Program with UIMA Framework

Annie Lukas - 2019.05.06

Text to XMI files

- Each line contains one statement, with Argumentative Discourse Unit Type (ADU)
- To convert the dataset in UIMA format (XMI) with annotated ADU

culture21-output.txt
~/wstud-visit-the-dome-ss19/data/webis-debate-16

```
Non-Argumentative       in our globalized world with an increasingly international media, beauty standards for both men and women are
becoming more internationalized - or, perhaps, westernized.
Non-Argumentative        this has included skin tone, with cultural values in many countries favouring lighter skin tones.
Non-Argumentative       one particular aspect of this has been creams aiming to whiten skin.
Non-Argumentative        this is not an issue that affects women, the key target of much of the fashion and beauty industry, exclusively
- some products are aimed at men.
Non-Argumentative        this is not a phenomenon unique to africa and communities of african origins.
Non-Argumentative        india is a large market for skin whitening creams, with a market worth just shy of half a billion us dollars
inand growing at nearly 20% per year.
Non-Argumentative        in hong kong, around a third of women have used such products.
Non-Argumentative        the issue of people seeking to look “whiter” or caucasian is not limited to any area or simply skin whitening
creams – there are also products and the possibility of surgery to mimic more european-looking eyelids available in east asia.
Non-Argumentative       west africa is a key market for these products, with over half of women in togo and nigeria claiming to use them.
Non-Argumentative        while some contain chemicals aimed at modifying the skin’s transfer of melanin, others use harsh chemicals such
as mercury, which can lead to medical conditions from eczema and infections to renal failure.
Non-Argumentative       a particular controversy has been generated by the launch of a new skin whitening product in early 2014, known
as “whitenicious”, by nigerian-cameroonian singer dencia, and her comments relating to the product, including a spat with actress lupita
nyong’o. nyong’o, who shot to fame outside her native kenya following her powerful performance in the popular and well received british-
american film “12 years a slave”, has been critical of skin lightening creams and the conception of lighter skin tones in african mind-
sets.
Non-Argumentative        in 2006, the us food and drug administration proposed a ban on over-the-counter sales of skin whitening creams.
Non-Argumentative        is this a solution for africa?
Argumentative    skin whitening creams often contain a wide variety of harmful ingredients -- in some cases , mercury .
Argumentative    these can cause various health problems ; mercury in particular causes renal , kidney , damage , major skin problems as
well as mental health issues .
Argumentative    states , throughout the world , ban consumer products because they are harmful regardless of whether this is for
consumption or for cosmetics .
Argumentative    this is just another case where that is appropriate in order to prevent the harm to health that may occur .
Argumentative    allowing the use of racial overtones -- the perception that a product will bring a person towards a `` white ideal ''
is harmful for several reasons .
Argumentative    it could cause communities to generate a form of inferiority complex , and it reinforces the structural difference
rather than aiming to minimize it .
Argumentative    while it may sound absurd , in the us darker-skinned african americans , and darker skinned latinos , are less well
educated and have lower incomes .
Argumentative    elsewhere in the western hemisphere , such as in brazil , race is seen as an issue of colour and socio-economic
background , not ancestry highlighting a much more obvious link between whitening creams and racism .
Argumentative    is it not the role of the state to reduce that discrimination , not to fuel it ?
Argumentative    banning such creams would help prevent such harmful effects by discouraging the notion that people should aim to make
themselves lighter skinned .
Argumentative    skin whitening can be seen as an attempt to fit in with a form of a neo-colonialist mind-set ; a form of cultural
imperialism driven by capitalism .
Argumentative    these products , often sold by big international fmcg , fast moving consumer goods , companies feed off a neo-
colonialist mind-set -- one of a cultural inferiority complex .
Argumentative    these products form part of the process of tying african people into a globalised consumer world where non-westerners
feel compelled to buy western products that they do n't need .
Argumentative    they are therefore kept in a colonial situation where they are dependent on the west both mentally and in terms of the
products they buy .
Argumentative    that is reason enough for nations that have been victims of colonialism by the global north to take action against
them .
```

Text to XMI Converter

- Created a Text to XMI Converter (used example), where defined:
- Input directory
- Output directory
- Collection Reader: Plain Text Reader
- AE: Dummy

```java
public class Text2XMIConverter {

    // ---------------------------------------------------------------------
    // PARAMETERS
    // ---------------------------------------------------------------------

    /**
     * The path of root directory of the files to be processed.
     */
    private static final String INPUT_COLLECTION_DIR =
            "data/webis-debate-16/";

    /**
     * The path of the XMI file of the collection reader to be used to iterate
     * over all files to be processed.
     */
    private static final String COLLECTION_READER_PATH =
            "../aitools4-ie-uima/conf/uima-descriptors/collection-readers/UIMAPlainTextReader.xml";

    /**
     * The path of the XMI file of the analysis engine to be used to process
     * the files.
     */

    private static final String ANALYSIS_ENGINE_PATH =
            "../aitools4-ie-uima/conf/uima-descriptors/primitive-AEs/template/"
            + "DummyAnalysisEngine.xml";

    /**
     * The path of the directory where the XMI files shall be written to
     */
    private static final String OUTPUT_COLLECTION_DIR =
            "data/webis-debate-16-xmi-output/";
```

XMI Output

- Split each line and extract the content to be document text
- Annotated the statement with corresponding ADU type

culture21_0-output.xmi
~/wstud-visit-the-dome-ss19/data/webis-debate-16-xmi-output


```
<?xml version="1.0" encoding="utf-8"?>
<xmi:XMI xmlns:pos="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/pos.ecore"
xmlns:tcas="http:///uima/tcas.ecore" xmlns:xmi="http://www.omg.org/XMI" xmlns:cas="http:///
uima/cas.ecore" xmlns:tweet="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/pos/
tweet.ecore" xmlns:dependency="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/
dependency.ecore" xmlns:infexba="http:///de/aitools/ie/uima/type/infexba.ecore"
xmlns:argumentation="http:///de/aitools/ie/uima/type/argumentation.ecore" xmlns:core="http:///
de/aitools/ie/uima/type/core.ecore" xmlns:type7="http:///de/tudarmstadt/ukp/dkpro/core/api/
semantics/type.ecore" xmlns:article="http:///de/aitools/ie/uima/type/article.ecore"
xmlns:supertype="http:///de/aitools/ie/uima/type/supertype.ecore" xmlns:type4="http:///de/
tudarmstadt/ukp/dkpro/core/api/metadata/type.ecore" xmlns:type2="http:///de/tudarmstadt/ukp/
dkpro/core/api/coref/type.ecore" xmlns:type8="http:///de/tudarmstadt/ukp/dkpro/core/api/
structure/type.ecore" xmlns:types="http:///de/tudarmstadt/ukp/dkpro/argumentation/types.ecore"
xmlns:morph="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/morph.ecore"
xmlns:type3="http:///de/tudarmstadt/ukp/dkpro/core/api/frequency/tfidf/type.ecore"
xmlns:type="http:///de/tudarmstadt/ukp/dkpro/core/api/anomaly/type.ecore" xmlns:type9="http:///
de/tudarmstadt/ukp/dkpro/core/api/syntax/type.ecore" xmlns:type10="http:///de/tudarmstadt/ukp/
dkpro/tc/type.ecore" xmlns:type5="http:///de/tudarmstadt/ukp/dkpro/core/api/ner/type.ecore"
xmlns:type6="http:///de/tudarmstadt/ukp/dkpro/core/api/segmentation/type.ecore"
xmlns:constituent="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/constituent.ecore"
xmlns:arguana="http:///de/aitools/ie/uima/type/arguana.ecore" xmlns:chunk="http:///de/
tudarmstadt/ukp/dkpro/core/api/syntax/type/chunk.ecore" xmi:version="2.0">
    <cas:NULL xmi:id="0"/>
    <tcas:DocumentAnnotation xmi:id="8" sofa="1" begin="0" end="169" language="x-unspecified"/>
    <argumentation:ArgumentativeDiscourseUnit xmi:id="13" sofa="1" begin="0" end="168"
unitType="Non-Argumentative"/>
    <cas:Sofa xmi:id="1" sofaNum="1" sofaID="_InitialView" mimeType="text" sofaString="in our
globalized world with an increasingly international media, beauty standards for both men and
women are becoming more internationalized - or, perhaps, westernized."/>
    <cas:View sofa="1" members="8 13"/>
</xmi:XMI>
```

Simple Pipeline

- Input: XMI Files
- Output: Statement with annotated token and part of speech tag
- Collection Reader: UIMA Annotation File Reader

```java
public class CollectionProcessor {

    // ---------------------------------------------------------------------
    // PARAMETERS
    // ---------------------------------------------------------------------

    /**
     * The path of root directory of the training files to be processed.
     */
    private static final String COLLECTION_ROOT_DIR =
            "data/webis-debate-16-xmi-output/";

    /**
     * The path of the XMI file of the collection reader to be used to iterate
     * over all files to be processed.
     */
    private static final String COLLECTION_READER_PATH =
            "../aitools4-ie-uima/conf/uima-descriptors/collection-readers/"
            + "UIMAAnnotationFileReader.xml";

    /**
     * The path of the XMI file of the analysis engine to be used to process
     * the files.
     */
    private static final String ANALYSIS_ENGINE_PATH =
            "conf/uima-descriptors/aggregate-AEs/"
            + "AggregateAE.xml";
    /**
     * The path of the directory where the XMI files shall be written to
     */
    private static final String OUTPUT_COLLECTION_DIR =
            "data/webis-debate-16-xmi-output-token-partofspeech/";
```

Aggregate Analysis Engine

- Contains
  - Sentence Splitter
  - Tokenizer
  - Lemma and Part of Speech Tagger

```
<delegateAnalysisEngine key="InfexBASentenceSplitter">
  <import location="../../../../aitools4-ie-uima/conf/uima-descriptors/
</delegateAnalysisEngine>
<delegateAnalysisEngine key="InfexBATokenizer">
  <import location="../../../../aitools4-ie-uima/conf/uima-descriptors/
</delegateAnalysisEngine>
<delegateAnalysisEngine key="TT4jLemmaAndPartOfSpeechTagger">
  <import location="../../../../aitools4-ie-uima/conf/uima-descriptors/
```

Infex BA Sentence Splitter

- Developed in the Infex BA project and improve in the Argument Analysis project later on
- Does not require any input annotations and produces sentence annotations

Infex BA Tokenizer

- Rather effective but still efficient rule-based tokenizer that was developed in the Infex BA project and improved in the Argument Analysis project later on
- Requires sentence annotations and produces token annotations

Part of Speech Tag

- POS tagging is process of marking up a word in a text as a corresponding to a part of speech, based on its definition and its context; e.g identification of words as nouns, verbs, adjectives, adverbs

Tag	Description	Example
CC	conjunction, coordinating	*and, or, but*
CD	cardinal number	*five, three, 13%*
DT	determiner	*the, a, these*
EX	existential there	*there were six boys*
FW	foreign word	*mais*
IN	conjunction, subordinating or preposition	*of, on, before, unless*
JJ	adjective	*nice, easy*
JJR	adjective, comparative	*nicer, easier*
JJS	adjective, superlative	*nicest, easiest*
LS	list item marker	
MD	verb, modal auxillary	*may, should*
NN	noun, singular or mass	*tiger, chair, laughter*
NNS	noun, plural	*tigers, chairs, insects*
NNP	noun, proper singular	*Germany, God, Alice*
NNPS	noun, proper plural	
PDT	predeterminer	
POS	possessive ending	
PRP	pronoun, personal	
PRP$	pronoun, possessive	*my, your, our*
RB	adverb	*extremely, loudly, hard*
RBR	adverb, comparative	*better*
RBS	adverb, superlative	*best*
RP	adverb, particle	*about, off, up*
SYM	symbol	*%*
TO	infinitival to	*what to do?*
UH	interjection	*oh, oops, gosh*
VB	verb, base form	*think*
VBZ	verb, 3rd person singular present	*she thinks*
VBP	verb, non-3rd person singular present	*I think*
VBD	verb, past tense	*they thought*
VBN	verb, past participle	*a sunken ship*
VBG	verb, gerund or present participle	*thinking is fun*
WDT	*wh*-determiner	*which, whatever, whichever*
WP	*wh*-pronoun, personal	*what, who, whom*
WP$	*wh*-pronoun, possessive	*whose, whosever*
WRB	*wh*-adverb	*where, when*
.	punctuation mark, sentence closer	*.;?**
,	punctuation mark, comma	*,*
:	punctuation mark, colon	*:*
(	contextual separator, left paren	*(*
)	contextual separator, right paren	*)*

Output XMI Files

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

culture21_0-output.xmi
~/wstud-visit-the-dome-ss19/data/webis-debate-16-xmi-output-token-partofspeech


```
<?xml version="1.0" encoding="utf-8"?>
<xmi:XMI xmlns:pos="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/pos.ecore" xmlns:tcas="http:///uima/tcas.ecore"
xmlns:xmi="http://www.omg.org/XMI" xmlns:cas="http:///uima/cas.ecore" xmlns:tweet="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/
type/pos/tweet.ecore" xmlns:dependency="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/dependency.ecore" xmlns:infexba="http:///
de/aitools/ie/uima/type/infexba.ecore" xmlns:argumentation="http:///de/aitools/ie/uima/type/argumentation.ecore" xmlns:core="http:///de/
aitools/ie/uima/type/core.ecore" xmlns:type7="http:///de/tudarmstadt/ukp/dkpro/core/api/semantics/type.ecore" xmlns:article="http:///de/
aitools/ie/uima/type/article.ecore" xmlns:supertype="http:///de/aitools/ie/uima/type/supertype.ecore" xmlns:type4="http:///de/
tudarmstadt/ukp/dkpro/core/api/metadata/type.ecore" xmlns:type2="http:///de/tudarmstadt/ukp/dkpro/core/api/coref/type.ecore"
xmlns:type8="http:///de/tudarmstadt/ukp/dkpro/core/api/structure/type.ecore" xmlns:types="http:///de/tudarmstadt/ukp/dkpro/argumentation/
types.ecore" xmlns:morph="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/morph.ecore" xmlns:type3="http:///de/tudarmstadt/ukp/
dkpro/core/api/frequency/tfidf/type.ecore" xmlns:type="http:///de/tudarmstadt/ukp/dkpro/core/api/anomaly/type.ecore"
xmlns:type9="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type.ecore" xmlns:type10="http:///de/tudarmstadt/ukp/dkpro/tc/type.ecore"
xmlns:type5="http:///de/tudarmstadt/ukp/dkpro/core/api/ner/type.ecore" xmlns:type6="http:///de/tudarmstadt/ukp/dkpro/core/api/
segmentation/type.ecore" xmlns:constituent="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/constituent.ecore"
xmlns:arguana="http:///de/aitools/ie/uima/type/arguana.ecore" xmlns:chunk="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/
chunk.ecore" xmi:version="2.0">
    <cas:NULL xmi:id="0"/>
    <tcas:DocumentAnnotation xmi:id="1" sofa="11" begin="0" end="169" language="x-unspecified"/>
    <argumentation:ArgumentativeDiscourseUnit xmi:id="6" sofa="11" begin="0" end="168" unitType="Non-Argumentative"/>
    <core:SourceDocumentInformation xmi:id="18" sofa="11" begin="0" end="0" uri="/home/ciso8478/wstud-visit-the-dome-ss19/data/webis-
debate-16-xmi-output/culture21_0-output.xmi" offsetInSource="0" documentSize="169" lastSegment="false"/>
    <core:Sentence xmi:id="26" sofa="11" begin="0" end="169"/>
    <core:Token xmi:id="30" sofa="11" begin="0" end="2" lemma="in" pos="IN"/>
    <core:Token xmi:id="41" sofa="11" begin="3" end="6" lemma="our" pos="PP$"/>
    <core:Token xmi:id="52" sofa="11" begin="7" end="17" lemma="globalize" pos="VVN"/>
    <core:Token xmi:id="63" sofa="11" begin="18" end="23" lemma="world" pos="NN"/>
    <core:Token xmi:id="74" sofa="11" begin="24" end="28" lemma="with" pos="IN"/>
    <core:Token xmi:id="85" sofa="11" begin="29" end="31" lemma="an" pos="DT"/>
    <core:Token xmi:id="96" sofa="11" begin="32" end="44" lemma="increasingly" pos="RB"/>
    <core:Token xmi:id="107" sofa="11" begin="45" end="58" lemma="international" pos="JJ"/>
    <core:Token xmi:id="118" sofa="11" begin="59" end="64" lemma="medium" pos="NNS"/>
    <core:Token xmi:id="129" sofa="11" begin="64" end="65" lemma="," pos=","/>
    <core:Token xmi:id="140" sofa="11" begin="66" end="72" lemma="beauty" pos="NN"/>
    <core:Token xmi:id="151" sofa="11" begin="73" end="82" lemma="standard" pos="NNS"/>
    <core:Token xmi:id="162" sofa="11" begin="83" end="86" lemma="for" pos="IN"/>
    <core:Token xmi:id="173" sofa="11" begin="87" end="91" lemma="both" pos="DT"/>
    <core:Token xmi:id="184" sofa="11" begin="92" end="95" lemma="man" pos="NNS"/>
    <core:Token xmi:id="195" sofa="11" begin="96" end="99" lemma="and" pos="CC"/>
    <core:Token xmi:id="206" sofa="11" begin="100" end="105" lemma="woman" pos="NNS"/>
    <core:Token xmi:id="217" sofa="11" begin="106" end="109" lemma="be" pos="VBP"/>
    <core:Token xmi:id="228" sofa="11" begin="110" end="118" lemma="become" pos="VVG"/>
    <core:Token xmi:id="239" sofa="11" begin="119" end="123" lemma="more" pos="RBR"/>
    <core:Token xmi:id="250" sofa="11" begin="124" end="141" lemma="internationalize" pos="VVN"/>
    <core:Token xmi:id="261" sofa="11" begin="142" end="143" lemma="-" pos=":"/>
    <core:Token xmi:id="272" sofa="11" begin="144" end="146" lemma="or" pos="CC"/>
    <core:Token xmi:id="283" sofa="11" begin="146" end="147" lemma="," pos=","/>
    <core:Token xmi:id="294" sofa="11" begin="148" end="155" lemma="perhaps" pos="RB"/>
    <core:Token xmi:id="305" sofa="11" begin="155" end="156" lemma="," pos=","/>
    <core:Token xmi:id="316" sofa="11" begin="157" end="168" lemma="westernize" pos="VVN"/>
    <core:Token xmi:id="327" sofa="11" begin="168" end="169" lemma="." pos="SENT"/>
    <cas:Sofa xmi:id="11" sofaNum="1" sofaID="_InitialView" mimeType="text" sofaString="in our globalized world with an increasingly
international media, beauty standards for both men and women are becoming more internationalized - or, perhaps, westernized."/>
    <cas:View sofa="11" members="1 6 18 26 30 41 52 63 74 85 96 107 118 129 140 151 162 173 184 195 206 217 228 239 250 261 272 283 294
305 316 327"/>
</xmi:XMI>
```

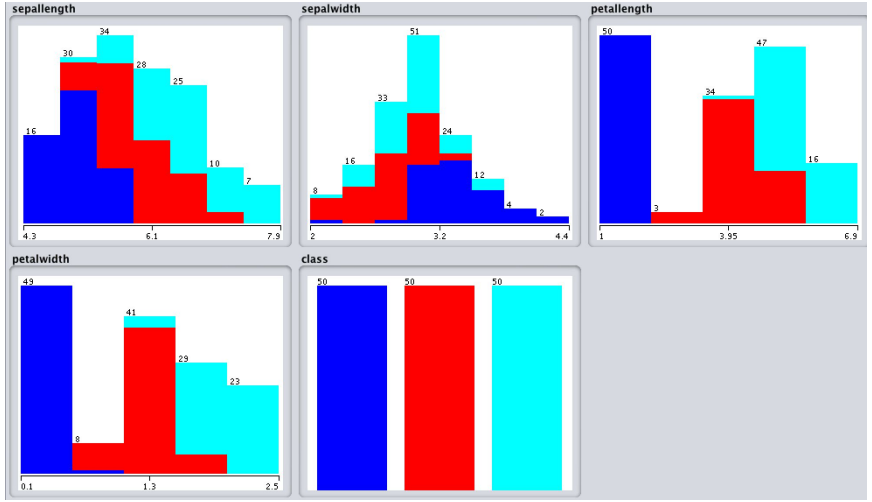

XML Tab Width: 8 Ln 2, Col 1369 INS

Weka

- An open source Java software that has a collection of machine learning algorithms for data mining tasks
- A powerful tool for understanding and visualizing machine learning algorithms on your local machine
- Contains tools for data preparation, classification, regression, clustering, and visualization
- Provides an easy way to apply many different algorithms to your data and see which one will give the best results

Visualise class distribution vs features

- Attribute information: sepal length, sepal width, petal length, petal width
- Class: Iris- Setosa, Versicolour, Virginica
- Petal length and petal width has high class correlation (0.95)
- Even class distribution (50 each)



Classify with Random Forest Classifier

- Random Forest is an ensemble learning algorithm that can be used for classification regression and other tasks. It works by constructing a multitude of decision trees at training time and outputting the predicted class.
- 10 fold cross validation: Form 10 test sets by splitting dataset into disjoint sets of similar size. Each time, train the classifier by 9 sets and test by the remaining set.

Summary of run information

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation
```

Result

- Accuracy: 95.3%

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         143               95.3333 %
Incorrectly Classified Instances         7                4.6667 %
Kappa statistic                          0.93
Mean absolute error                      0.0408
Root mean squared error                  0.1621
Relative absolute error                  9.19   %
Root relative squared error             34.3846 %
Total Number of Instances              150
```

Result (continue)

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 0.940    0.040    0.922      0.940    0.931      0.896    0.991     0.984     Iris-versicolor
                 0.920    0.030    0.939      0.920    0.929      0.895    0.991     0.982     Iris-virginica
Weighted Avg.    0.953    0.023    0.953      0.953    0.953      0.930    0.994     0.989

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	TN (d)

Precision:

$$\frac{a}{a+b}$$

Recall:

$$\frac{a}{a+c}$$

F-measure:

$$F_\alpha = \frac{1+\alpha}{\frac{1}{precision} + \frac{\alpha}{recall}}$$

$\alpha = 1$	harmonic mean
$\alpha \in (0;1)$	favor precision over recall
$\alpha > 1$	favor recall over precision