Note sul formato `svmlight-tk`

29/06/2024

Carlo Rosso

TABLE OF CONTENTS

1 INTRODUCTION 1

2 THE svmlight-tk FORMAT 1

3 THE Sentiment Penn Treebank FORMAT 1

4 THE CONVERSION PROCESS 1

1 Introduction

As stated in tmy intership plan, I need to convert the "Sentiment Penn Treebank" into the `svmlight-tk` format. The next step consists in training models based on the kernel methods using the just cited treebank.

In this document I will:

- describe the `svmlight-tk` format Section 2;
- describe the format of "Sentiment Penn Treebank" Section 3;
- describe the conversion process Section 4.

2 The `svmlight-tk` format

Follows the syntax of the `svmlight` format:

<line> := <target><blanket><set-of-trees>

<target> := +1 | −1 | 0 | <float>

<blanket> := " " (i.e. one space)

<set-of-trees> := <begin-tree><blank><tree><blank>..<begin-tree><blank><tree><blank><end-tree>

<begin-tree> := "|BT|"

<end-tree> := "|ET|"

<tree> := <full-tree> | <blank>

<full-tree> := (<root><blank><full-tree>..<full-tree>) | (<root><blank><leaf>)

<root> := <string>

<leaf> := <string>

—

<line> := <target>" |BT|"<full-tree>" |ET|"

<target> := <sentiment>

<full-tree> := (<root>" "<full-tree>..<full-tree>) | (<root>" "<word>)

<sentiment> := "0" | "1" | "2" | "3" | "4"

<word> := a word from the sentence

3 The `Sentiment Penn Treebank` format

Follows what I got of the `Sentiment Penn Treebank`'s format.

<line> := "("<target><blank><set-of-trees>")"

<set-of-trees> := <node> | "("<sentiment><blank><node><blank><set-of-trees>")"

<node> := "("<sentiment><blank><word>")" | "("<sentiment><blank><set-of-trees>")"

<target> := <sentiment>

<sentiment> := "0" | "1" | "2" | "3" | "4", where 0 is the most negative and 4 is the most positive, therefore 2 is neutral;

<word> := a word from the sentence, for example "hello".

<blanket> := " " (i.e. one space)

4 The conversion process