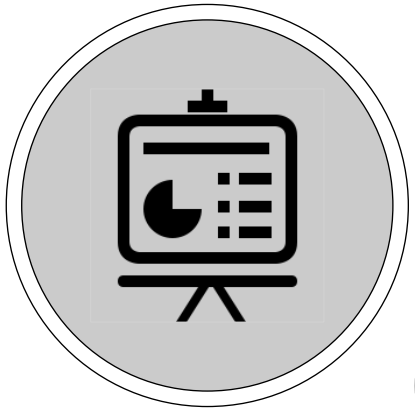
UNIVERSITÉ DU LUXEMBOURG

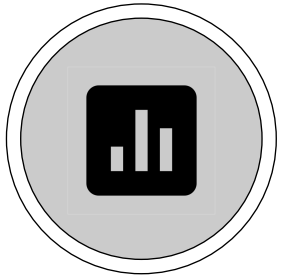
Advanced Content Analysis for Web Pages in Apache Spark

Dioni REBONATO ENDRINGER

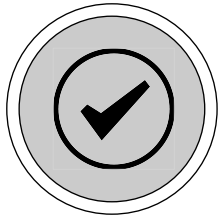
Introduction



Methodology



Results



Conclusion

university of luxembourg receives grant from eu

Recherche Google J'ai de la chance

Google disponible en : Deutsch


Google
university of luxembourg receives grant from eu master
Tous
Actualités
Images
Vidéos
Maps
Plus
Paramètres
Outils
Environ 1.820.000 résultats (0,64 secondes)
Conseil : Recherchez des résultats uniquement en français. Vous pouvez indiquer votre langue de recherche sur la page Préférences.
University receives 2.5 million euros for research in data science
https://wwwen.uni.lu/university/.../university_receives_2_5_millio...
Traduire cette page
13 sept. 2018 - The University of Luxembourg has recently been awarded a prestigious ERA (European Research Area) Chair grant under the EU's Horizon 2020 ... In addition, new teaching programmes at Master and PhD levels will be ...

Introduction

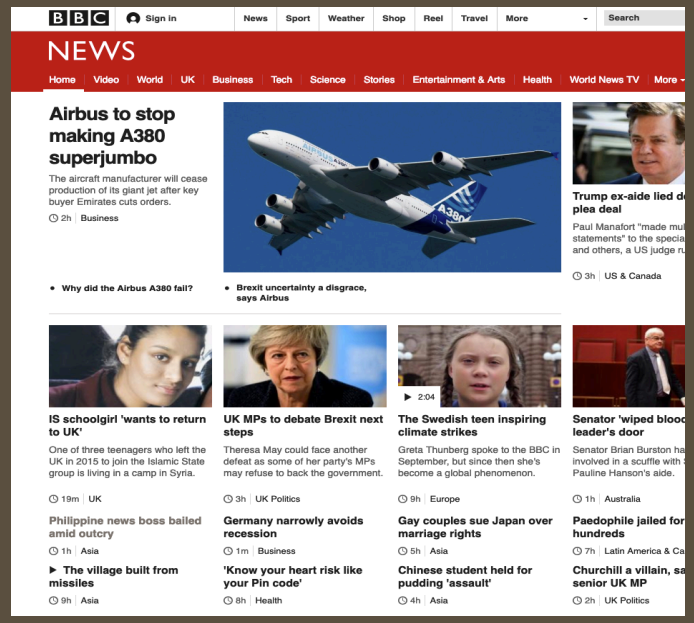
Research Questions

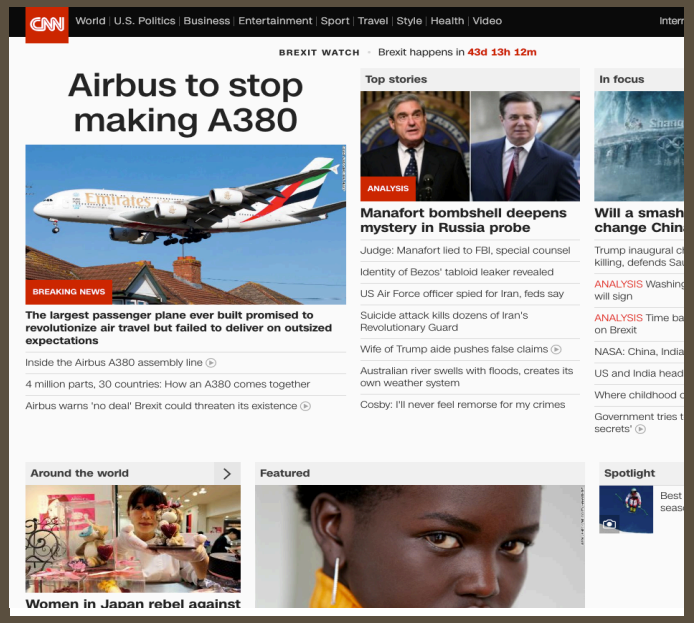
Is it possible to build a Machine Learning model to extract the information from some web pages and analyse them in order to classify the best content part in a web page?

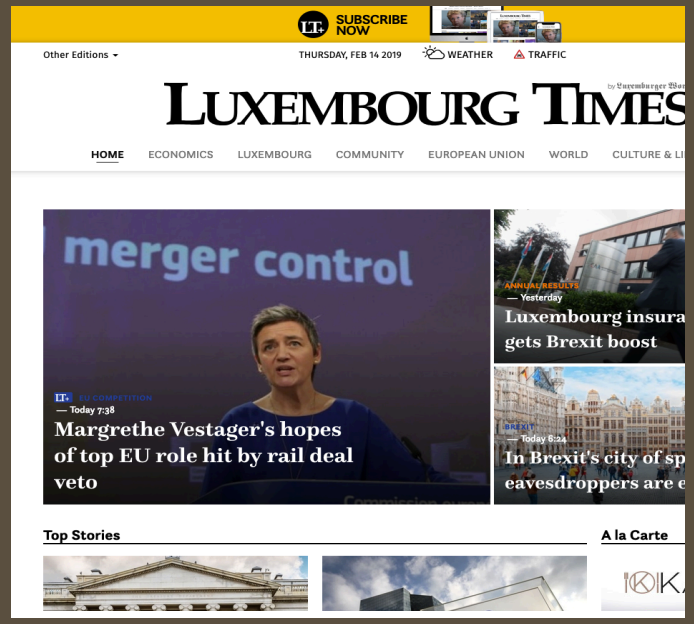
If so, how could we evaluate the performance or the precision of the model?

Is it possible to apply the model to all types of web pages?

What could be the best collection of features for analysing and classifying the web content?



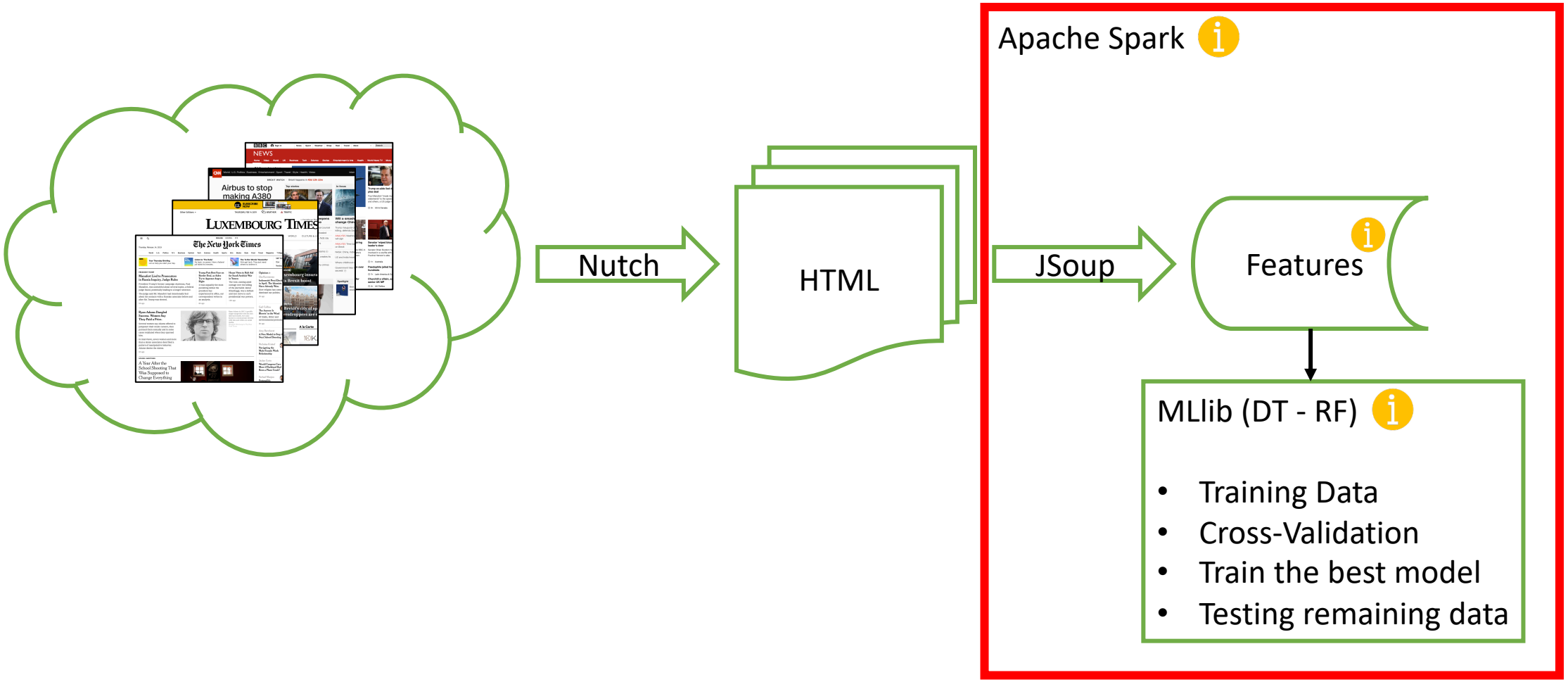




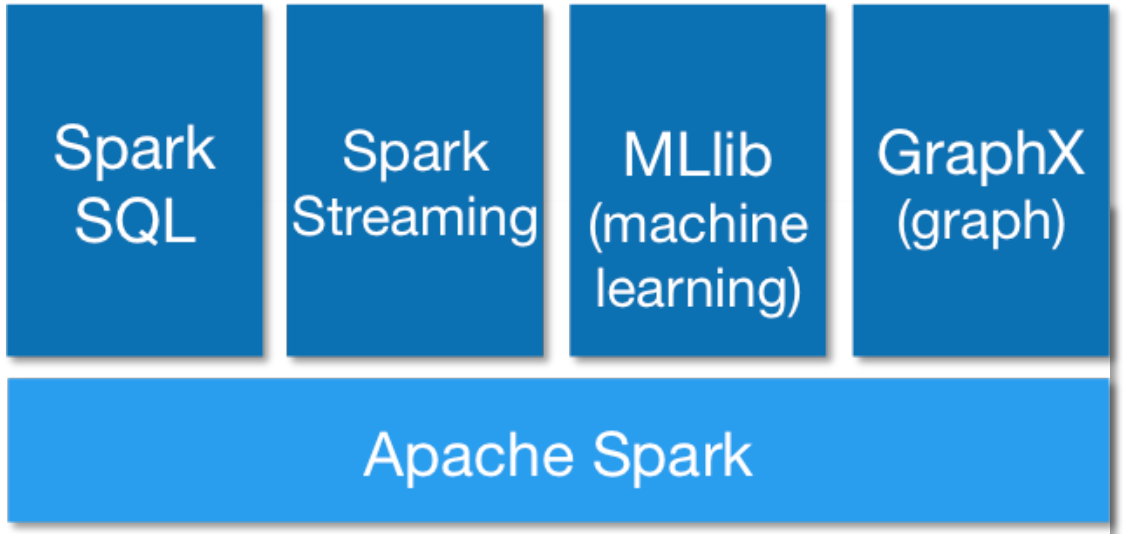
Methodology

BBC – https://www.bbc.com/news
CNN – https://edition.cnn.com/
LUX – https://luxtimes.lu/
NYT – https://www.nytimes.com/

Methodology



Apache Spark



Spark Master at spark://iris-149:7077

URL: spark://iris-149:7077
Alive Workers: 8
Cores in use: 223 Total, 223 Used
Memory in use: 996.5 GB Total, 960.0 GB Used
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (8)

Worker Id	Address	State	Cores	Memory
worker-20190421011021-172.17.6.153-33413	172.17.6.153:33413	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.149-42645	172.17.6.149:42645	ALIVE	27 (27 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.151-43136	172.17.6.151:43136	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.154-38000	172.17.6.154:38000	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.155-36900	172.17.6.155:36900	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.156-33098	172.17.6.156:33098	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.157-45500	172.17.6.157:45500	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)
worker-20190421011022-172.17.6.158-37667	172.17.6.158:37667	ALIVE	28 (28 Used)	124.6 GB (120.0 GB Used)

Extracted Features

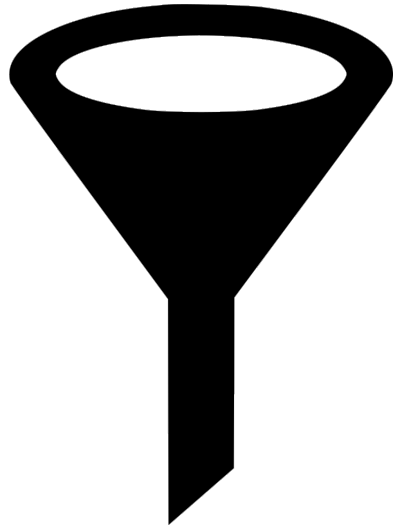
- docTitle
- docLinks
- docWords
- docDensity
- docParag
- docImages
- docChildSize
- docDiv

Document Features

- divLinks
- divWords
- divStopWords
- divDensity
- divParag
- divImages
- divChildSize
- divClass

Element Features

- goodBad



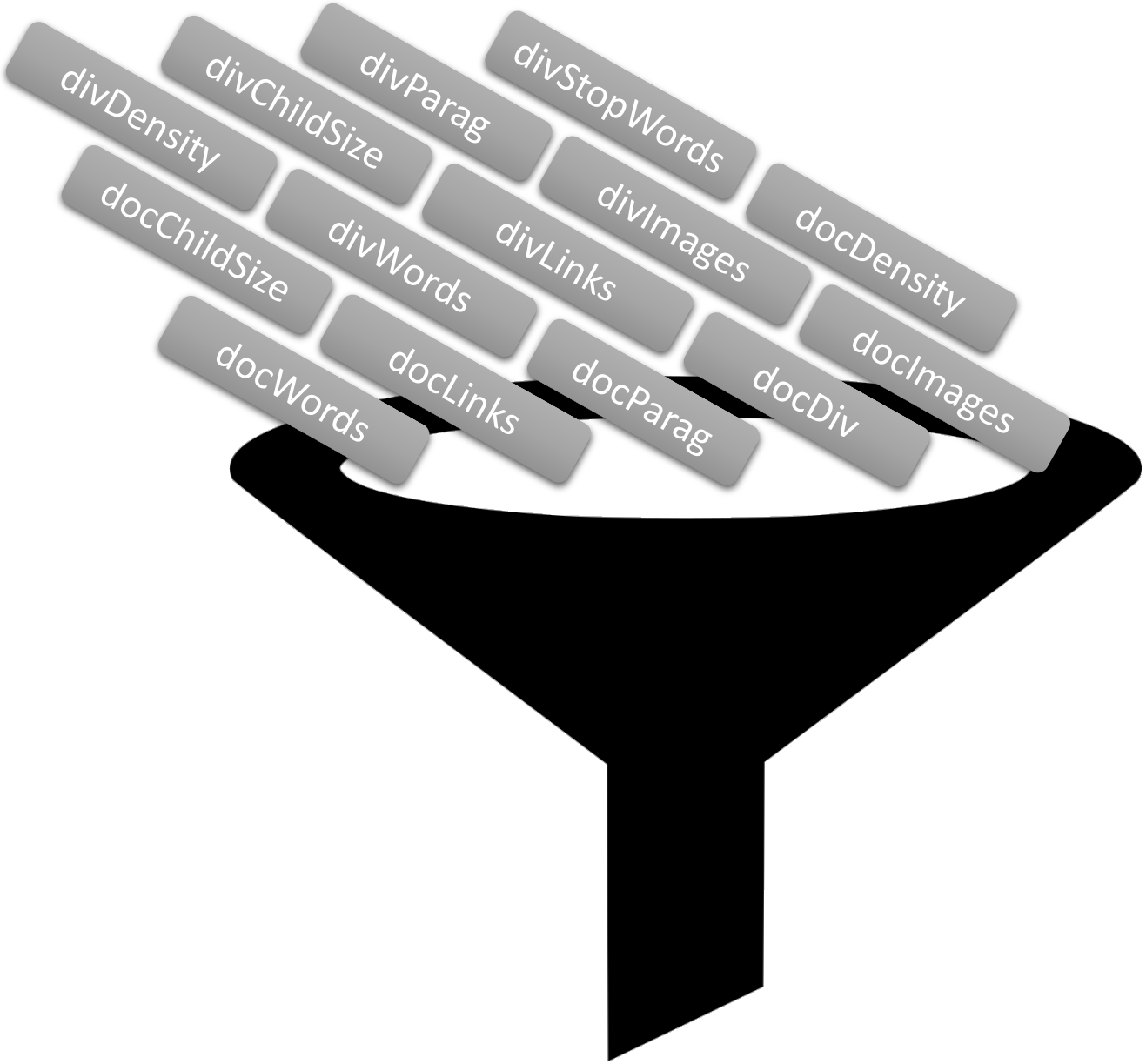
Extracted Features

- docTitle
- divClass

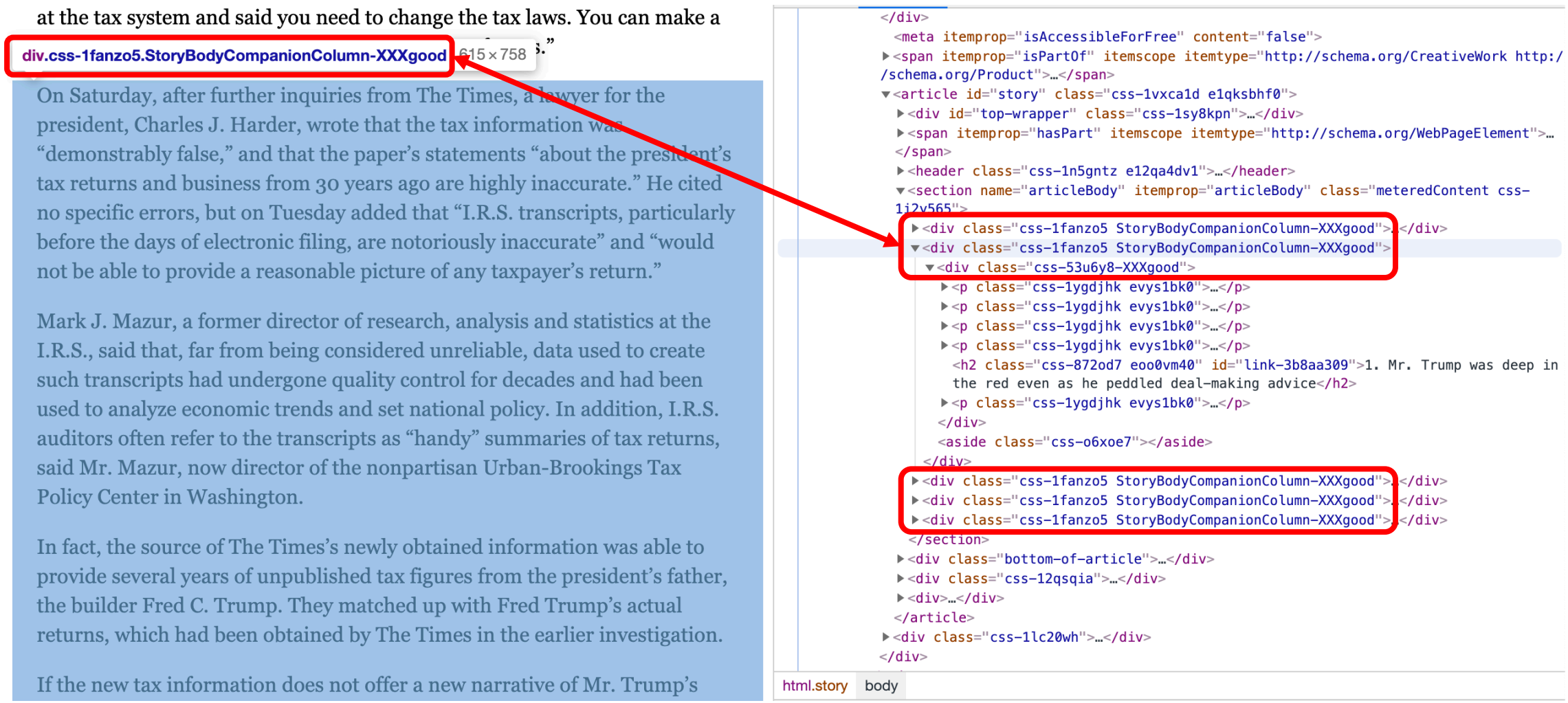
Future Identification

- goodBad

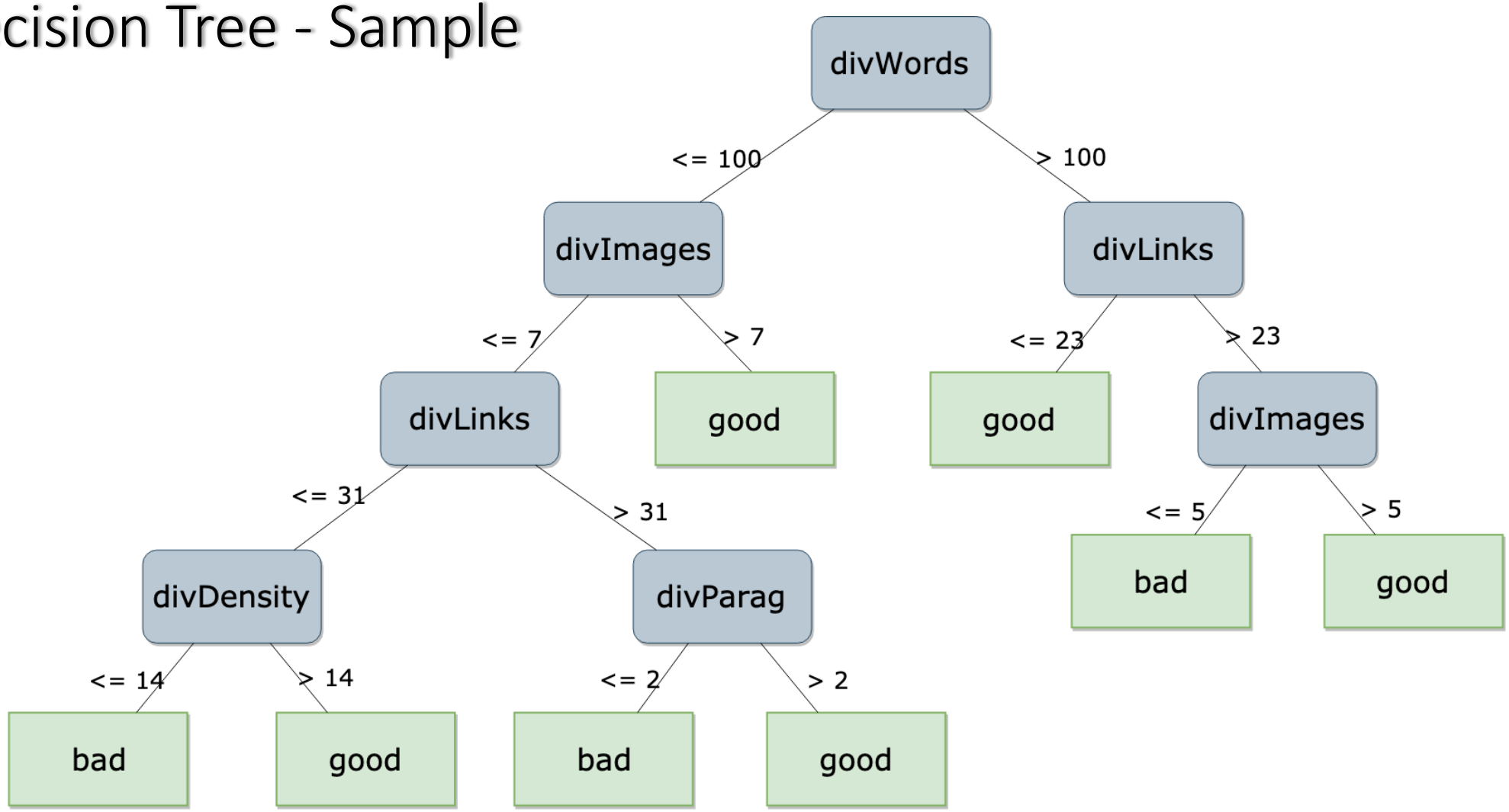
Labelling Point



Labelling Dataset



Decision Tree - Sample



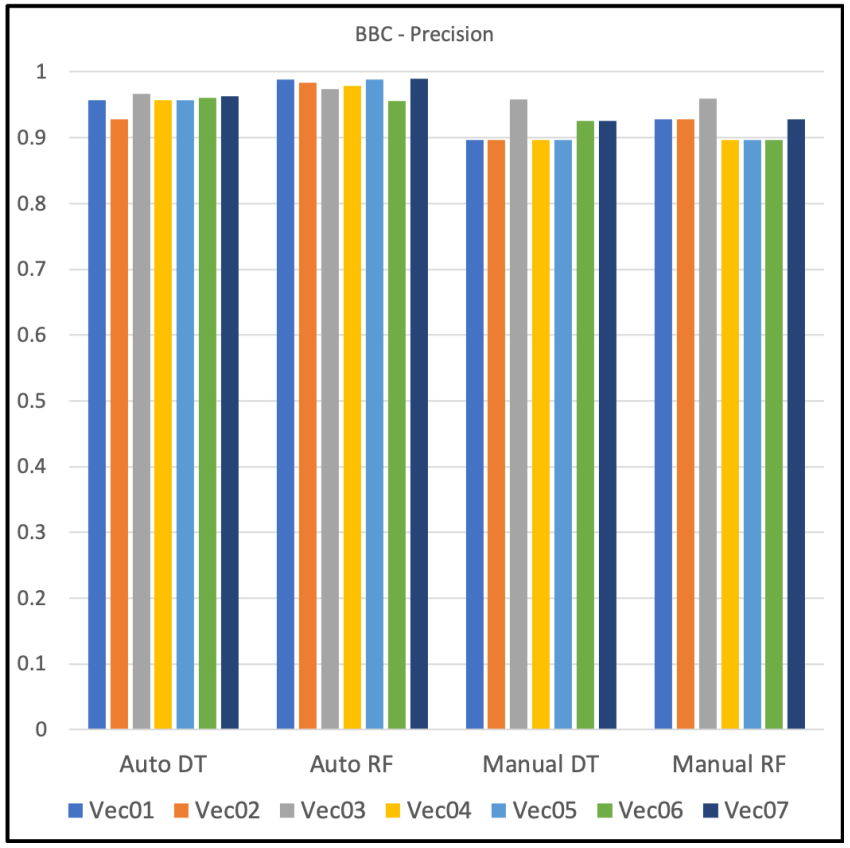
Datasets

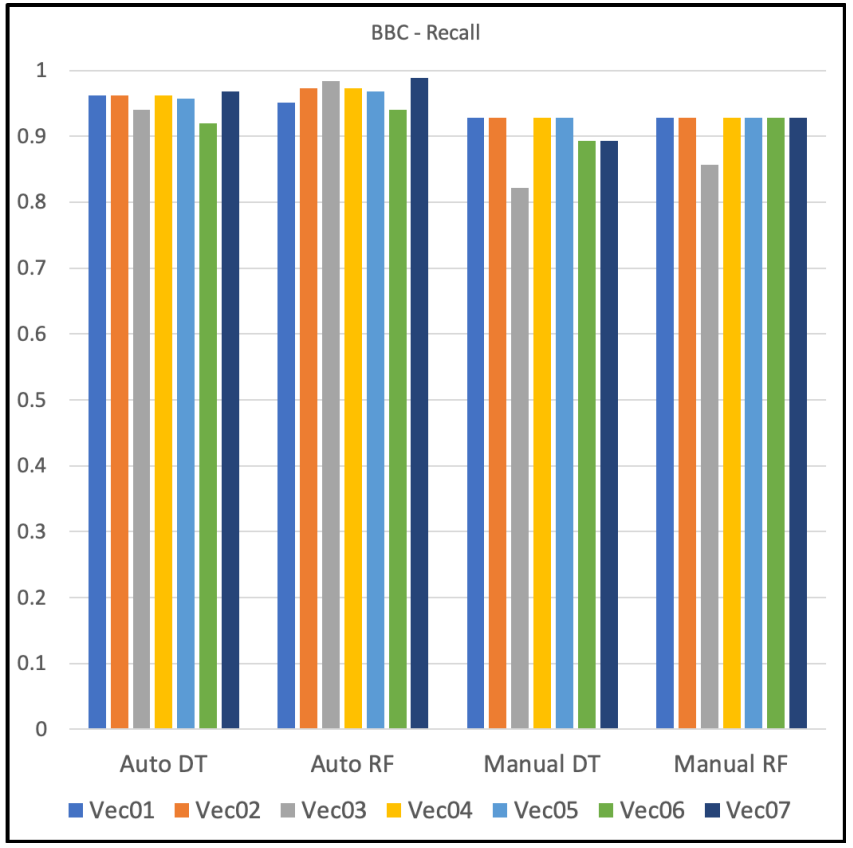
Site	Automatic	Manual
BBC	390	60
CNN	420	42
LUX	200	60
NYT	340	62

Vector	Features
Vec01	All 14 features (document and element)
Vec02	Removed Words features
Vec03	Removed Links features
Vec04	Removed Density features
Vec05	Removed Images features
Vec06	Removed Structural features
Vec07	Only “element level” features

Results

BBC – https://www.bbc.com/news



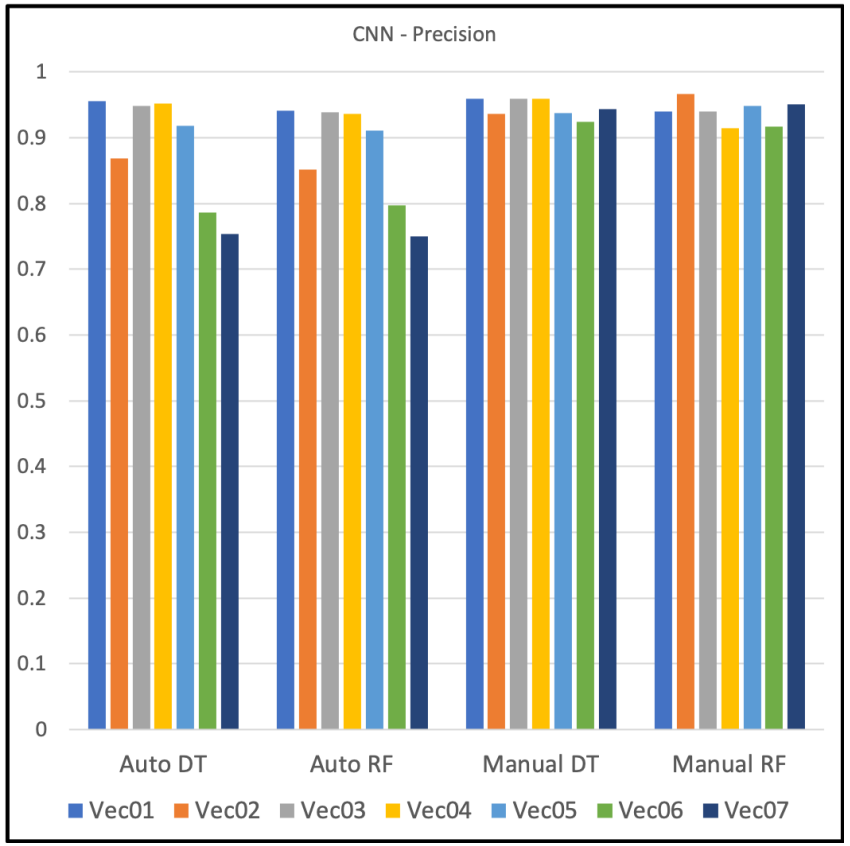


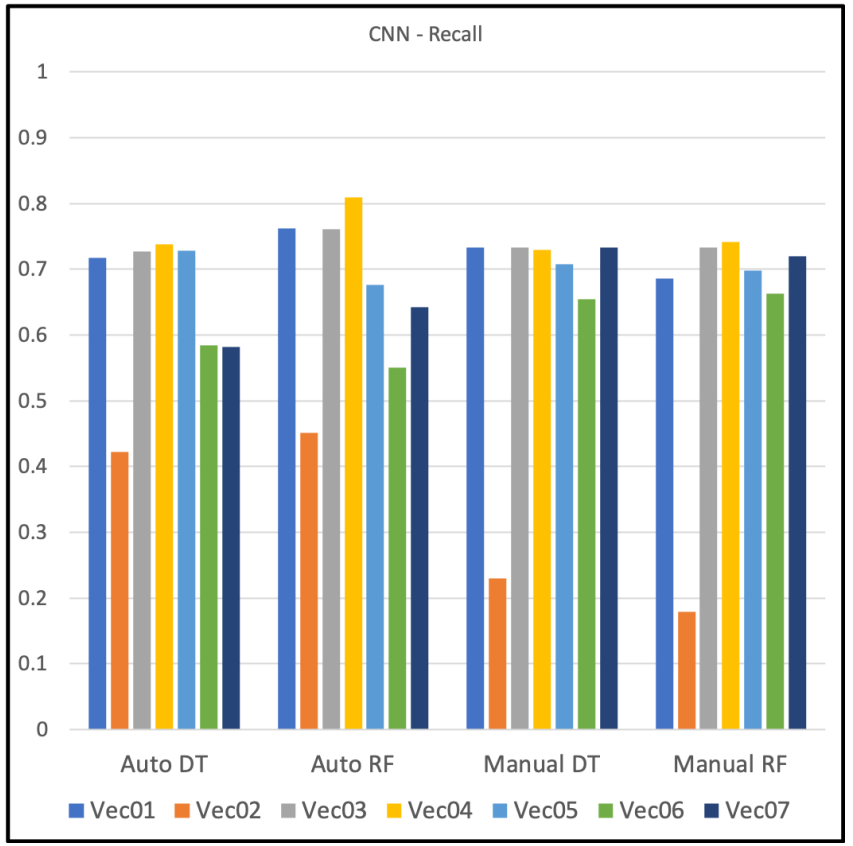
Auto	Manual
390	60

Vector	Features
Vec01	All 14 features
Vec02	No Words
Vec03	No Links
Vec04	No Density
Vec05	No Images
Vec06	No Structural
Vec07	Only Element

Results

CNN – https://edition.cnn.com/



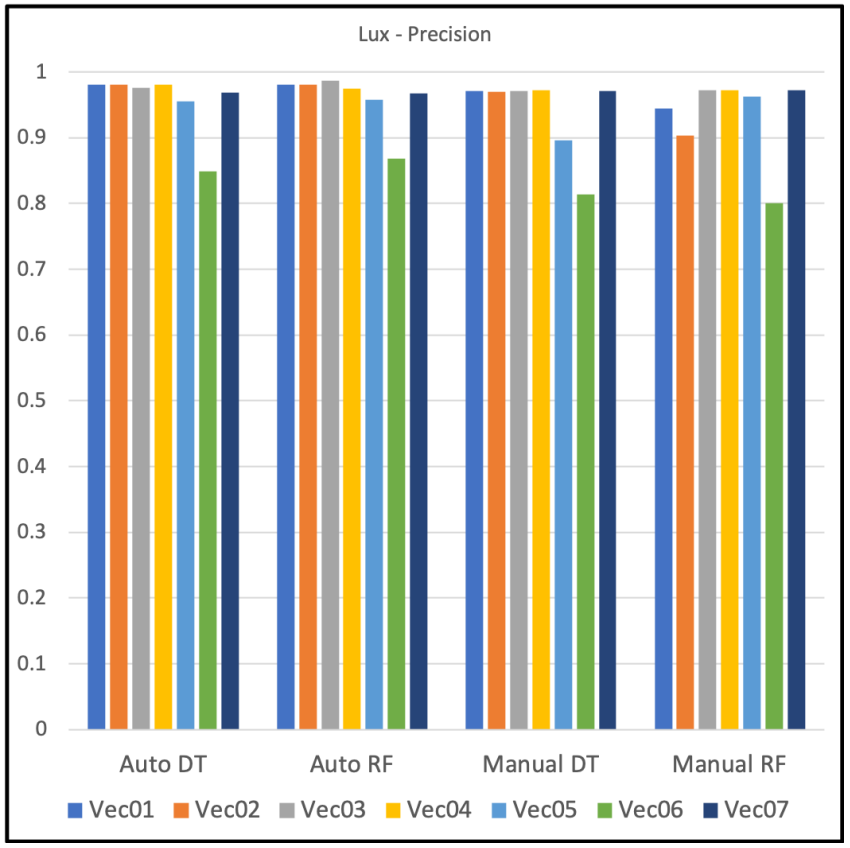


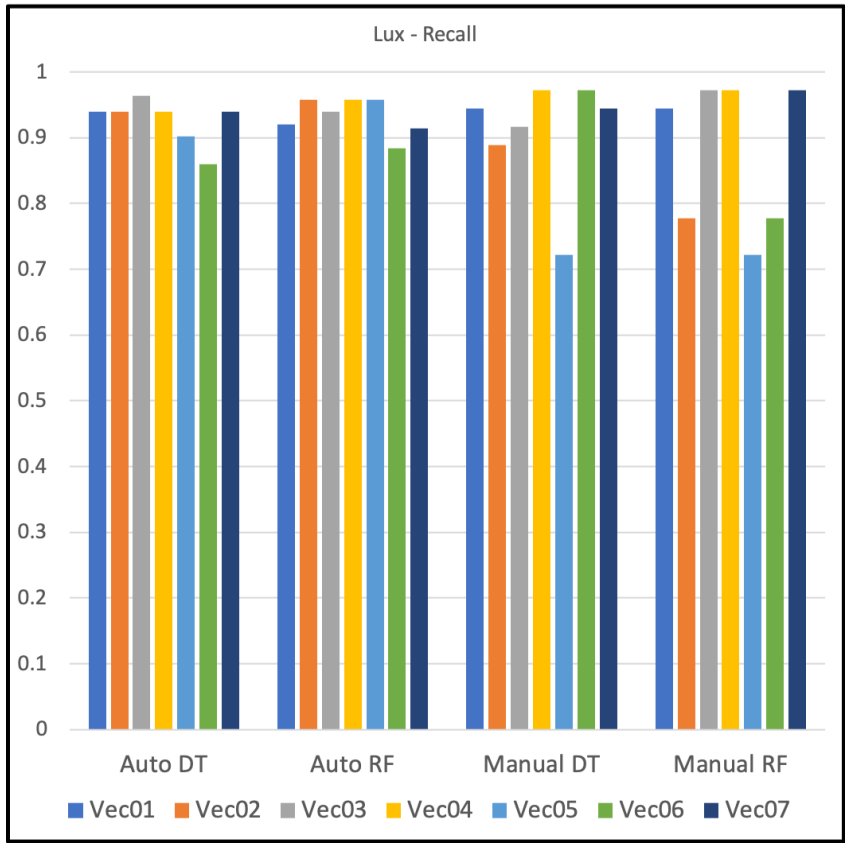
Auto	Manual
420	42

Vector	Features
Vec01	All 14 features
Vec02	No Words
Vec03	No Links
Vec04	No Density
Vec05	No Images
Vec06	No Structural
Vec07	Only Element

Results

LUX – https://luxtimes.lu/



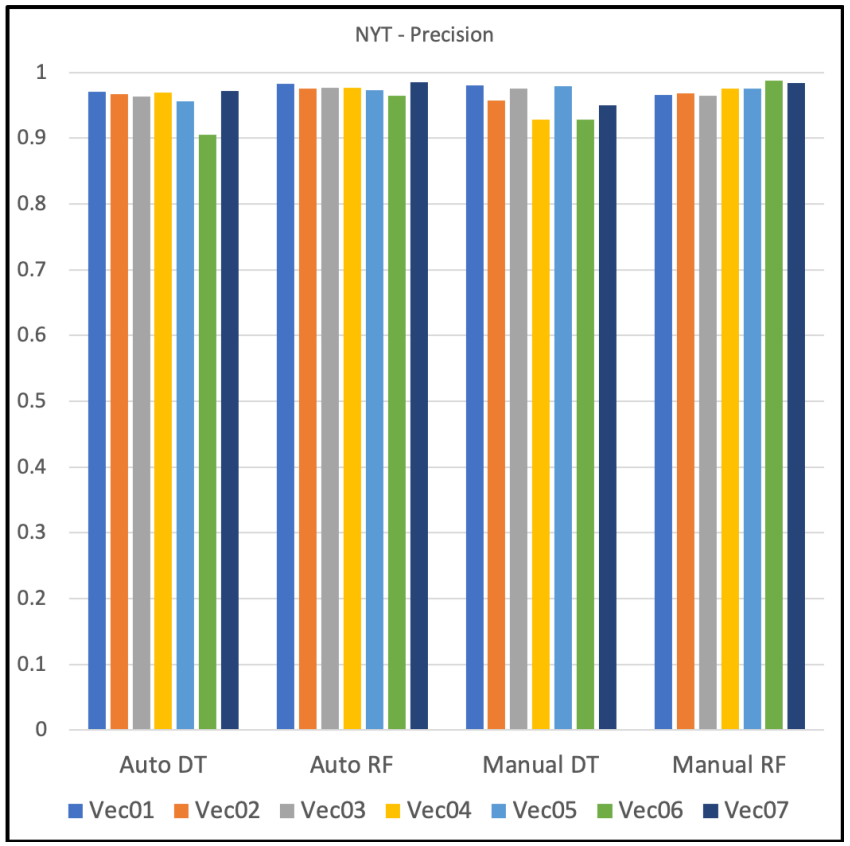


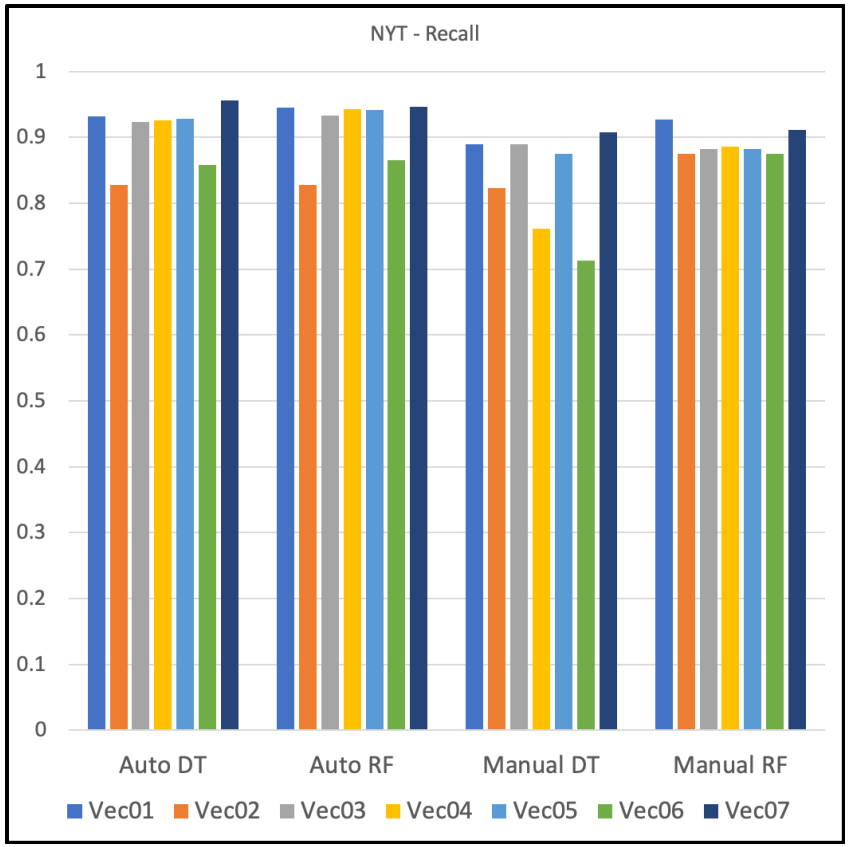
Auto	Manual
200	60

Vector	Features
Vec01	All 14 features
Vec02	No Words
Vec03	No Links
Vec04	No Density
Vec05	No Images
Vec06	No Structural
Vec07	Only Element

Results

NYT – https://www.nytimes.com/





Auto	Manual
340	62

Vector	Features
Vec01	All 14 features
Vec02	No Words
Vec03	No Links
Vec04	No Density
Vec05	No Images
Vec06	No Structural
Vec07	Only Element

Conclusion

Research Questions

Is it possible to build a Machine Learning model to extract the information from some web pages and analyse them in order to classify the best content part in a web page?

If so, how could we evaluate the performance or the precision of the model?

Is it possible to apply the model to all types of web pages?

What could be the best collection of features for analysing and classifying the web content?

Conclusion

Challenges

"Noise" pages in the datasets.

Labelling tasks require too much effort

Nutch documentation

Future Work

Improve Features

(mainly String Features)

Acquiring deeper knowledge in Cross-Validation k-fold in Spark

Thank You!

Find more: https://github.com/drendrin88/master-thesis