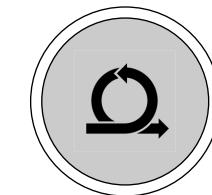
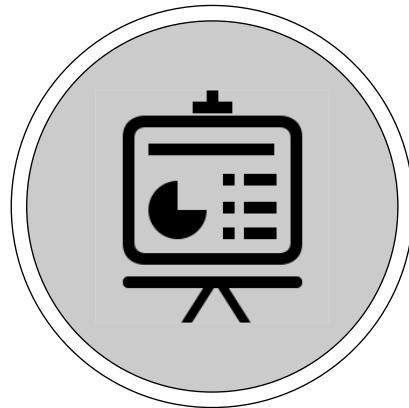




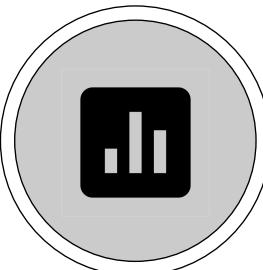
# Advanced Content Analysis for Web Pages in Apache Spark

# Dioni REBONATO ENDRINGER

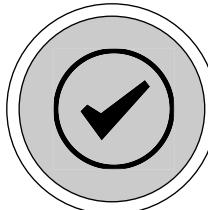
# Introduction



Methodology



Results



Conclusion



university of luxembourg receives grant from eu



Recherche Google

J'ai de la chance

Google disponible en : [Deutsch](#)



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...](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?

BBC Sign in News Sport Weather Shop Reel Travel More Search

## NEWS

Home Video World UK Business Tech Science Stories Entertainment & Arts Health World News TV More

**Airbus to stop making A380 superjumbo**

The aircraft manufacturer will cease production of its giant jet after key buyer Emirates cuts orders.

2h Business

Why did the Airbus A380 fail? Brexit uncertainty a disgrace, says Airbus



Trump ex-aide lied during plea deal



Paul Manafort "made multiple statements" to the special counsel and others, a US judge ruled.

3h US & Canada

IS schoolgirl 'wants to return to UK'



One of three teenagers who left the UK in 2015 to join the Islamic State group is living in a camp in Syria.

19m UK

UK MPs to debate Brexit next steps



Theresa May could face another defeat as some of her party's MPs may refuse to back the government.

3h UK Politics

The Swedish teen inspiring climate strikes



Greta Thunberg spoke to the BBC in September, but since then she's become a global phenomenon.

2:04

Senator 'wiped blood from leader's door'



Senator Brian Burston has become involved in a scuffle with Paula Hanson's aide.

1h Australia

Pedophile jailed for hundreds



Gay couples sue Japan over marriage rights



Chinese student held for pudding 'assault'



Churchill a villain, says senior UK MP



'Know your heart risk like your Pin code'



The village built from missiles



Philippines news boss bailed amid outcry



Germany narrowly avoids recession



Asian

Best schoolgirl 'wants to return to UK'



UK MPs to debate Brexit next steps



The Swedish teen inspiring climate strikes



Senator 'wiped blood from leader's door'



Pedophile jailed for hundreds



Gay couples sue Japan over marriage rights



Chinese student held for pudding 'assault'



Churchill a villain, says senior UK MP



'Know your heart risk like your Pin code'



The village built from missiles



Philippines news boss bailed amid outcry



Germany narrowly avoids recession



Asian

Best schoolgirl 'wants to return to UK'



UK MPs to debate Brexit next steps



The Swedish teen inspiring climate strikes



Senator 'wiped blood from leader's door'



Pedophile jailed for hundreds



Gay couples sue Japan over marriage rights



Chinese student held for pudding 'assault'



Churchill a villain, says senior UK MP



'Know your heart risk like your Pin code'



The village built from missiles



Philippines news boss bailed amid outcry



Germany narrowly avoids recession



Asian

CNN World U.S. Politics Business Entertainment Sport Travel Style Health Video

BREXIT WATCH Brexit happens in 43d 13h 12m

## Airbus to stop making A380

Top stories



Manafort bombshell deepens mystery in Russia probe

In focus

Will a smash change China

AIRBUS TO STOP MAKING A380

The largest passenger plane ever built promised to revolutionize air travel but failed to deliver on outsized expectations

Inside the Airbus A380 assembly line

4 million parts, 30 countries: How an A380 comes together

Airbus warns 'no deal' Brexit could threaten its existence

Around the world

Featured

Women in Japan rebel against

ANALYSIS Washington will sign

ANALYSIS Time before Brexit

NASA: China, India

US and India head

Where childhood c

Government tries t

Spotlight

SUBSCRIBE NOW

THURSDAY, FEB 14 2019 WEATHER TRAFFIC

Other Editions

LUXEMBOURG TIMES

HOME ECONOMICS LUXEMBOURG COMMUNITY EUROPEAN UNION WORLD CULTURE & LIFE

merger control



Margrethe Vestager's hopes of top EU role hit by rail deal veto

ANNUAL RESULTS Yesterday Luxembourg insurance gets Brexit boost

BREXIT Today 8:00 In Brexit's city of spies eavesdroppers are e

Top Stories

A la Carte

KOK

ENGLISH ESPAÑOL 中文

Thursday, February 14, 2019

Your Thursday Briefing Listen to 'The Daily' The 'In Her Words' Newsletter

am Your Thursday Briefing Let us help you start your day. Listen to 'The Daily' No heat, no power: How a federal jail failed its inmates. The 'In Her Words' Newsletter Girls get tech. They just need others to believe it.

PRESIDENT TRUMP Manafort Lied to Prosecutors in Russia Inquiry, Judge Rules

President Trump's former campaign chairman, Paul Manafort, was untruthful about several topics, a federal judge found, potentially leading to a longer sentence.

The judge said Mr. Manafort had intentionally lied about his contacts with a Russian associate before and after Mr. Trump was elected.

5h ago

Trump Puts Best Face on Border Deal, as Aides Try to Appease Angry Right

It was arguably the most punishing defeat the president has experienced in office, our correspondent writes in an analysis.

10h ago

House Votes to Halt Aid for Saudi Arabia's War in Yemen

The vote, coming amid outrage over the killing of the journalist Jamal Khashoggi, was a defiant and rare move to curb presidential war powers.

6h ago

Opinion > Elka Kurniawan Indonesia's Next Election in April. The Islamists Have Already Won. How religion has come to dominate our politics.

3h ago

Ryan Adams Dangled Success, Women Say They Paid a Price

Several women say Adams offered to jumpstart their music careers, then pursued them sexually and in some cases retaliated when they spurned him.

In interviews, seven women and more than a dozen associates described a pattern of manipulative behavior. Adams denies the claims.

9h ago

SCHOOL SHOOTINGS A Year After the School Shooting That Was Supposed to Change Everything

Ryan Adams in 2017. A prolific singer-songwriter with his own label, he is well-known to communicate directly with fans and critics on social media. Elizabeth Weinstein for The New York Times

Gail Collins The Answer Is Blowing in the Wind Of Stalin, Hitler and Environmental Protection

8h ago

Amy Barnhorst A New Model to Stop Next School Shooting

Nicholas Kristof Navigating the Male-Female Work Relationship

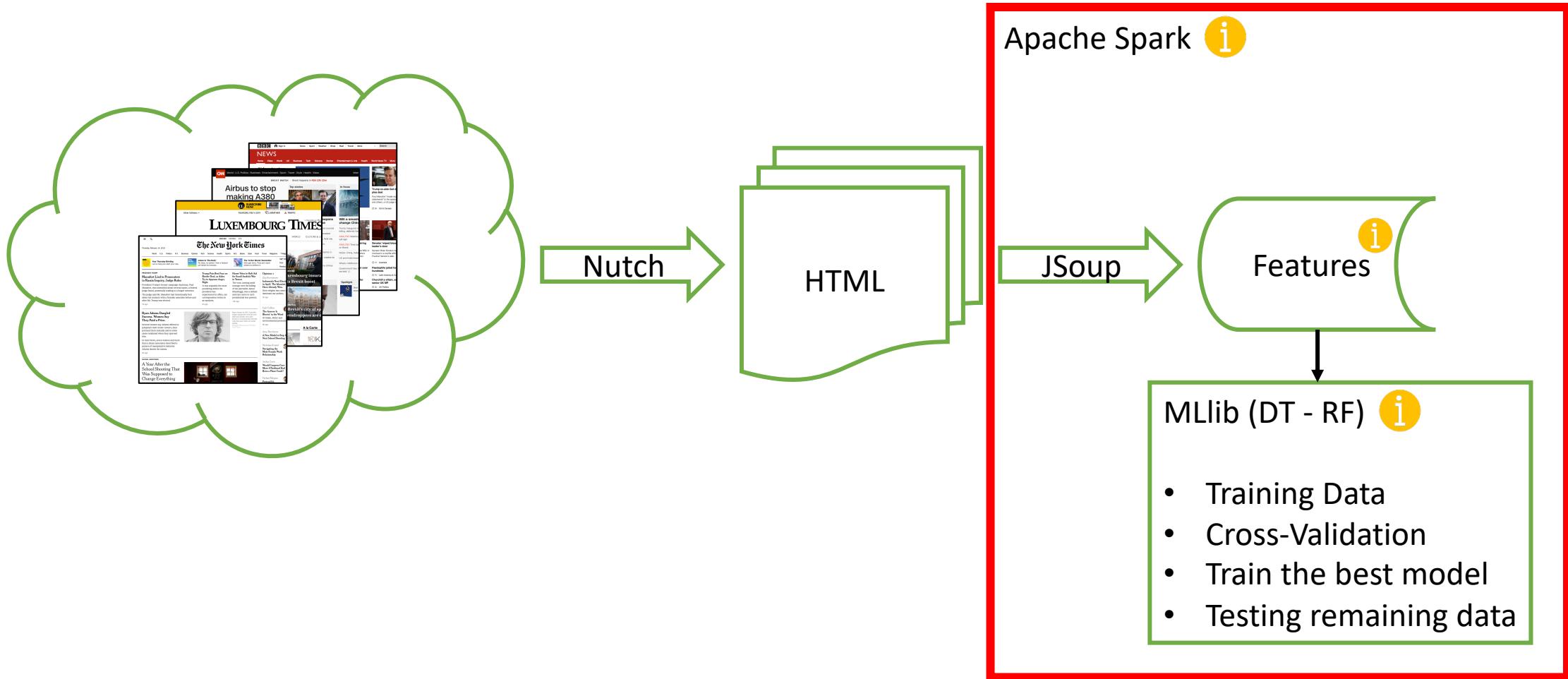
Jaclyn Corin Would Congress Care More if Parkland Had Been a Plane Crash?

Farhad Manjoo Pretend It's

# 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

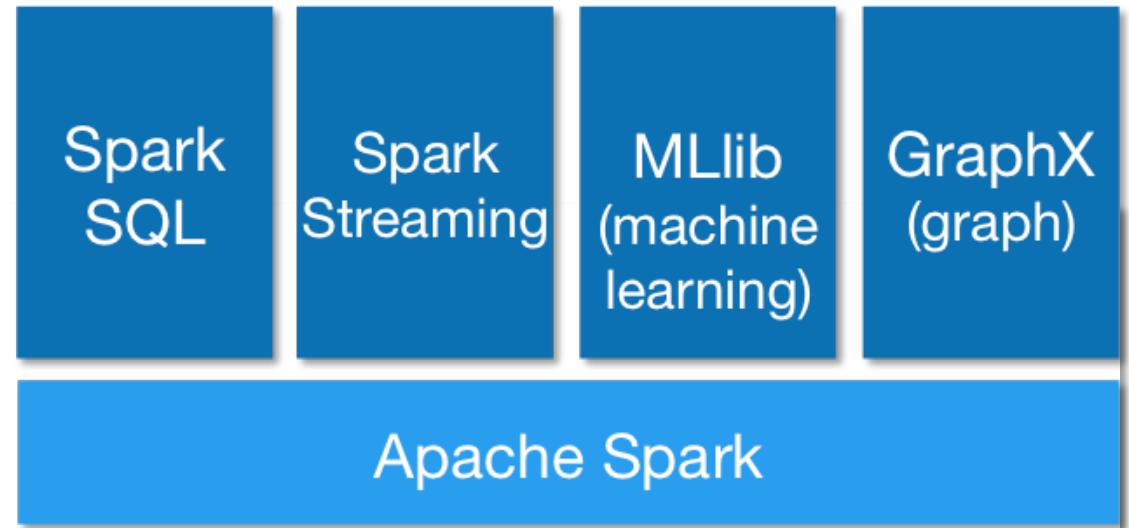
**Applications:** 1 [Running](#), 0 [Completed](#)

**Drivers:** 0 Running, 0 Completed

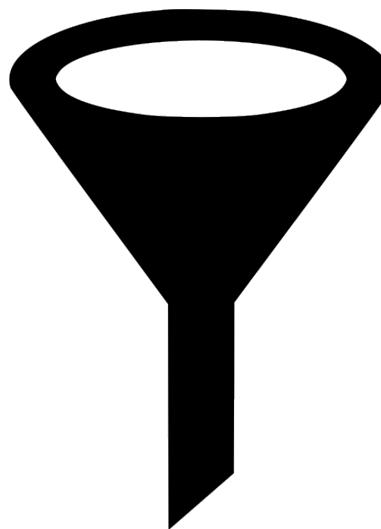
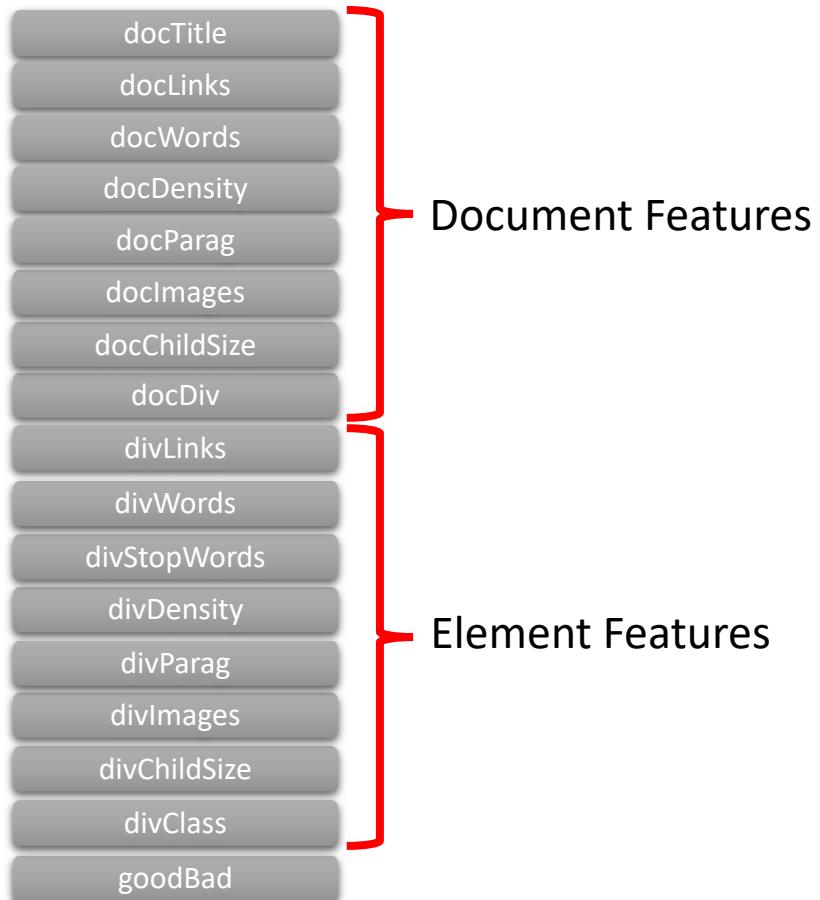
**Status:** ALIVE

## ▼ Workers (8)

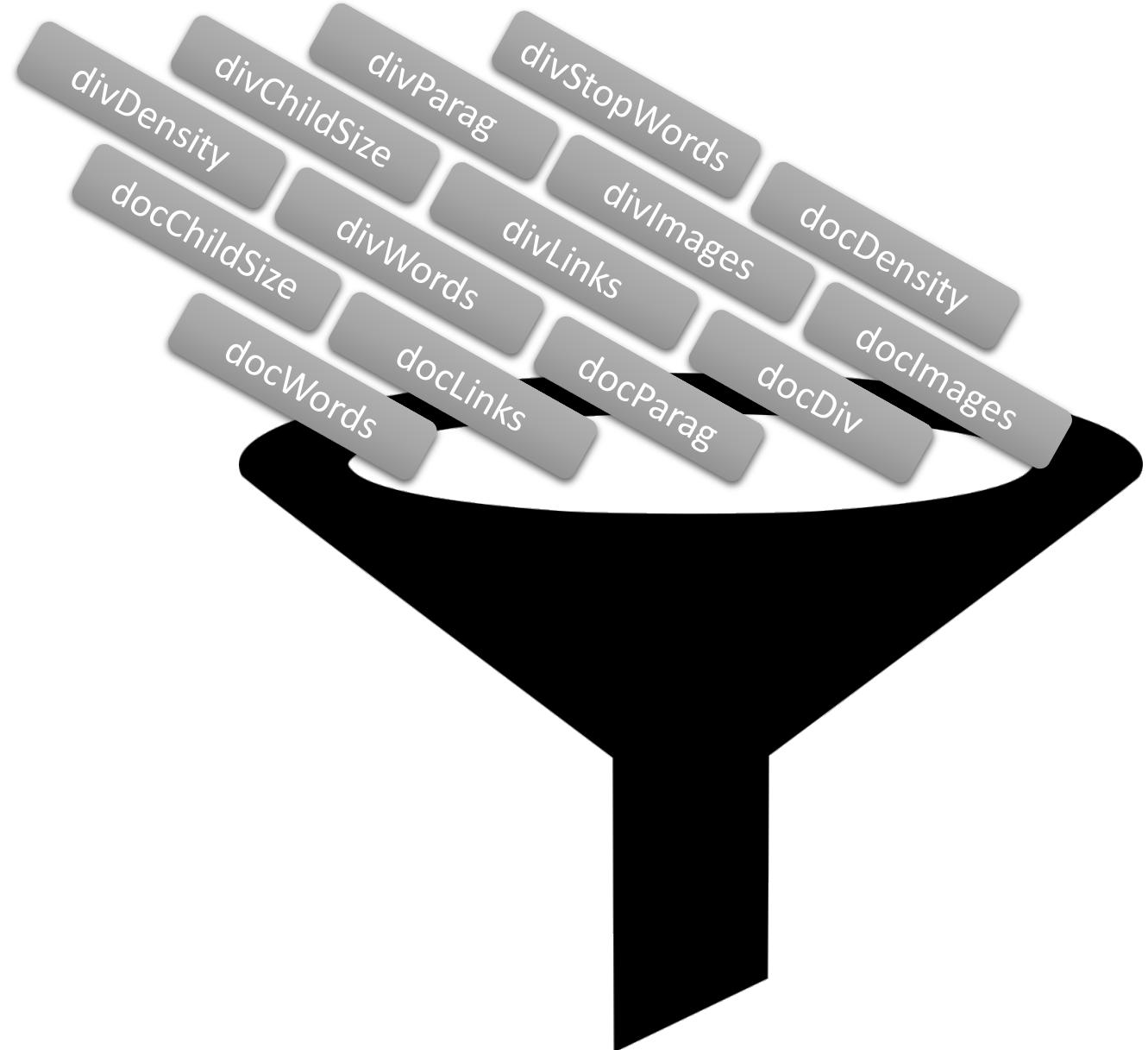
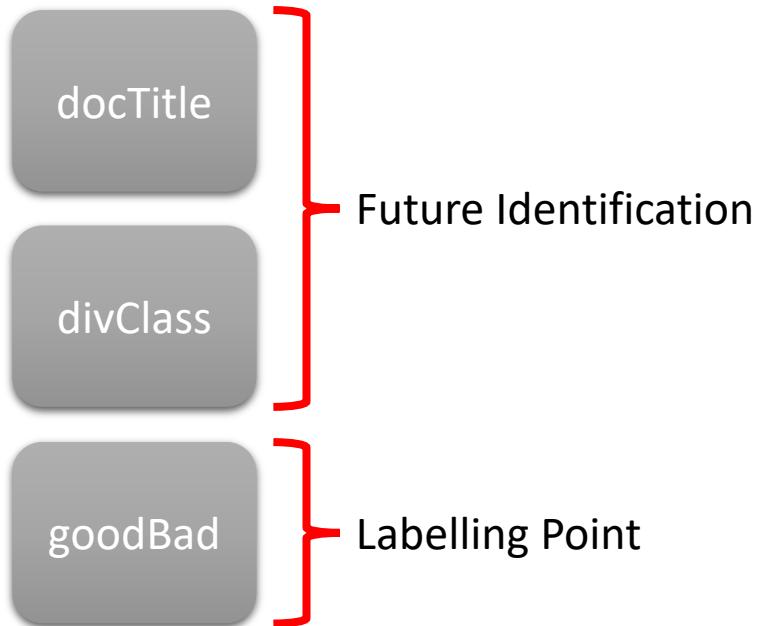
| Worker Id  | Address            | State | Cores        | Memory                   |
|--|--------------------|-------|--------------|--------------------------|
| <a href="#">worker-20190421011021-172.17.6.153-33413</a> | 172.17.6.153:33413 | ALIVE | 28 (28 Used) | 124.6 GB (120.0 GB Used) |
| <a href="#">worker-20190421011022-172.17.6.149-42645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



# Extracted Features



# Extracted Features



# Labelling Dataset

at the tax system and said you need to change the tax laws. You can make a

`div.css-1fanzo5.StoryBodyCompanionColumn-XXXgood` 615 x 758 .”

On Saturday, after further inquiries from The Times, a lawyer for the president, Charles J. Harder, wrote that the tax information was “demonstrably false,” and that the paper’s statements “about the president’s tax returns and business from 30 years ago are highly inaccurate.” He cited no specific errors, but on Tuesday added that “I.R.S. transcripts, particularly before the days of electronic filing, are notoriously inaccurate” and “would not be able to provide a reasonable picture of any taxpayer’s return.”

Mark J. Mazur, a former director of research, analysis and statistics at the I.R.S., said that, far from being considered unreliable, data used to create such transcripts had undergone quality control for decades and had been used to analyze economic trends and set national policy. In addition, I.R.S. auditors often refer to the transcripts as “handy” summaries of tax returns, said Mr. Mazur, now director of the nonpartisan Urban-Brookings Tax Policy Center in Washington.

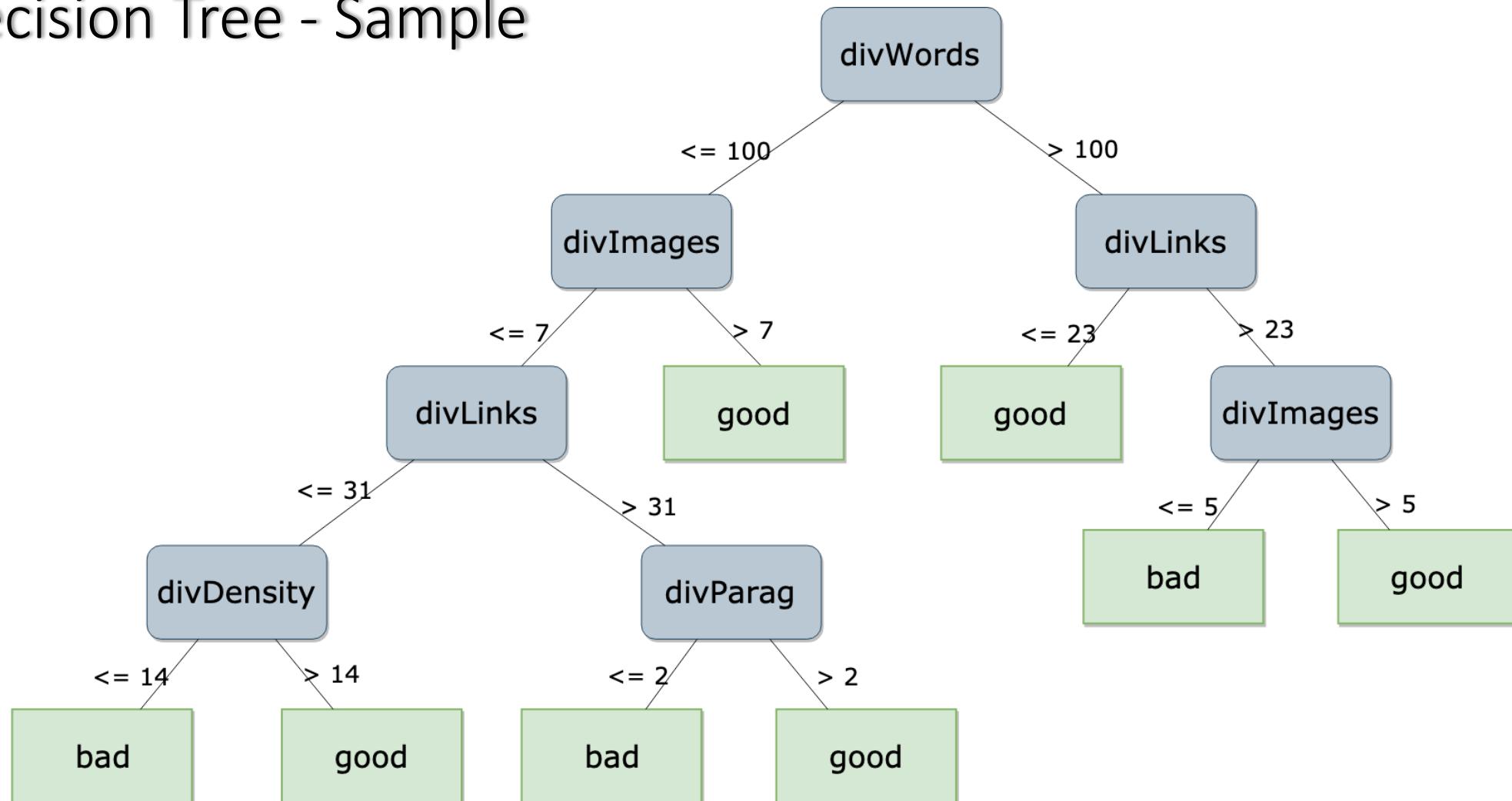
In fact, the source of The Times’s newly obtained information was able to provide several years of unpublished tax figures from the president’s father, the builder Fred C. Trump. They matched up with Fred Trump’s actual returns, which had been obtained by The Times in the earlier investigation.

If the new tax information does not offer a new narrative of Mr. Trump’s

```
</div>
<meta itemprop="isAccessibleForFree" content="false">
▶<span itemprop="isPartOf" itemscope itemtype="http://schema.org/CreativeWork http://schema.org/Product">...</span>
▼<article id="story" class="css-1vxca1d e1qksbf0">
  ▶<div id="top-wrapper" class="css-1sy8kpn">...</div>
  ▶<span itemprop="hasPart" itemscope itemtype="http://schema.org/WebPageElement">...
  </span>
  ▶<header class="css-1n5gntz e12qa4dv1">...</header>
  ▶<section name="articleBody" itemprop="articleBody" class="meteredContent css-1j2v565">
    ▶<div class="css-1fanzo5 StoryBodyCompanionColumn-XXXgood"></div>
    ▶<div class="css-1fanzo5 StoryBodyCompanionColumn-XXXgood">
      ▶<div class="css-53u6y8-XXXgood">
        ▶<p class="css-1ygdjhk evys1bk0">...</p>
        ▶<p class="css-1ygdjhk evys1bk0">...</p>
        ▶<p class="css-1ygdjhk evys1bk0">...</p>
        ▶<p class="css-1ygdjhk evys1bk0">...</p>
        <h2 class="css-872od7 e00vm40" id="link-3b8aa309">1. Mr. Trump was deep in the red even as he peddled deal-making advice</h2>
        ▶<p class="css-1ygdjhk evys1bk0">...</p>
      </div>
      <aside class="css-o6xoe7"></aside>
    </div>
    ▶<div class="css-1fanzo5 StoryBodyCompanionColumn-XXXgood"></div>
    ▶<div class="css-1fanzo5 StoryBodyCompanionColumn-XXXgood"></div>
    ▶<div class="css-1fanzo5 StoryBodyCompanionColumn-XXXgood"></div>
  </section>
  ▶<div class="bottom-of-article">...</div>
  ▶<div class="css-12qsquia">...</div>
  ▶<div>...</div>
</article>
▶<div class="css-1lc20wh">...</div>
</div>
```

html.story body

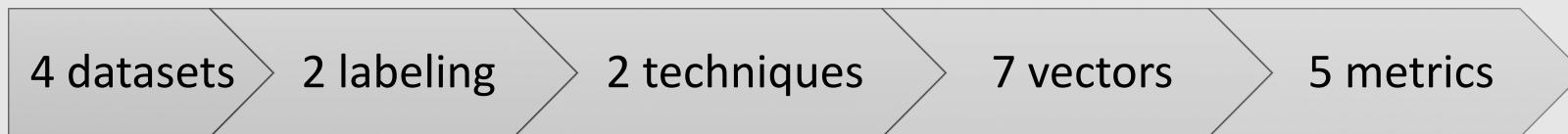
# Decision Tree - Sample



# Datasets

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

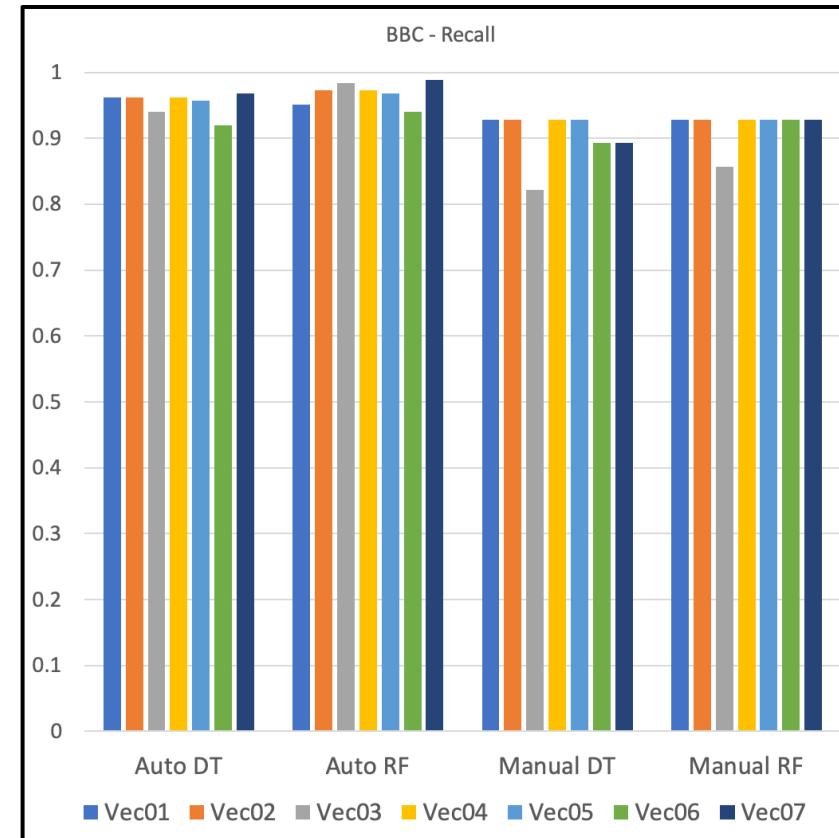
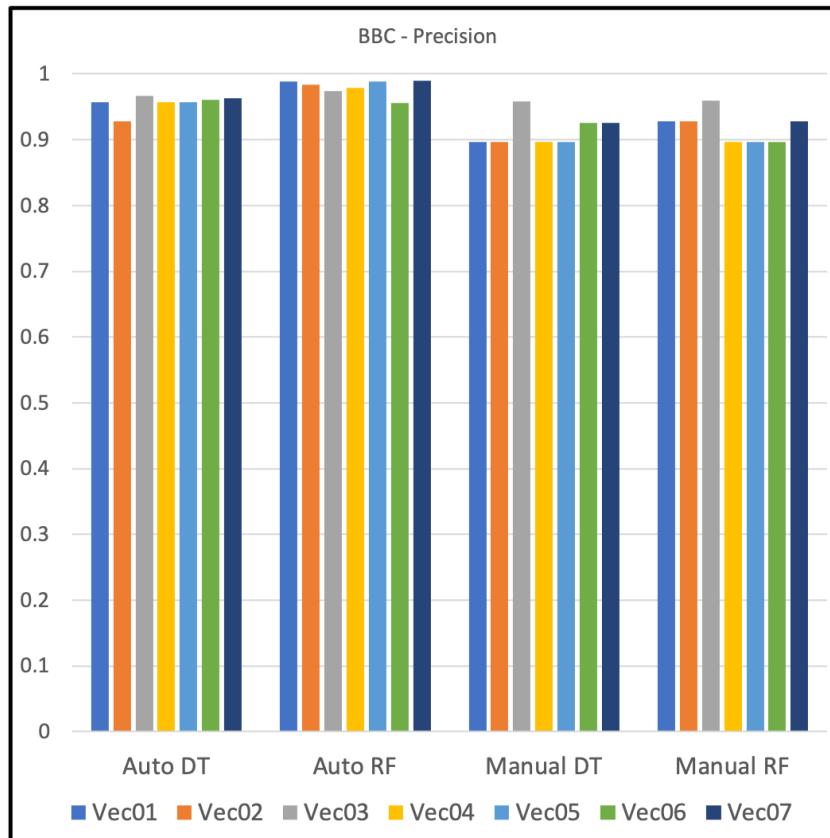
| Vector | Features                               |
|--------|--|
| Vec01  | All 14 features (document and element) |
| Vec02  | Removed Words features                 |
| Vec03  | Removed Links features                 |
| Vec04  | Removed Density features               |
| Vec05  | Removed Images features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



**560  
results**

# Results

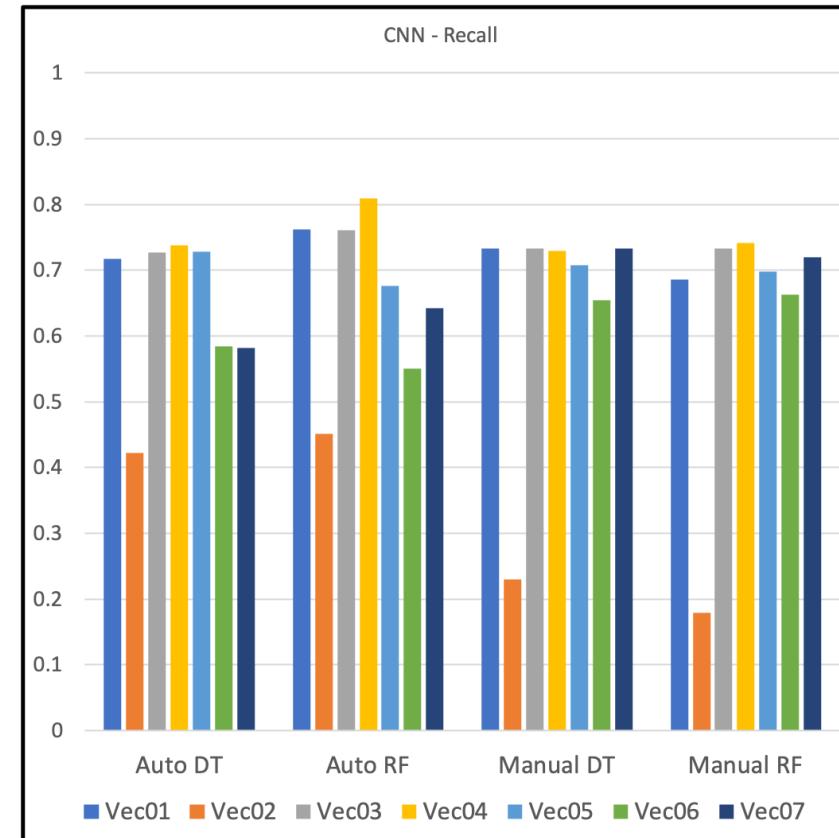
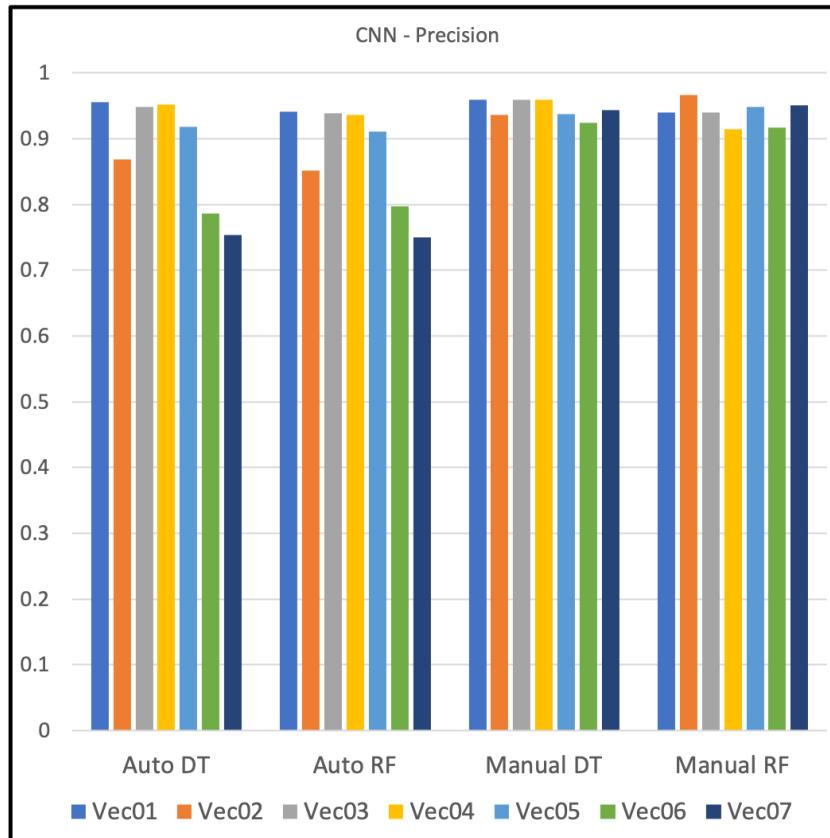
BBC – <https://www.bbc.com/news>



| 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/>

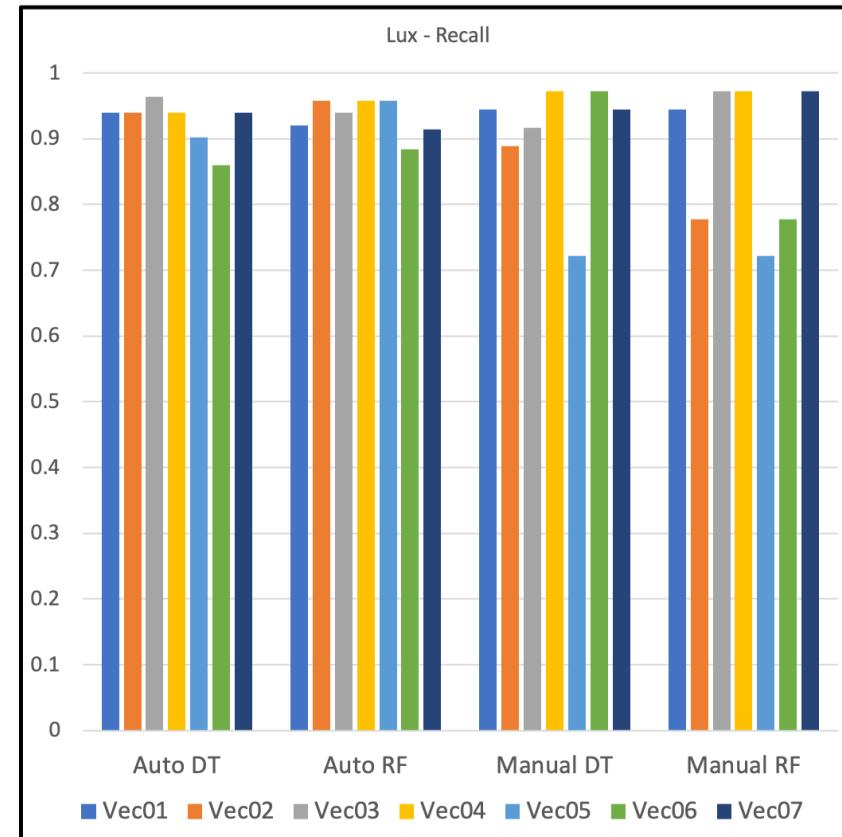
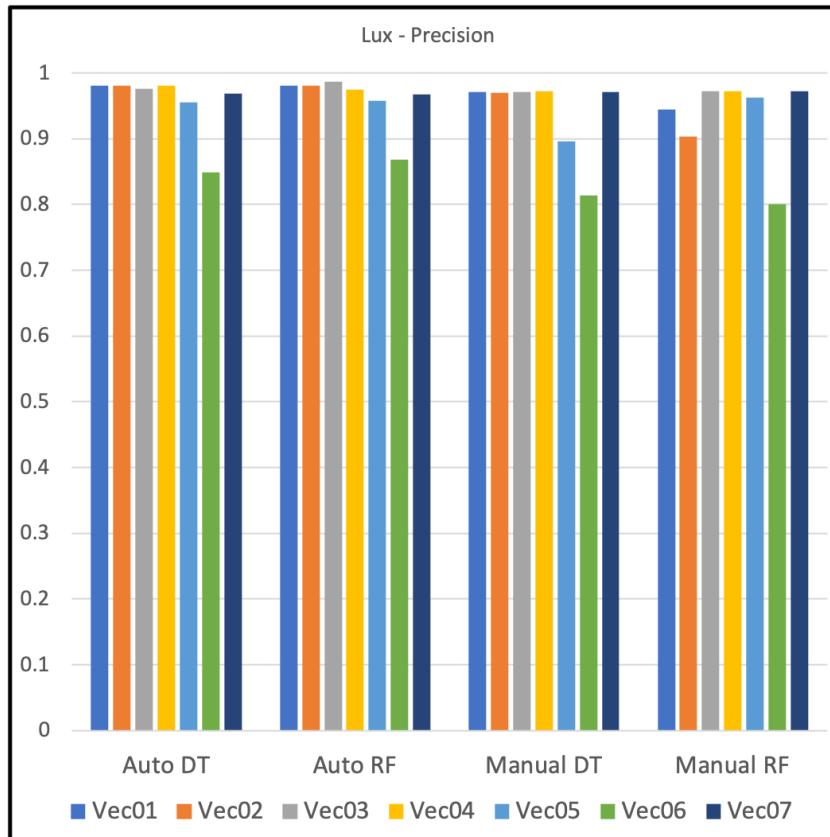


| Auto | Manual |
|------|--------|
| 420  | 42     |

| 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

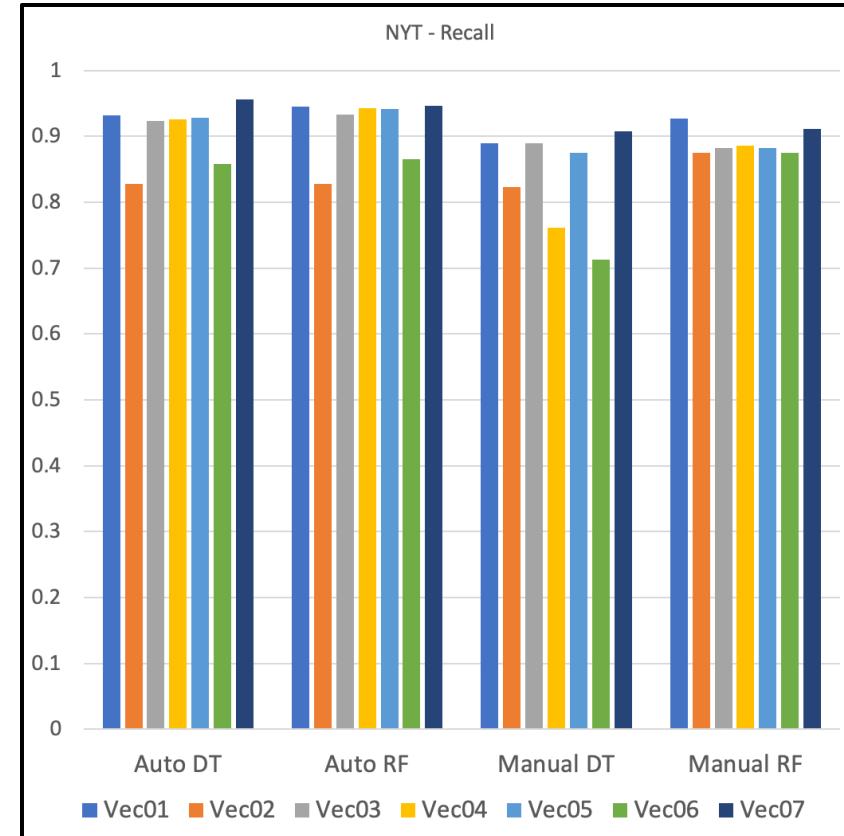
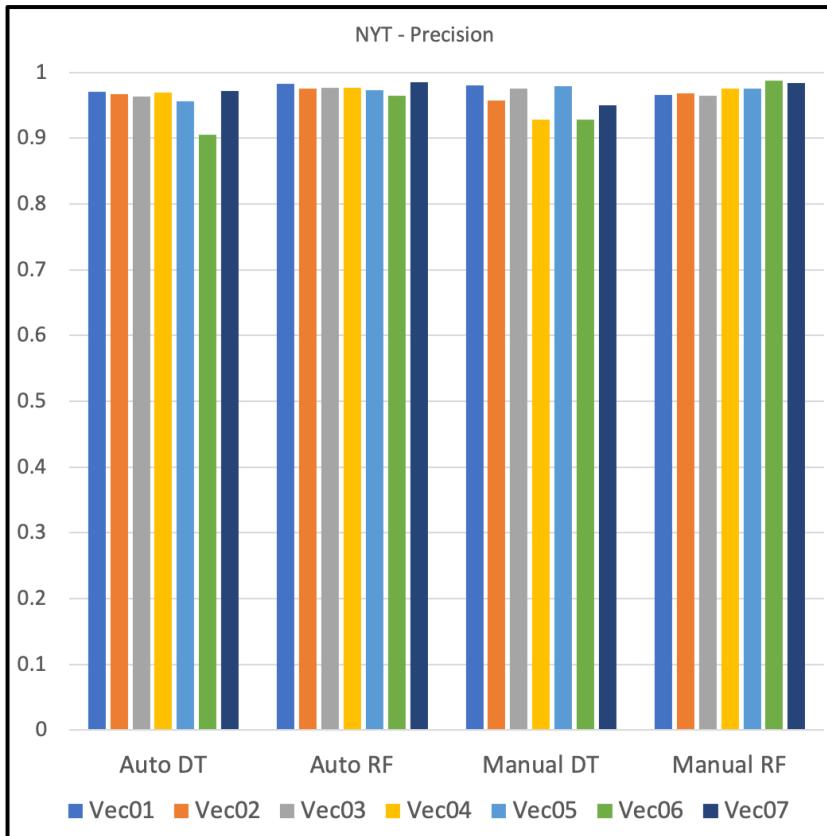
LUX – <https://luxtimes.lu/>



| 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>