ANDROID APP FOR AUTOMATIC WEB PAGE CLASSIFICATION

ANALYSIS OF TEXT AND VISUAL FEATURES

ERASMUS STUDENT: DIEGO SALAS UGALDE

JURY

PROF. DOUTOR FERNANDO MANUEL DOS SANTOS PERDIGÃO

PROF. DOUTOR NUNO MIGUEL MENDONÇA DA SILVA GONÇALVES

PROF. DOUTOR PAULO JOSÉ MONTEIRO PEIXOTO

JULY 2015

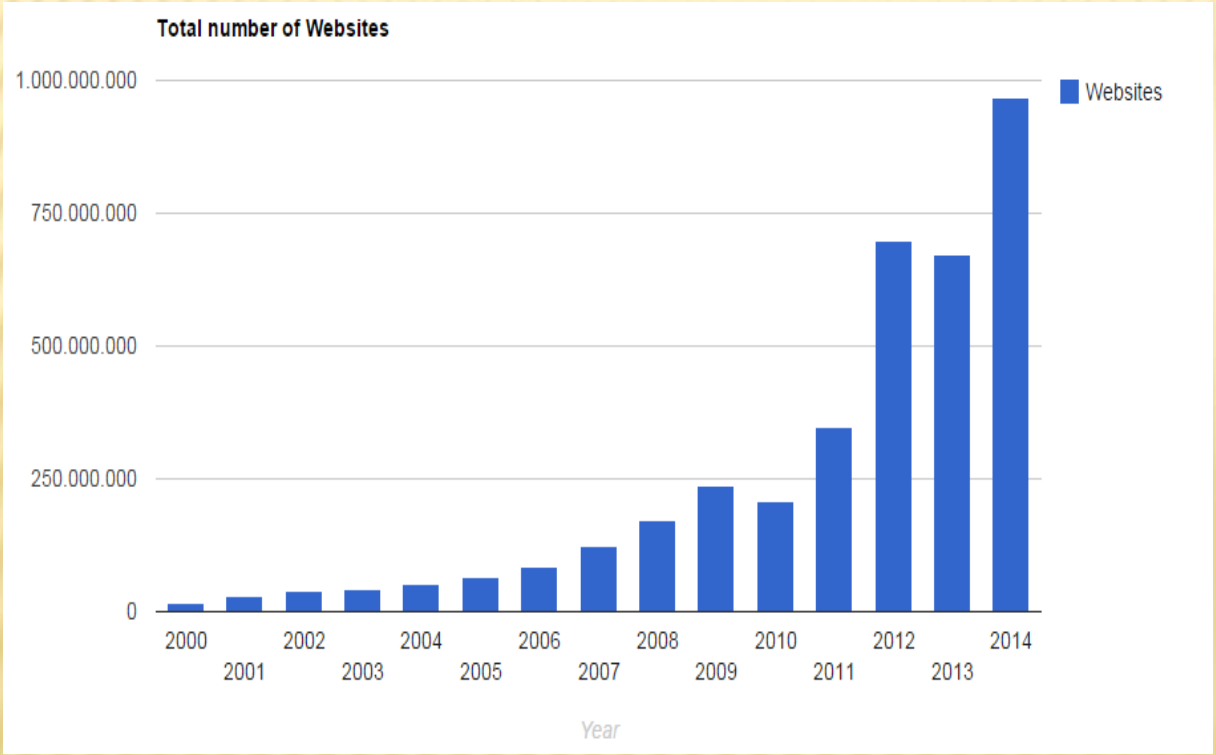
UNIVERSIDADE DE COIMBRA

OUTLINE

- INTRODUCTION
- AIMS AND OBJECTIVES
- MOTIVATION
- WEB PAGE CLASSIFICATION
  - Feature Extraction
  - Feature Fusion
  - Machine Learning ⟶ WEKA
- ANDROID:OUR APP
- RESULTS
  - RESULTS: APP PERFORMANCE
- CONCLUSIONS
- FUTURE IMPROVEMENTS

INTRODUCTION

- Information Technology is evolving fastly
  - Number of Web Pages in the World Wide Web



→ WEB PAGE CLASSIFICATION

INTRODUCTION

- New Operating Systems: ANDROID

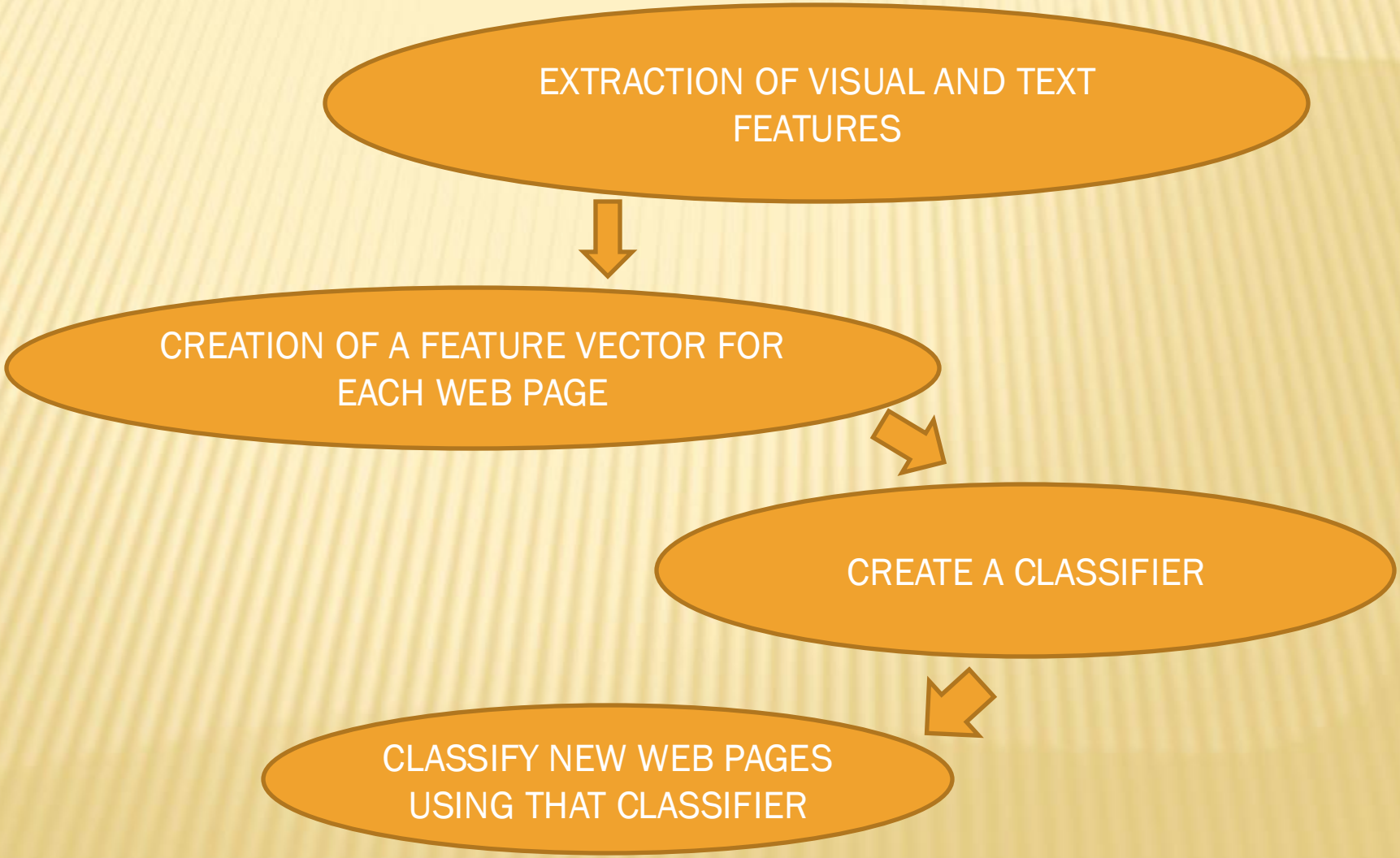
AIMS AND OBJECTIVES

- Android application
  - Web Page Classification using Visual and Text Features

MOTIVATION

- Viktor de Boer, Maarten van Someren, and Tiberiu Lupascu. ***Classifying web pages with visual features***. In WEBIST (1), pages 245-252, 2010.
- Gonçalves and Videira. ***Automatic web page classification using visual content***. In International Conference on Web Information Systems and Technologies. WEBIST,2013.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten.***The weka data mining software: an update***. SIGKDD Explor. Newsl., pages 10–18, 2009.

WEB PAGE CLASSIFICATION



WEB PAGE CLASSIFICATION

- Feature Extraction
  - Visual Features

    A vector of 160 attributes
    - Color Histogram -32 att
    - Edge Histogram -80 att
    - Gabor Features -36 att
    - Tamura Feautes -12 att

WEB PAGE CLASSIFICATION

- Feature Extraction
  - Text Features
    - TF-IDF
    - BoW (Bag Of Words)
    - Vector Space Model

WEB PAGE CLASSIFICATION

- TF-IDF
  - TF: It measures how frequently a term in a document occurs

$$TF(t) = \frac{NT_t}{T_tD}$$

$NT_t = Number\ of\ times\ a\ term\ called\ t\ appears\ in\ a\ document$

$T_tD = Total\ number\ of\ terms\ in\ the\ document$

  - IDF: It measure how important a term is

$$IDF(t) = \log \frac{TD}{ND_t}$$

TD = Total number of documents.

$ND_t = Number\ of\ documents\ with\ the\ term\ t.$

WEB PAGE CLASSIFICATION

- BoW (Bag-Of-Words)
  - Based on TF-IDF values a dictionary of words is built
- VSM (Vector Space Model)
  - Algebraic model to represent text as a vector of 160 attributes

  It is built based on the absolute frequency of a term in the BoW

WEB PAGE CLASSIFICATION

DOCUMENT 1
Where
glass
Suddenly
Because
university

DOCUMENT TO EXTRACT THE VECTOR FROM
Where
School
university
glass
university
Because
glass

The final vector would be:

[1,2,0,1,2]

WEB PAGE CLASSIFICATION

- Feature Fusion
  - Fuse visual-feature-vector and text-feature-vector
  - Vectors of 320 attributes (VISUAL+TEXT)

WEB PAGE CLASSIFICATION

- Machine Learning
  - WEKA
    - Collection of machine learning algorithms for data mining tasks
    - Free software
    - Advantages
      - Available in platforms as Android Studio

WEB PAGE CLASSIFICATION

- WEKA
  - Dataset
    - Basic concept
    - Implemented by weka.core.Instances
    - ARFF file

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
```

WEB PAGE CLASSIFICATION

- WEKA
  - Classifiers
    - Derived from the class weka.classifiers.Classifier class
    - In this work:
      - J48 - It builds decision trees using the concept of Information Entropy
      - NAIVE BAYES – Based on Bayes' theorem
      - ADA BOOST - Adaptive Boosting, machine learning meta-algorithm

ANDROID

- Mobile Operating System
  - Developed by Google
  - Based on the Linux Kernel

ANDROID: OUR APP

- Name: WebClass
- Weight: 47.08MB
- Functionality: Perform Web Page Classification with three different classifiers using Text, Visual or both features

ANDROID: OUR APP

- WEKA library
- OpenCV library to extract Visual features
- Jsoup library to extract Text features

ANDROID: OUR APP

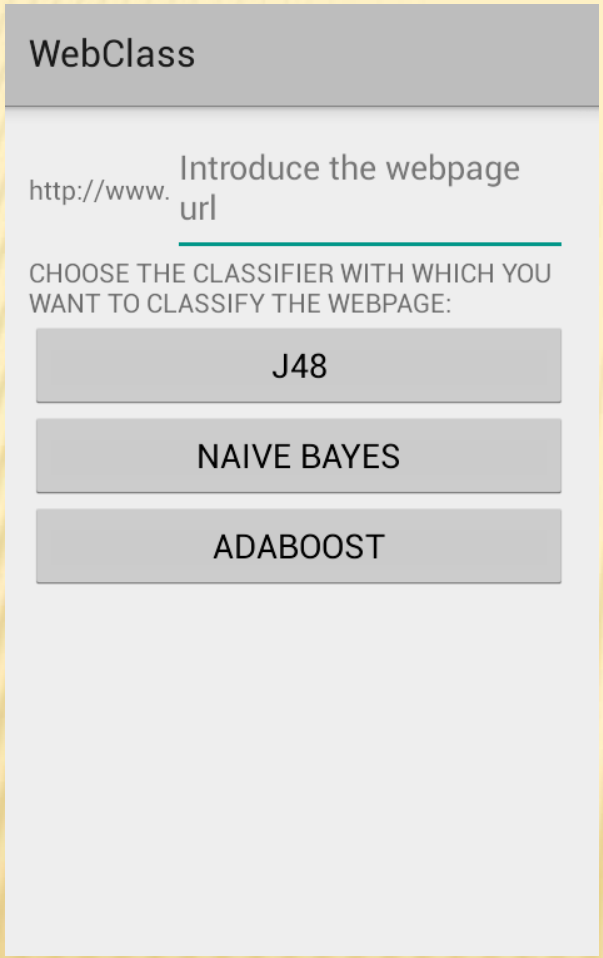
- WHEN VECTORS BUILT…
- ARFF file to classify
  - Another ARFF file will be obtained with “?” labeled

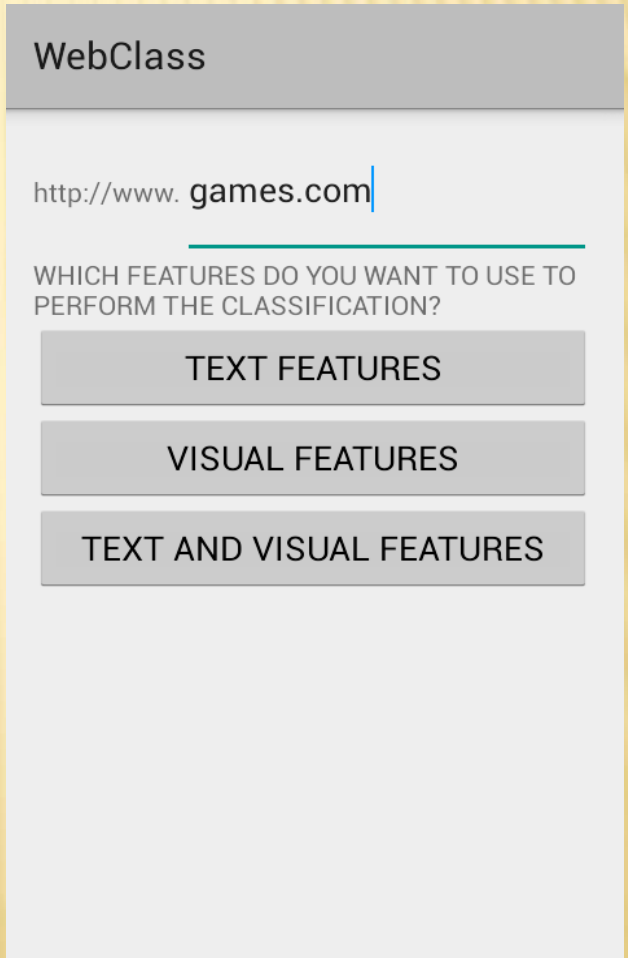
```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
4.3,3.7,1.2,0.3,?
```

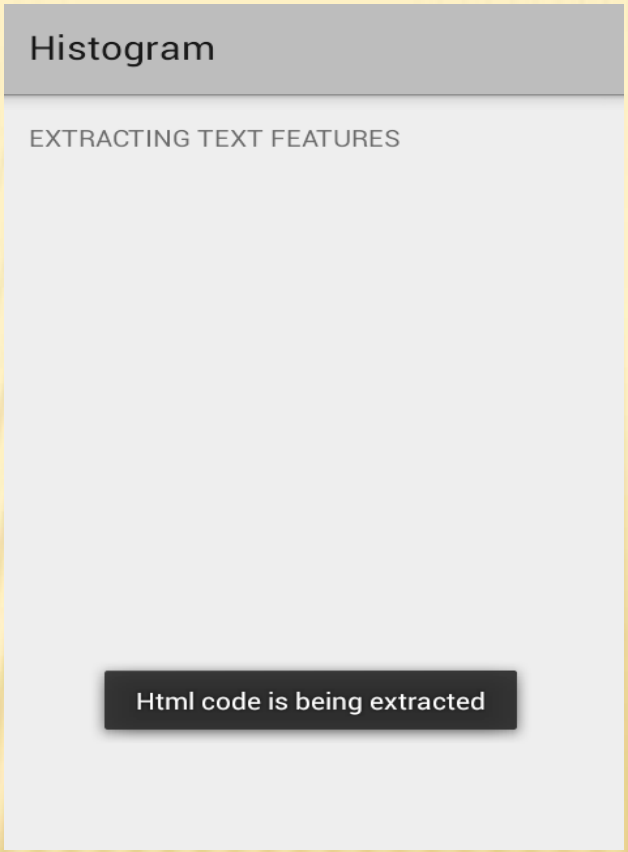
ANDROID: OUR APP

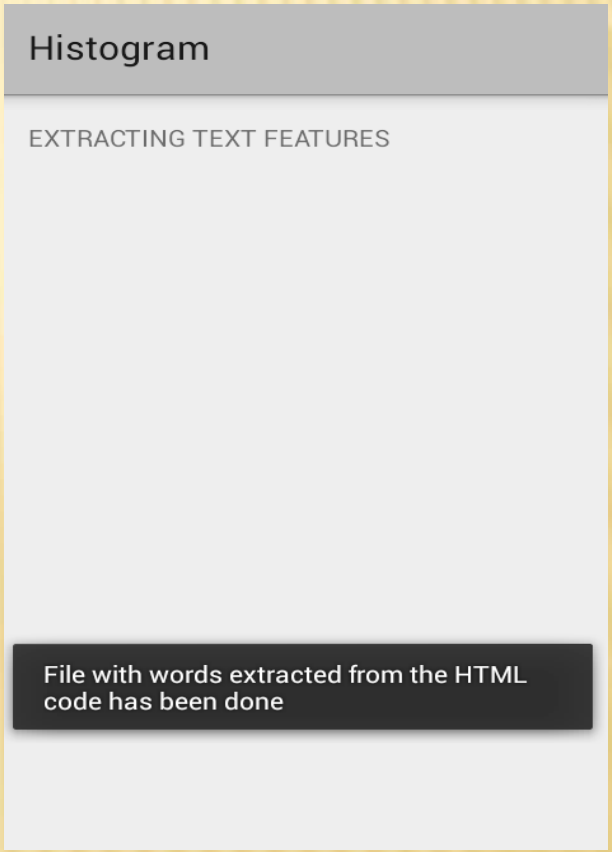




ANDROID: OUR APP

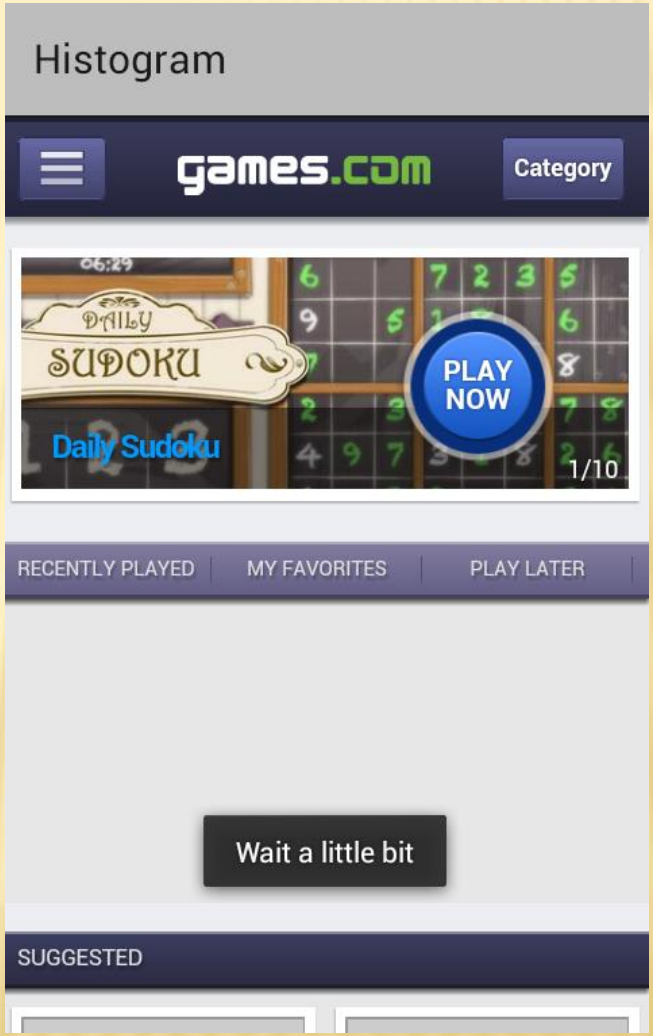
- Text features



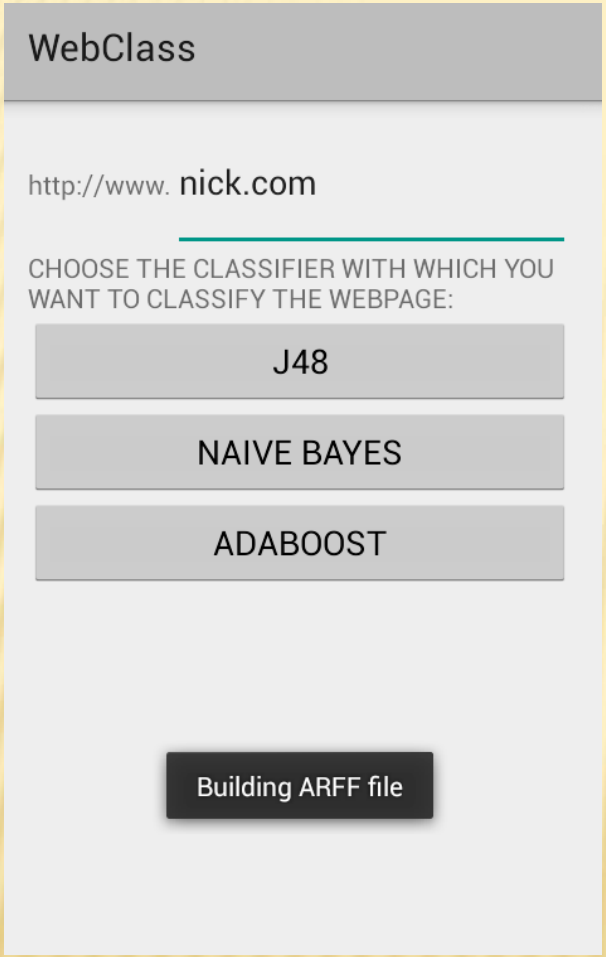


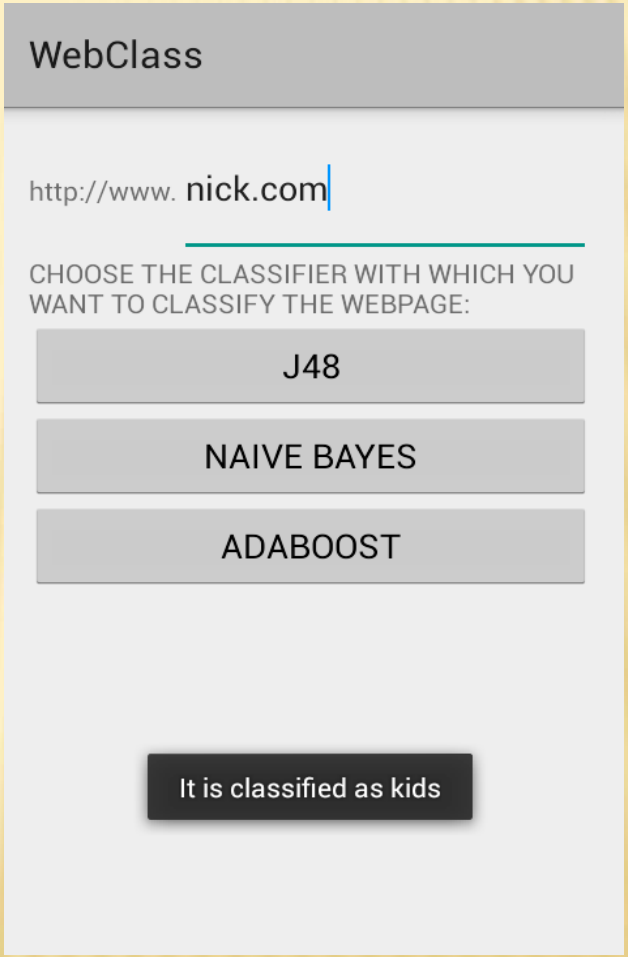
ANDROID: OUR APP

- Visual features



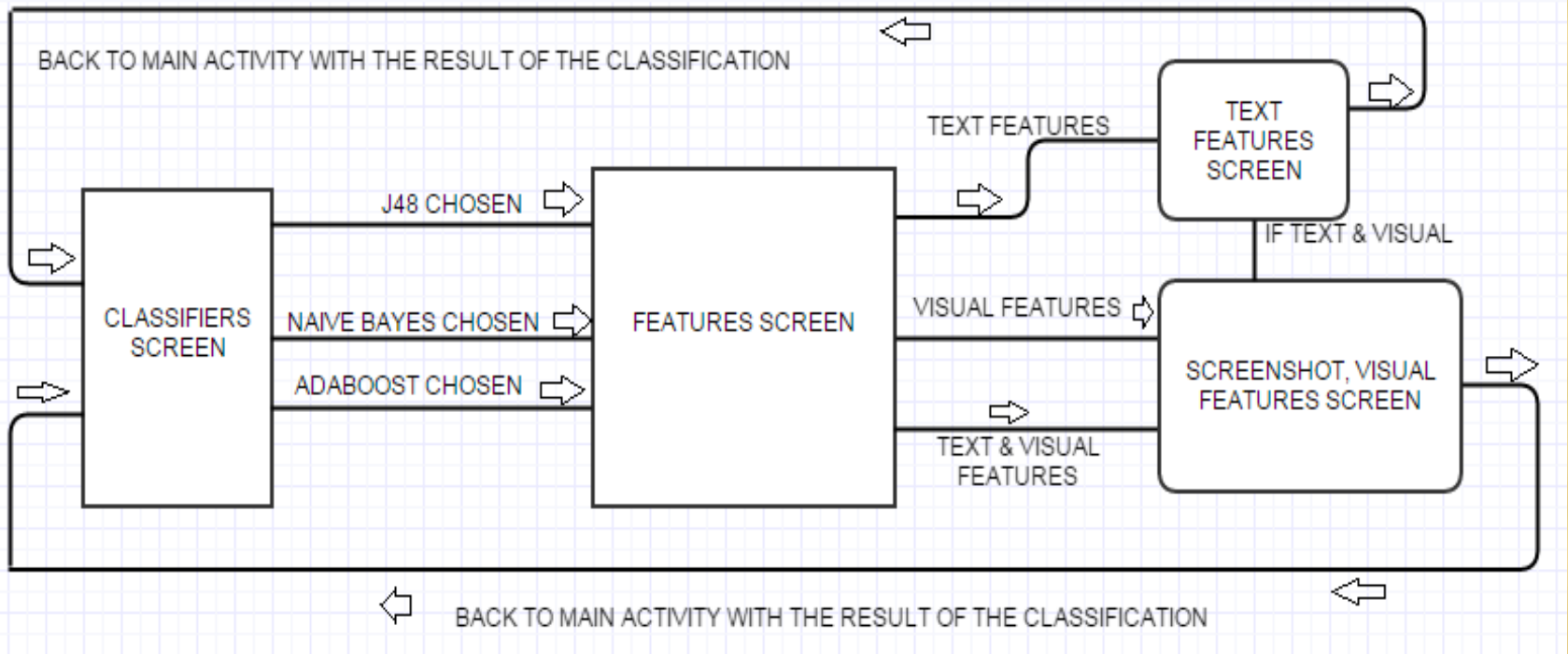
ANDROID: OUR APP





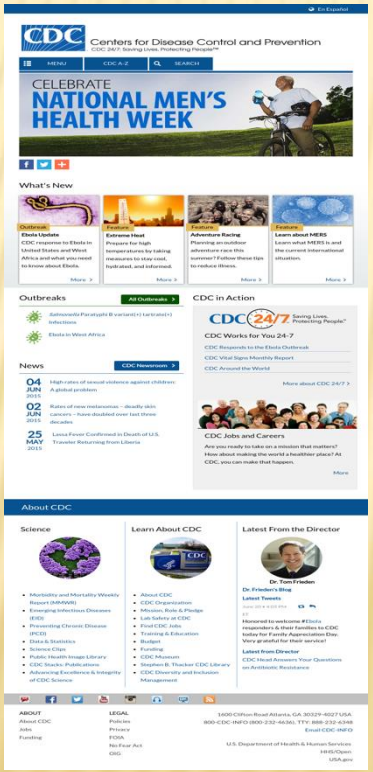
ANDROID: OUR APP

- Flow chart of the app



RESULTS

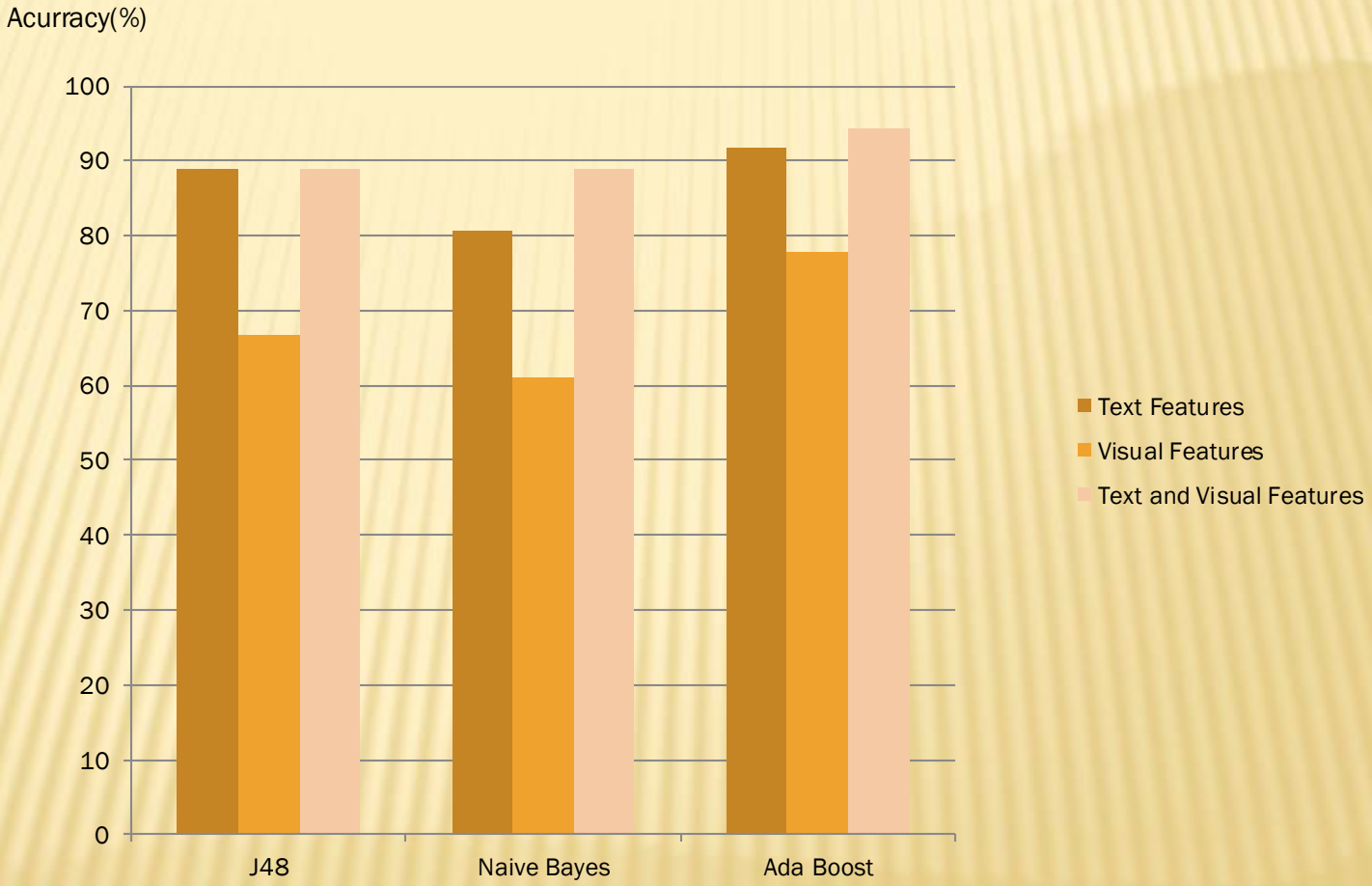
- Binary Classification: Adults & Kids
  - Adults: News, banks, universities…
  - Kids: cartoon, TV series…



RESULTS

- TRAIN AND TEST PHASE
  - TRAIN PHASE
    - 382 web pages (193 Adults, 189 Kids)
  - TEST PHASE
    - 36 web pages (18 Adults, 18 Kids)

RESULTS


Acurracy(%)
100
90
80
70
60
50
40
30
20
10
0
J48
Naive Bayes
Ada Boost
Text Features
Visual Features
Text and Visual Features

RESULTS

- J48
  - 66.67%

VISUAL FEATURES	ADULTS	KIDS
ADULTS	14	4
KIDS	8	10

  - 88.89%

TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	3	15

  - 88.89%

VISUAL AND TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	3	15

RESULTS

- NAIVE BAYES

+ 61.11%

VISUAL FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	13	5

+ 80.56%

TEXT FEATURES	ADULTS	KIDS
ADULTS	14	4
KIDS	3	15

+ 88.89%

VISUAL AND TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	3	15

RESULTS

- ADABOOST

+ 77.78%

VISUAL FEATURES	ADULTS	KIDS
ADULTS	16	2
KIDS	6	12

+ 91.67%

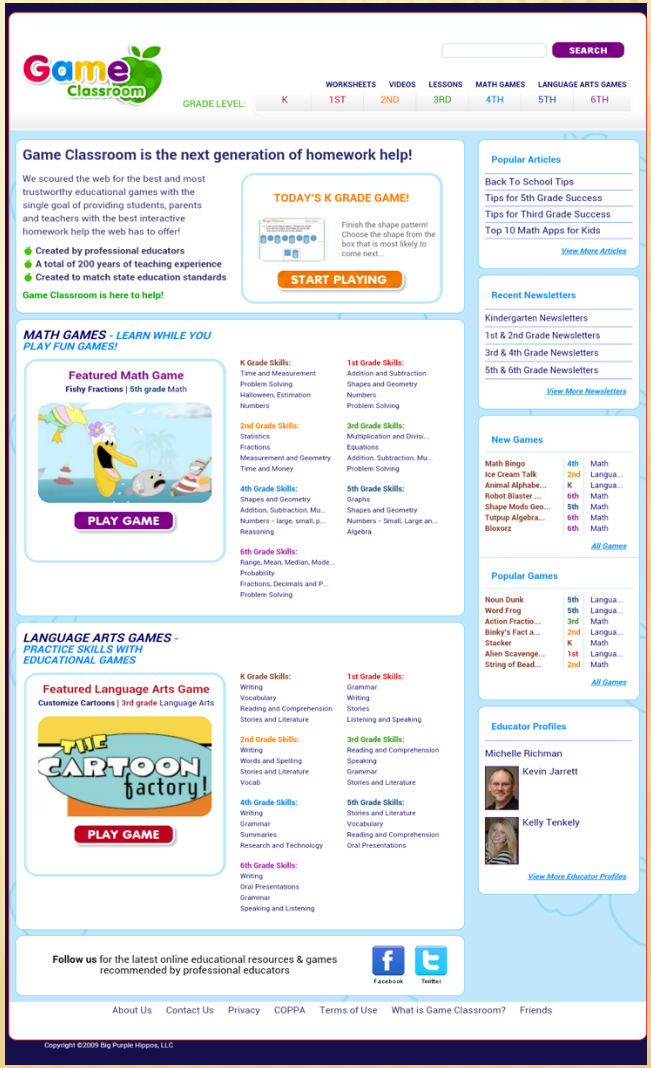
TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	2	16

+ 94.44%

VISUAL AND TEXT FEATURES	ADULTS	KIDS
ADULTS	18	0
KIDS	2	16

RESULTS

- Kids web page classified as Adults



RESULTS

- BEST RESULT
  - ADABOOST CLASSIFIER:

    ACCURACY OF 94.44% WHEN USING BOTH FEATURES

RESULTS: APP PERFORMANCE

WEB PAGE	www.nick.com	www.games.com	www.su.se
CLASSIFICATION WITH TEXT FEATURES	25sec	26sec	26sec
CLASSIFICATION WITH VISUAL FEATURES	46sec	105sec	159sec
CLASSIFICATION WITH TEXT AND VISUAL FEATURES	61s	117sec	181sec

WEB PAGE	www.nick.com	www.games.com	www.su.se
WORDS EXTRACTED FROM THE HTML CODE	430	421	188
WEIGHT OF THE SCREENSHOT	1.61KB	91.5KB	375KB

CONCLUSIONS

- VISUAL FEATURES FROM WEB PAGES IMPROVE THE CLASSIFICATION AND THEY SHOULD NOT BE IGNORED
- CLASSIFICATION OF ADULTS WEB PAGES SEEMS TO BE EASIER TO PERFORM THAN KIDS
- ONLY THE WEIGHT OF THE SCREENSHOT MATTERS FOR THE TIME OF EXECUTION OF THE APP

FUTURE IMPROVEMENTS

- Enhance the app execution time by solving the problem of the weight of the screenshots
- Add binary and multi-label classification to the app

ACKNOWLEDGEMENTS

- THANK YOU FOR YOUR ATTENTION

· U C ·

UNIVERSIDADE DE COIMBRA