

Notebook de acompañamiento a

Autoría y estilo: Análisis estilométrico mediante clasificación de la Conquista de Jerusalén

Artículo realizado por:

- Juan Cerezo Soler
- José Calvo Tello

Presentado y aceptado en la revista *Anales Cervantinos*

Importación de herramientas

```
In [1]: import rpy2.objects as ro  
R = ro.r
```

```
In [2]: R.librarv("stylo")
```

```
/usr/local/lib/python3.4/dist-packages/rpy2/rinterface/__init__  
.py:186: RRuntimeWarning:  
### stylo version: 0.6.5 ###
```

If you plan to cite this software (please do!), use the following reference:

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R:
a package for computational text analysis. R Journal 8(1): 107-121.
<<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>>

To get full BibTeX entry, type: `citation("stylo")`

```
warnings.warn(x, RRuntimeWarning)
```

```
Out[2]: R object with classes: ('character',) mapped to:  
<StrVector - Python:0x7f780698f148 / R:0x82d4820>  
['stylo', 'tools', 'stats', ..., 'data...', 'meth...', 'base']
```

```
In [3]: R('''
        cosinedelta <- function(x){
          # z-scoring the input matrix of frequencies
          x = scale(x)
          # computing cosine dissimilarity
          y = as.dist( x %*% t(x) / (sqrt(rowSums(x^2) %*% t(rowSu
          # then, turning it into cosine similarity
          z = 1 - y
          # getting the results
          return(z)
        })
        distances = ["dist.eder", "dist.delta", "cosinedelta"]
```

Exploración del corpus mediante clustering

Realizamos dendrograma sin el texto discutido. En el artículo mostramos el dendrograma de Eder's Delta, 3000 MFW.

```
In [4]: R.setwd("/home/jose/Desktop/analysis/corpus0/")
```

```
Out[4]: R object with classes: ('character',) mapped to:
<StrVector - Python:0x7f780699a188 / R:0x872b3a8>
['/home/jose/Desktop/analysis']
```

```
In [5]: for distance in distances:
        results = R.stylo(
            **{
                "gui" : False,
                "analyzed.features" : "w",
                "ngram.size" : 1,
                "preserve.case" : False,
                "mfw.min" : 1000,
                "mfw.max" : 5000,
                "mfw.incr" : 1000,
                "mfw.list.cutoff" : 5000,
                "analysis.type" : "CA",
                "distance.measure" : distance,
                "sampling" : "no.sampling",
                "display.on.screen" : False,
                "write.png.file" : True,
                "save.distance.tables" : True,
                "save.analyzed.features" : True,
                "save.analyzed.freqs" : True,
                "use.existing.freq.tables" : True,
                "use.custom.list.of.files" : True,
                "use.existing.wordlist" : True,
            }
        )
```

```
/usr/local/lib/python3.4/dist-packages/rpy2/rinterface/__init__
.py:186: RRuntimeWarning: Read 5005 items
```

```
warnings.warn(x, RRuntimeWarning)
```

```
using current directory...
```

```
reading a custom set of features from a file...
```

```
reading a file containing frequencies...
```

Clustering

Clustering de texto discutido con textos similares de Cervantes

Realizamos dendrograma con el texto discutido. En el artículo mostramos el dendrograma de Eder's Delta, 3000 MFW.

```
In [6]: R.setwd("/home/jose/Desktop/analysis/corpus1/")
```

```
Out[6]: R object with classes: ('character',) mapped to:  
<StrVector - Python:0x7f780699cccc8 / R:0x83a3628>  
['/home/jose/Desktop/analysis/corpus0']
```

```
In [7]: for distance in distances:
        results = R.stylo(
            **{
                "gui" : False,
                "analyzed.features" : "w",
                "ngram.size" : 1,
                "preserve.case" : False,
                "mfw.min" : 1000,
                "mfw.max" : 5000,
                "mfw.incr" : 1000,
                "mfw.list.cutoff" : 5000,
                "analysis.type" : "CA",
                "distance.measure" : distance,
                "sampling" : "no.sampling",
                "display.on.screen" : False,
                "write.png.file" : True,
                "save.distance.tables" : True,
                "save.analyzed.features" : True,
                "save.analyzed.freqs" : True,
                "use.existing.freq.tables" : True,
                "use.custom.list.of.files" : True,
                "use.existing.wordlist" : True,
            }
        )
```

using current directory...

reading a custom set of features from a file...

reading a file containing frequencies...

Building a

Realización de Consensus Tree con mismo corpus

```
In [8]: for distance in distances:
        results = R.stylo(
            **{
                "gui" : False,
                "analyzed.features" : "w",
                "ngram.size" : 1,
                "preserve.case" : False,
                "mfw.min" : 500,
                "mfw.max" : 5000,
                "mfw.incr" : 500,
                "mfw.list.cutoff" : 5000,
                "analysis.type" : "BCT",
                "distance.measure" : distance,
                "sampling" : "no.sampling",
                "display.on.screen" : False,
                "write.png.file" : True,
                "save.distance.tables" : True,
                "save.analyzed.features" : True,
                "save.analyzed.freqs" : True,
                "use.existing.freq.tables" : True,
                "use.custom.list.of.files" : True,
                "use.existing.wordlist" : True,
            }
        )
```

using current directory...

reading a custom set of features from a file...

reading a file containing frequencies...

Building a

Clustering de texto discutido con textos menos similares de Cervantes

```
In [9]: R.setwd("/home/jose/Desktop/analysis/corpus2/")
```

```
Out[9]: R object with classes: ('character',) mapped to:
<StrVector - Python:0x7f78069a89c8 / R:0xb6bfd58>
['/home/jose/Desktop/analysis/corpus1']
```

```
In [10]: for distance in distances:
          results = R.stylo(
              **{
                  "gui" : False,
                  "analyzed.features" : "w",
                  "ngram.size" : 1,
                  "preserve.case" : False,
                  "mfw.min" : 1000,
                  "mfw.max" : 5000,
                  "mfw.incr" : 1000,
                  "mfw.list.cutoff" : 5000,
                  "analysis.type" : "CA",
                  "distance.measure" : distance,
                  "sampling" : "no.sampling",
                  "display.on.screen" : False,
                  "write.png.file" : True,
                  "save.distance.tables" : True,
                  "save.analyzed.features" : True,
                  "save.analyzed.freqs" : True,
                  "use.existing.freq.tables" : True,
                  "use.custom.list.of.files" : True,
                  "use.existing.wordlist" : True,
              }
          )
```

using current directory...

reading a custom set of features from a file...

reading a file containing frequencies...

colling 0

```
In [11]: R.setwd("/home/jose/Desktop/analysis/corpus3/")
```

```
Out[11]: R object with classes: ('character',) mapped to:
<StrVector - Python:0x7f78069a7d08 / R:0x9ff68a8>
['/home/jose/Desktop/analysis/corpus2']
```

```
In [12]: for distance in distances:
          results = R.stylo(
              **{
                  "gui" : False,
                  "analyzed.features" : "w",
                  "ngram.size" : 1,
                  "preserve.case" : False,
                  "mfw.min" : 1000,
                  "mfw.max" : 5000,
                  "mfw.incr" : 1000,
                  "mfw.list.cutoff" : 5000,
                  "analysis.type" : "CA",
                  "distance.measure" : distance,
                  "sampling" : "no.sampling",
                  "display.on.screen" : False,
                  "write.png.file" : True,
                  "save.distance.tables" : True,
                  "save.analyzed.features" : True,
                  "save.analyzed.freqs" : True,
                  "use.existing.freq.tables" : True,
                  "use.custom.list.of.files" : True,
                  "use.existing.wordlist" : True,
              }
          )
```

using current directory...

reading a custom set of features from a file...

reading a file containing frequencies...

colling 0

Clasificación


```
In [13]: R.setwd("/home/jose/Desktop/analysis/corpus4/")
results = R.classify(
    **{
        "gui" : False,
        "analyzed.features" : "w",
        "ngram.size" : 1,
        "preserve.case" : False,
        "mfw.min" : 500,
        "mfw.max" : 5000,
        "mfw.incr" : 500,
        "mfw.list.cutoff" : 5000,
        "distance.measure" : "dist.eder",
        "sampling" : "no.sampling",
        "display.on.screen" : False,
        "save.distance.tables" : True,
        "save.analyzed.features" : True,
        "save.analyzed.freqs" : True,
        "use.existing.freq.tables": True,
        "use.custom.list.of.files": True,
        "use.existing.wordlist": True,
    }
)

scores corpus4 = results[6]
using current directory...
```

reading a custom set of features from a file...

reading a file containing frequencies...

Training set successfully loaded.

```
In [14]: R.setwd("/home/jose/Desktop/analysis/corpus5/")
results = R.classify(
    **{
        "gui" : False,
        "analyzed.features" : "w",
        "ngram.size" : 1,
        "preserve.case" : False,
        "mfw.min" : 500,
        "mfw.max" : 5000,
        "mfw.incr" : 500,
        "mfw.list.cutoff" : 5000,
        "distance.measure" : "dist.eder",
        "sampling" : "no.sampling",
        "display.on.screen" : False,
        "save.distance.tables" : True,
        "save.analyzed.features" : True,
        "save.analyzed.freqs" : True,
        "use.existing.freq.tables": True,
        "use.custom.list.of.files": True,
        "use.existing.wordlist": True,
    }
)
scores corpus5 = results[6]
using current directory...
```

reading a custom set of features from a file...

reading a file containing frequencies...

Training set successfully loaded.

```
In [15]: R.setwd("/home/jose/Desktop/analysis/corpus6/")
results = R.classify(
    **{
        "gui" : False,
        "analyzed.features" : "w",
        "ngram.size" : 1,
        "preserve.case" : False,
        "mfw.min" : 500,
        "mfw.max" : 5000,
        "mfw.incr" : 500,
        "mfw.list.cutoff" : 5000,
        "distance.measure" : "dist.eder",
        "sampling" : "no.sampling",
        "display.on.screen" : False,
        "save.distance.tables" : True,
        "save.analyzed.features" : True,
        "save.analyzed.freqs" : True,
        "use.existing.freq.tables": True,
        "use.custom.list.of.files": True,
        "use.existing.wordlist": True,
    }
)
scores corpus6 = results[6]
using current directory...
```

reading a custom set of features from a file...

reading a file containing frequencies...

Training set successfully loaded.

```
In [16]: list(scores corpus4)
```

```
Out[16]: [100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0,
100.0]
```

```
In [17]: import pandas as pd
columns = ["500", "1000", "1500", "2000", "2500", "3000", "3500", "
dataframe = pd.DataFrame([list(scores_corpus4), list(scores_corp
```

```
In [18]: dataframe
```

Out[18]:

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
corpus4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
corpus 5	50.0	75.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
corpus 6	80.0	80.0	80.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

```
In [19]: import matplotlib.pyplot as plt
import numpy as np

results_list = [item[0] + "-" + item[1] for item in list(zip(col

%matplotlib inline
dataframe.boxplot(medianprops = dict(linewidth=5), figsize=(5,5)
plt.ylim((0,110))
plt.xticks(np.arange(1, len(columns)+1, 1), results_list, rotati
plt.ylabel("Resultado de evaluación")
plt.xlabel("MFW y autor clasificado")
plt.title("Resultados de clasificación\nde La conquista de Jerus
plt.tight_layout()
plt.savefig("/home/jose/Desktop/analysis/results.png", dpi=300)
plt.show()
```



