrante do grupo deve enviar um relatório como sua tarefa no Classroom, detalhando sua porcentagem de participação no trabalho e a porcentagem de participação dos demais integrantes do grupo. Exemplo (supondo grupo de 3): se você fez metade do trabalho e cada um dos seus colegas contribuiu com 25%, reporte qual foi a sua contribuição correspondente aos 50% e no que seus colegas contribuíram, com 25% cada