Práctica 9

Procesamiento de Lenguaje Natural

Javier Pellejero Ortega & Zhaoyan Ni

Inteligencia Artificial

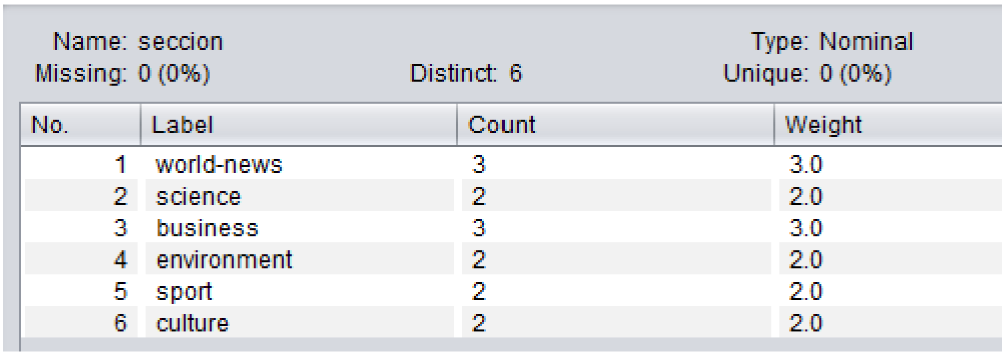
Grupo 11

Doble grado Matemáticas e ingeniería informática

PARTE 2:

Para realizar esta parte, hemos buscado párrafos de noticias que están publicadas en periódicos digitales de 6 secciones diferentes: *World-new, Science, Business, Environment, Sport and Culture*.

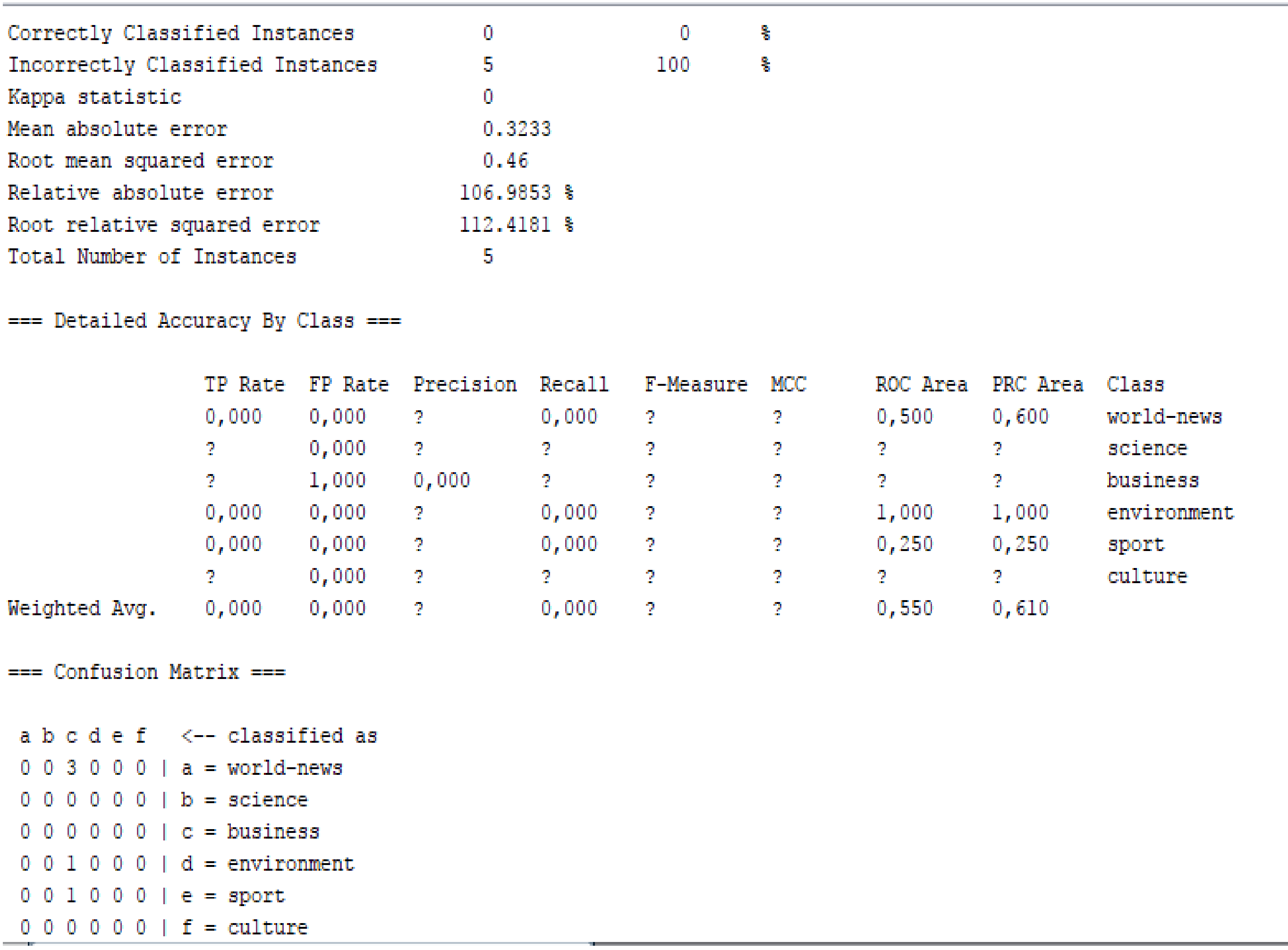
En el archivo .arff, están 14 instancias, de las cuales, 3 son de la clase *world-news*, 2 de la clase *science*, 3 de la clase *business*, 2 de la clase *environment*, 2 de la clase *sport* y 2 de la clase *culture*.



Name: seccion
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 0 (0%)

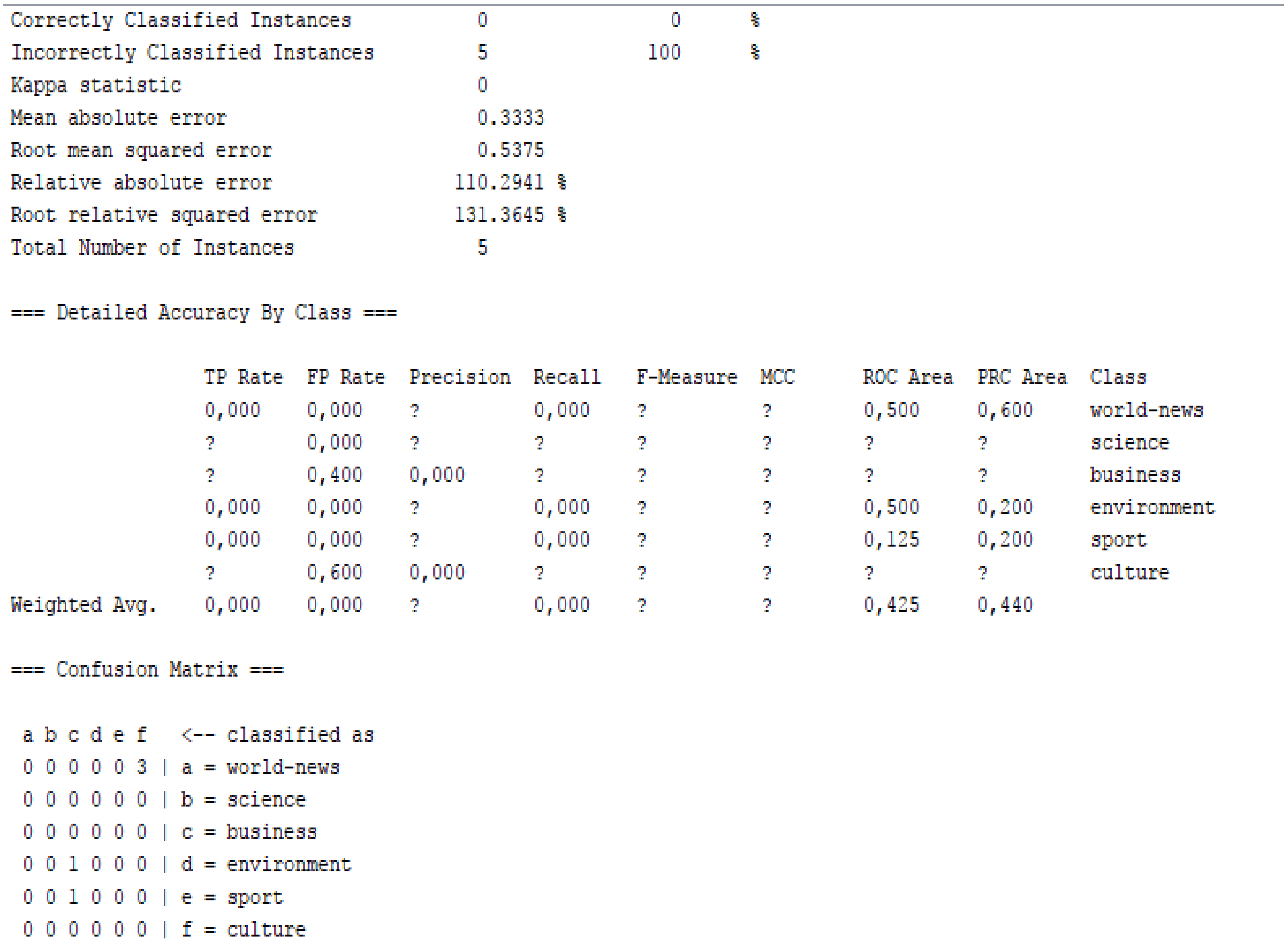
No.	Label	Count	Weight
1	world-news	3	3.0
2	science	2	2.0
3	business	3	3.0
4	environment	2	2.0
5	sport	2	2.0
6	culture	2	2.0

Hemos intentado clasificar estas 14 instancias con 5 clasificadores, en todos ellos, hemos utilizado dos tercios de los datos para entrenar un clasificador y un tercio de los datos sirve como el conjunto de validación. Los resultados obtenidos son siguientes:



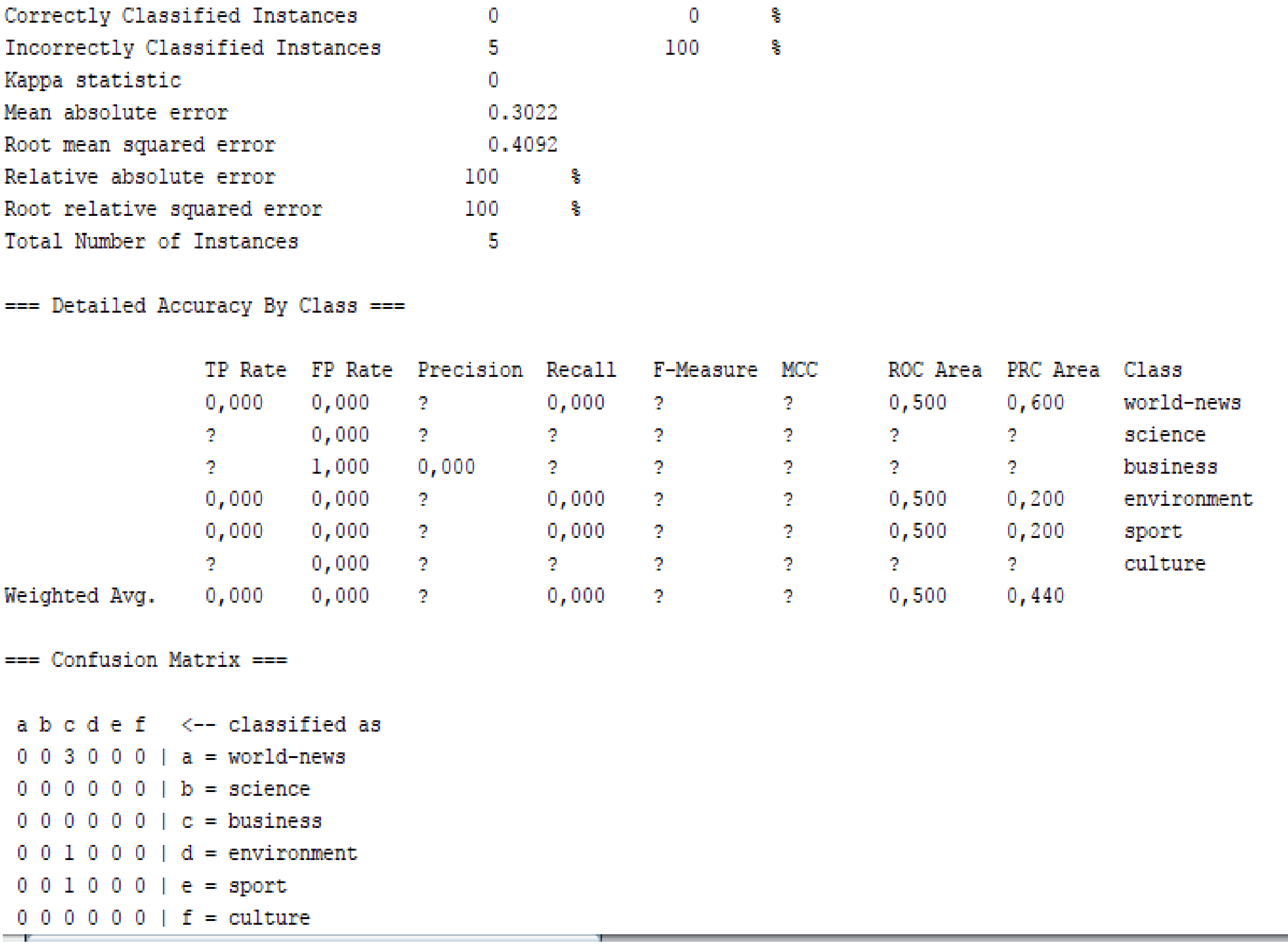
```
Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         5              100      %
Kappa statistic                          0
Mean absolute error                      0.3233
Root mean squared error                  0.46
Relative absolute error                106.9853 %
Root relative squared error            112.4181 %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,600     world-news
                 ?        0,000    ?          ?        ?          ?        ?         ?         science
                 ?        1,000    0,000      ?        ?          ?        ?         ?         business
                 0,000    0,000    ?          0,000    ?          ?        1,000     1,000     environment
                 0,000    0,000    ?          0,000    ?          ?        0,250     0,250     sport
                 ?        0,000    ?          ?        ?          ?        ?         ?         culture
Weighted Avg.    0,000    0,000    ?          0,000    ?          ?        0,550     0,610

=== Confusion Matrix ===

 a b c d e f   <-- classified as
 0 0 3 0 0 0 | a = world-news
 0 0 0 0 0 0 | b = science
 0 0 0 0 0 0 | c = business
 0 0 1 0 0 0 | d = environment
 0 0 1 0 0 0 | e = sport
 0 0 0 0 0 0 | f = culture
```

Resultado obtenido con clasificador RandomForest(trees)



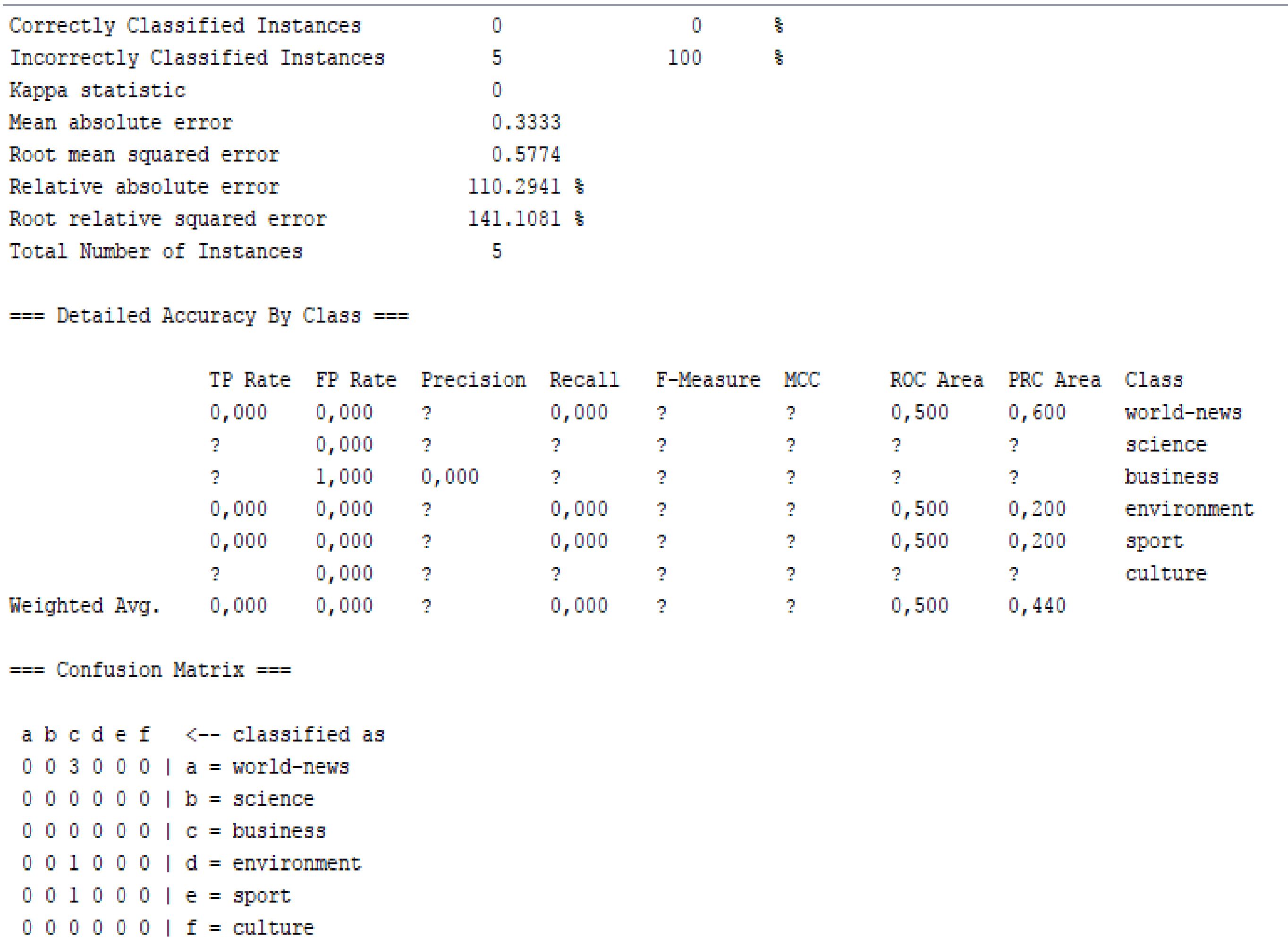
```
Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         5              100      %
Kappa statistic                          0
Mean absolute error                      0.3333
Root mean squared error                  0.5375
Relative absolute error                110.2941 %
Root relative squared error            131.3645 %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,600     world-news
                 ?        0,000    ?          ?        ?          ?        ?         ?         science
                 ?        0,400    0,000      ?        ?          ?        ?         ?         business
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     environment
                 0,000    0,000    ?          0,000    ?          ?        0,125     0,200     sport
                 ?        0,600    0,000      ?        ?          ?        ?         ?         culture
Weighted Avg.    0,000    0,000    ?          0,000    ?          ?        0,425     0,440

=== Confusion Matrix ===

 a b c d e f   <-- classified as
 0 0 0 0 0 3 | a = world-news
 0 0 0 0 0 0 | b = science
 0 0 0 0 0 0 | c = business
 0 0 1 0 0 0 | d = environment
 0 0 1 0 0 0 | e = sport
 0 0 0 0 0 0 | f = culture
```

Resultados obtenidos aplicando J48 (tree)



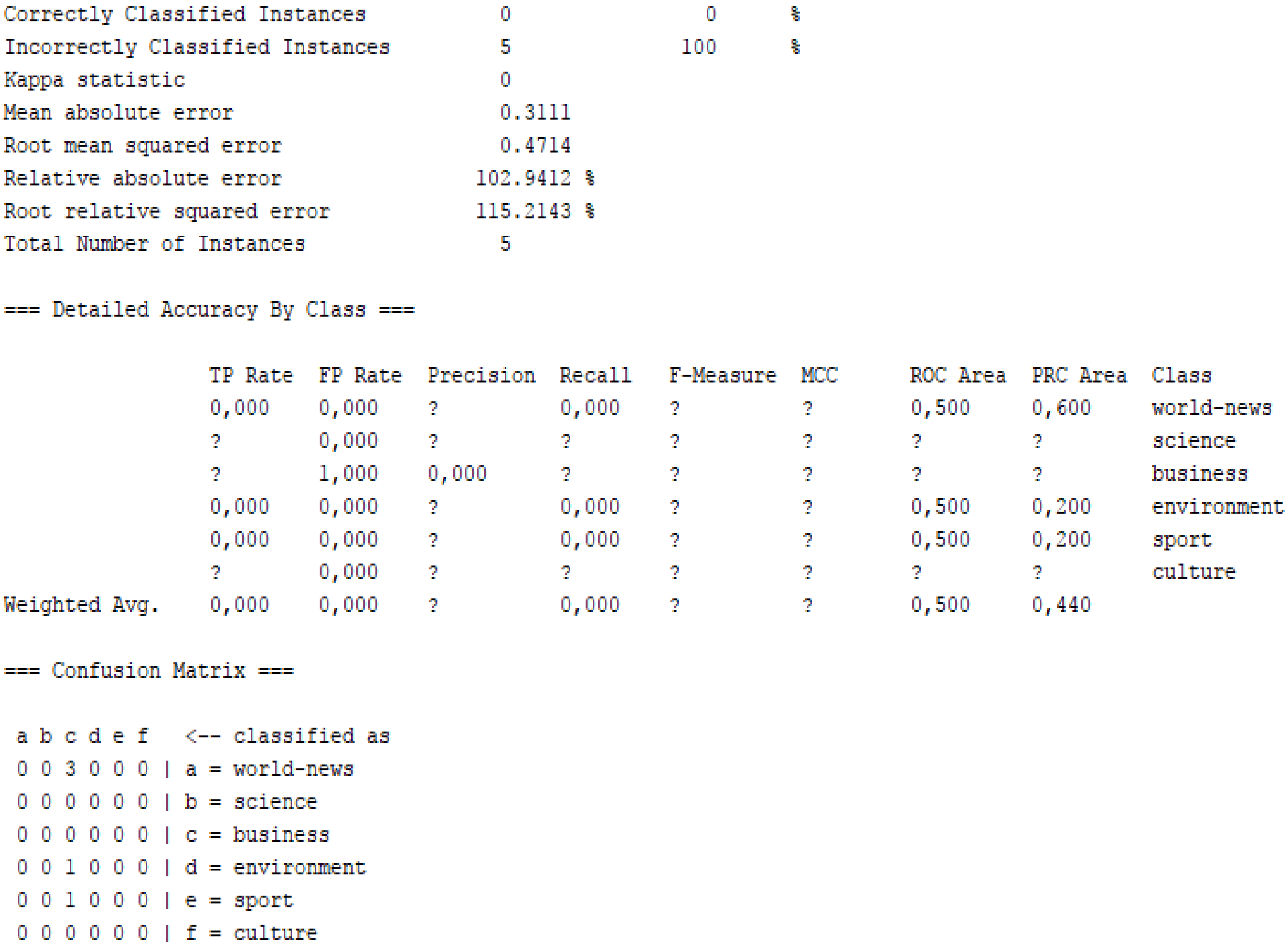
```
Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         5              100      %
Kappa statistic                          0
Mean absolute error                      0.3022
Root mean squared error                  0.4092
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,600     world-news
                 ?        0,000    ?          ?        ?          ?        ?         ?         science
                 ?        1,000    0,000      ?        ?          ?        ?         ?         business
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     environment
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     sport
                 ?        0,000    ?          ?        ?          ?        ?         ?         culture
Weighted Avg.    0,000    0,000    ?          0,000    ?          ?        0,500     0,440

=== Confusion Matrix ===

 a b c d e f   <-- classified as
 0 0 3 0 0 0 | a = world-news
 0 0 0 0 0 0 | b = science
 0 0 0 0 0 0 | c = business
 0 0 1 0 0 0 | d = environment
 0 0 1 0 0 0 | e = sport
 0 0 0 0 0 0 | f = culture
```

Resultados obtenidos aplicando ZeroR (rules)



```
Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         5              100      %
Kappa statistic                          0
Mean absolute error                      0.3333
Root mean squared error                  0.5774
Relative absolute error                110.2941 %
Root relative squared error            141.1081 %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,600     world-news
                 ?        0,000    ?          ?        ?          ?        ?         ?         science
                 ?        1,000    0,000      ?        ?          ?        ?         ?         business
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     environment
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     sport
                 ?        0,000    ?          ?        ?          ?        ?         ?         culture
Weighted Avg.    0,000    0,000    ?          0,000    ?          ?        0,500     0,440

=== Confusion Matrix ===

 a b c d e f   <-- classified as
 0 0 3 0 0 0 | a = world-news
 0 0 0 0 0 0 | b = science
 0 0 0 0 0 0 | c = business
 0 0 1 0 0 0 | d = environment
 0 0 1 0 0 0 | e = sport
 0 0 0 0 0 0 | f = culture
```

Resultados obtenidos aplicando OneR (rules)



```
Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         5              100      %
Kappa statistic                          0
Mean absolute error                      0.3111
Root mean squared error                  0.4714
Relative absolute error                102.9412 %
Root relative squared error            115.2143 %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,600     world-news
                 ?        0,000    ?          ?        ?          ?        ?         ?         science
                 ?        1,000    0,000      ?        ?          ?        ?         ?         business
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     environment
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,200     sport
                 ?        0,000    ?          ?        ?          ?        ?         ?         culture
Weighted Avg.    0,000    0,000    ?          0,000    ?          ?        0,500     0,440

=== Confusion Matrix ===

 a b c d e f   <-- classified as
 0 0 3 0 0 0 | a = world-news
 0 0 0 0 0 0 | b = science
 0 0 0 0 0 0 | c = business
 0 0 1 0 0 0 | d = environment
 0 0 1 0 0 0 | e = sport
 0 0 0 0 0 0 | f = culture
```

Resultados obtenidos aplicando lazy (IBK)

Como podemos observar, en todos los resultados, no hemos clasificador bien ninguna instancia. Esto puede ser debido tenemos demasiados atributos comparando con las instancias, puesto que solo tenemos 14 instancias, y el número de atributos depende de los textos que hemos elegido porque hemos utilizado el filtro *StringtoWordVector* para convertir los textos en vectores de palabras. Además, la aparición de los nombres propios también puede dificultar la clasificación.

BIBLIOGRAFÍA:

https://www.theguardian.com/science/2018/jun/11/trials-begin-of-a-saliva-test-for-prostate-cancer

https://www.theguardian.com/business/2018/jun/11/ryanairs-uk-cabin-crew-to-be-represented-by-union-for-first-time-unite

https://www.theguardian.com/technology/2018/jun/11/bitcoin-price-cryptocurrency-hacked-south-korea-coincheck

https://www.theguardian.com/science/2018/jun/06/when-a-dinosaur-fossil-is-gone-its-gone-forever

https://www.theguardian.com/sport/2018/jun/11/wales-wayne-pivac-shortlist-to-succeed-warren-gatland-rugby-union

https://www.theguardian.com/sport/2018/jun/11/chris-froome-lizzie-deignan-doping-cases-reputations

https://www.theguardian.com/world/2018/jun/12/new-zealand-coalition-under-strain-as-jacinda-ardern-prepares-for-maternity-leave

https://www.theguardian.com/world/2018/jun/12/us-de-facto-embassy-in-taiwan-reopens-as-symbol-of-strength-of-ties

https://www.theguardian.com/politics/2018/jun/08/acerbic-and-firm-mary-wilson-remembered-fondly-after-death-at-102

https://www.theguardian.com/commentisfree/2018/jun/11/brexit-uk-fishermen-fishing-industry-quotas-uk-government

https://www.theguardian.com/games/2018/jun/11/e3-2018-bethesda-and-microsoft-unveil-fallout-elder-scrolls-vi-halo-and-gears-of-war

https://www.theguardian.com/business/2018/jun/11/poundworld-administration-jobs-high-street

https://www.theguardian.com/science/2018/jun/09/asteroid-mining-space-prospectors-precious-resources-fuelling-future-among-stars

https://www.theguardian.com/science/2018/jun/11/trials-begin-of-a-saliva-test-for-prostate-cancer

https://www.theguardian.com/world/2018/jun/12/ireland-blasphemy-referendum