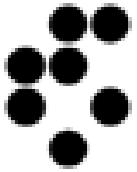


URL Tree: Efficient Unsupervised Content Extraction from Streams of Web Documents

Borut Sluban and Miha Grčar

Jožef Stefan Institute, Slovenia

Identifying Content

27 October 2013 Last updated at 17:49 ET

Share

Twitter share sale trial 'a success', says exchange

The New York Stock Exchange (NYSE) says a test run of Twitter's share sale was 'a success', as it tries to avoid the debacle surrounding Facebook's flotation on the Nasdaq.

Twitter says it has over 200 million active users but has yet to turn a profit

Traders simulated buying and selling shares on the exchange on Saturday, to try to clear up any technical hitches that it may face when shares go public.

Twitter will be the biggest tech company to go public since Facebook.

It is thought that trading will begin in early November.

In May, Facebook's first day on a rival exchange, the Nasdaq, was fraught with problems. A glitch in the system meant that traders did not know for hours, and in some cases days, whether or not their trades had successfully gone through.

The Nasdaq was later fined $10m (£6.18m) by the regulator, the Securities and Exchange Commission, for the problems.

The New York Stock Exchange is keen to avoid a similar fate.

In a statement, a spokeswoman for the exchange said: " This morning's systems test was successful, and we're grateful to all the firms that chose to participate.

"We're being very methodical in our planning for Twitter's IPO [initial public offering], and are working together with the industry to ensure a world-class experience for Twitter, retail investors and all market participants."

Last week Twitter said it planned to sell 70 million shares priced between $17 and $20 (£10 - £12) to raise up to $1.4bn (£865m).

According to its IPO documents, Twitter now has 218 million monthly users and 500 million tweets are sent a day.

However, all those users and tweets have not yet resulted in a profit.

Twitter made a loss of $69m in the first six months of 2013, on revenues of $254m.

Related Stories

Twitter float could raise $1.4bn

NYSE beats Nasdaq in Twitter listing

How does Twitter make money?

Top Stories

US admits spying constraints needed

Thirteen die as storm crosses Europe

Libyan van robbers snatch $54m

Top Islamists 'killed' in Somalia

Sahara migrants 'die of thirst'

Features

$90 a month

Why do Americans pay so much for home broadband?

Macho mayor

A Russian politician at war with drugs - and ready to break the law

No Woman, No Drive

Bob Marley's hit reworked as Saudi women defy driving ban

Day in pictures

Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Storm: Tree crushes Amsterdam woman

2 Top Islamists 'killed' in Somalia

3 Why is broadband more expensive in the US?

4 Accused ex-BBC driver found dead

5 Woman flown to Grenada, not Granada

6 US: Spying 'constraints needed'

Identifying Content

27 October 2013 Last updated at 17:49 ET

Share

Twitter share sale trial 'a success', says exchange

The New York Stock Exchange (NYSE) says a test run of Twitter's share sale was 'a success', as it tries to avoid the debacle surrounding Facebook's flotation on the Nasdaq.

Twitter says it has over 200 million active users but has yet to turn a profit

Traders simulated buying and selling shares on the exchange on Saturday, to try to clear up any technical hitches that it may face when shares go public.

Twitter will be the biggest tech company to go public since Facebook.

It is thought that trading will begin in early November.

In May, Facebook's first day on a rival exchange, the Nasdaq, was fraught with problems. A glitch in the system meant that traders did not know for hours, and in some cases days, whether or not their trades had successfully gone through.

The Nasdaq was later fined $10m (£6.18m) by the regulator, the Securities and Exchange Commission, for the problems.

The New York Stock Exchange is keen to avoid a similar fate.

In a statement, a spokeswoman for the exchange said: " This morning's systems test was successful, and we're grateful to all the firms that chose to participate.

"We're being very methodical in our planning for Twitter's IPO [initial public offering], and are working together with the industry to ensure a world-class experience for Twitter, retail investors and all market participants."

Last week Twitter said it planned to sell 70 million shares priced between $17 and $20 (£10 - £12) to raise up to $1.4bn (£865m).

According to its IPO documents, Twitter now has 218 million monthly users and 500 million tweets are sent a day.

However, all those users and tweets have not yet resulted in a profit.

Twitter made a loss of $69m in the first six months of 2013, on revenues of $254m.

Related Stories

- Twitter float could raise $1.4bn
- NYSE beats Nasdaq in Twitter listing
- How does Twitter make money?

Top Stories

US admits spying constraints needed

- Thirteen die as storm crosses Europe
- Libyan van robbers snatch $54m
- Top Islamists 'killed' in Somalia
- Sahara migrants 'die of thirst'

Features

$90 a month
Why do Americans pay so much for home broadband?

Macho mayor
A Russian politician at war with drugs - and ready to break the law

No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban

Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

Storm: Tree crushes Amsterdam woman	1
Top Islamists 'killed' in Somalia	2
Why is broadband more expensive in the US?	3
Accused ex-BBC driver found dead	4
Woman flown to Grenada, not Granada	5
US: Spying 'constraints needed'	6

Identifying Content

type="text/javascript">$render("advert","advert-leaderboard");</script> <script type="text/javascript">$render("advert-post-script-load");</script> <!-- START CPS_SITE CLASS: story --> <!-- No EWA -->   25 September 2011 Last updated at 10:06 GMT   <h3>Share this page</h3> <li class="delicious"> Delicious <li class="digg"> Digg <li class="facebook"> Facebook <li class="reddit"> reddit <li class="stumbleupon"> StumbleUpon <li class="twitter"> Twitter <li class="email"> Email <li class="print"> Print   <!-- Social media icons by Paul Annet | http://nicepaul.com/icons --> <script type="text/javascript"> <!-- $render("page-bookmark-links","page-bookmark-links-head",{ useForgeShareTools:"true", position:"top", site:'News', headline:'BBC News - Anti-Wall Street protesters arrested in New York', storyId:'15051554', sectionId:'99127', url:'http://www.bbc.co.uk/news/world-us-canada-15051554', edition:'International' }); --> </script> <h1 class="story-header">Anti-Wall Street protesters arrested in New York</h1> The protesters' placards carried demands about a range of causes At least 80 people have been arrested during an anti-Wall Street march in New York's financial district. Several hundred people took part in Saturday's march, which was intended to draw attention to "corporate greed and corrupt politics" in the US. Participants carried banners supporting a range of other issues, including healthcare reform, an end to US wars and the scrapping of the death penalty. The march came after a week of protests by the Occupy Wall Street campaign. The loosely organised group says it is defending 99% of the US population against the wealthiest 1%, and had called for 20,000 people to "flood into lower Manhattan" on 17 September and remain there for "a few months". Protesters, who are mostly young, initially numbered some 1,500 but their numbers had fallen to about 200 by Saturday's march. There was a heavy security presence in the district, with police deploying nets to block off major roads including Fifth Avenue and to protect the New York Stock Exchange. Corporate 'enemy' One protester, 21-year-old Ryan Reed, said he joined in "because what I see - and what I feel most people in this country see - is an economy and a system that's collapsing". Police said most of the arrests were for disorderly conduct "The enemy is the big business leaders of Wall Street, the big oil company leaders, the coal company leaders, the big military industrial leaders." A number of placards also called for "justice for Troy Davies", the US man executed in Georgia last week amid widespread criticism. Police said most of Saturday's arrests were for disorderly conduct and blocking traffic, but one person was charged with assaulting a police officer. One officer also suffered a shoulder injury, said police. They have

Streaming setting

BBC News Sport Weather Capital Culture Autos TV Radio More... Search

NEWS BUSINESS

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health | Sci/Environment | Tech | Entertainment | Video

Market Data | Economy | Entrepreneurship | Business of Sport | Companies | Technology of Business | Knowledge Economy

27 October 2013 Last updated at 17:49 ET

Share

Twitter share sale trial 'a success', says exchange

The New York Stock Exchange (NYSE) says a test run of Twitter's share sale was 'a success', as it tries to avoid the debacle surrounding Facebook's flotation on the Nasdaq.

Twitter says it has over 200 million active users but has yet to turn a profit

Traders simulated buying and selling shares on the exchange on Saturday, to try to clear up any technical hitches that it may face when shares go public.

Twitter will be the biggest tech company to go public since Facebook.

It is thought that trading will begin in early November.

In May, Facebook's first day on a rival exchange, the Nasdaq, was fraught with problems. A glitch in the system meant that traders did not know for hours, and in some cases days, whether or not their trades had successfully gone through.

The Nasdaq was later fined $10m (£6.18m) by the regulator, the Securities and Exchange Commission, for the problems.

The New York Stock Exchange is keen to avoid a similar fate.

In a statement, a spokeswoman for the exchange said: " This morning's systems test was successful, and we're grateful to all the firms that chose to participate.

"We're being very methodical in our planning for Twitter's IPO [initial public offering], and are working together with the industry to ensure a world-class experience for Twitter, retail investors and all market participants."

Last week Twitter said it planned to sell 70 million shares priced between $17 and $20 (£10 - £12) to raise up to $1.4bn (£865m).

According to its IPO documents, Twitter now has 218 million monthly users and 500 million tweets are sent a day.

However, all those users and tweets have not yet resulted in a profit.

Twitter made a loss of $69m in the first six months of 2013, on revenues of $254m.

Related Stories

- Twitter float could raise $1.4bn
- NYSE beats Nasdaq in Twitter listing
- How does Twitter make money?

Top Stories

US admits spying constraints needed

- Thirteen die as storm crosses Europe
- Libyan van robbers snatch $54m
- Top Islamists 'killed' in Somalia
- Sahara migrants 'die of thirst'

Features

$90 a month
Why do Americans pay so much for home broadband?

Macho mayor
A Russian politician at war with drugs - and ready to break the law

No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban

Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1. Storm: Tree crushes Amsterdam woman
2. Top Islamists 'killed' in Somalia
3. Why is broadband more expensive in the US?
4. Accused ex-BBC driver found dead
5. Woman flown to Grenada, not Granada
6. US: Spying 'constraints needed'

Streaming setting

BBC News Sport Weather Capital Culture Autos TV Radio More... Search

NEWS BUSINESS

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health | Sci/Environment | Tech | Entertainment | Video

Market Data | Economy | Entrepreneurship | Business of Sport | Companies | Technology of Business | Knowledge Economy

28 October 2013 Last updated at 19:04 ET

Share

Apple profits decline despite iPhone sales boost

Apple has finished releasing its new products for the year

Apple has reported a fourth quarter profit of $7.5bn (£4.6bn), beating analysts' expectations.

Although iPhone sales surged 26% to 33.8 million, profits slid for a third consecutive quarter as margins fell.

On an annual basis, Apple ended its fiscal year with its first earnings decline in 11 years as net profit slipped to $37bn.

Nonetheless, Apple boss Tim Cook insisted that "business was stronger than ever" on an earnings call.

But analysts were disappointed by forward guidance given by the company, which indicated that future profits might continue to slide.

Apple said it expected revenue next year to be somewhere between $55 - $58bn.

It indicated that the gross margin on its products - the difference between the amount the company spends to make a products versus how much consumers pay - would be between 36.5% - 37.5%, less than analysts were expecting.

Peter Oppenheimer, Apple's chief financial officer, explained that this was partly because it cost more to manufacture the company's new iPads and Macbooks and that prices weren't boosted accordingly.

Related Stories

Samsung Electronics profit jumps 26%

Mac v PC - the battle resumes

Apple unveils slimmer iPad Air

Top Stories

US admits spying constraints needed

Thirteen die as storm crosses Europe

Libyan van robbers snatch $54m

Top Islamists 'killed' in Somalia

Sahara migrants 'die of thirst'

Features

$90 a month
Why do Americans pay so much for home broadband?

Macho mayor
A Russian politician at war with drugs - and ready to break the law

No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban

Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Storm: Tree crushes Amsterdam woman
2 Top Islamists 'killed' in Somalia
3 Why is broadband more expensive in the US?
4 Accused ex-BBC driver found dead
5 Woman flown to Grenada, not Granada
6 US: Spying 'constraints needed'

Streaming setting

BBC News Sport Weather Capital Culture Autos TV Radio More... Search

NEWS BUSINESS

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health | Sci/Environment | Tech | Entertainment | Video

Market Data | Economy | Entrepreneurship | Business of Sport | Companies | Technology of Business | Knowledge Economy

28 October 2013 Last updated at 06:51 ET

Share

Rio Tinto to sell stake in Australian coal mine for $1bn

Rio Tinto is to sell its 50.1% share in Australia's Clermont thermal coal mine to Swiss commodities giant Glencore Xstrata and Japan's Sumitomo.

Rio says it remains committed to mining in Queensland

Anglo-Australian mining giant Rio will receive $1.02bn (£630m) from the sale of the facility in Queensland state.

"The sale... will allow us to realise value for our shareholders as we continue optimising our portfolio," a Rio statement said.

Coal mined from Clermont is burned to produce electricity.

Rio said it remained "committed to a long-term future in central Queensland" at its other mines.

The firm has announced or completed divestments so far this year of some $2.92bn.

Sell-offs include a controlling stake in Australia's Northparkes gold and copper mine to China Molybdenum, the sale of a majority holding in South African copper producer Palabora to a Chinese consortium and transfer of the Eagle nickel and copper project in the US to Canada-based Lundin Mining.

This summer Rio reported a 71% fall in half-year profits, partly due to reduced demand from China.

Earlier in October, it said it was making good progress on a cost-cutting programme.

Related Stories

Glencore takes $7.7bn Xstrata hit

Rio halts aluminium business sale

More on This Story

Related Stories

Glencore takes $7.7bn Xstrata hit 20 AUGUST 2013, BUSINESS

Rio halts aluminium business sale 08 AUGUST 2013, BUSINESS

From other news sites

Top Stories


US admits spying constraints needed

Thirteen die as storm crosses Europe

Libyan van robbers snatch $54m

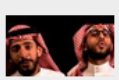
Top Islamists 'killed' in Somalia

Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Thirteen die as storm crosses Europe

2 Woman flown to Grenada, not Granada

3 Trending: No Woman, No Drive

4 US admits spying constraints needed

5 Why is broadband more expensive in the US?

6 Four die as storm hits southern UK

Top Stories


US admits spying constraints needed

Thirteen die as storm crosses Europe

Libyan van robbers snatch $54m

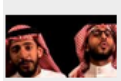
Top Islamists 'killed' in Somalia

Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Storm: Tree crushes Amsterdam woman

2 Top Islamists 'killed' in Somalia

3 Why is broadband more expensive in the US?

4 Accused ex-BBC driver found dead

5 Woman flown to Grenada, not Granada

6 US: Spying 'constraints needed'

Top Stories


US admits spying constraints needed

Thirteen die as storm crosses Europe

Libyan van robbers snatch $54m

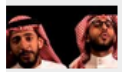
Top Islamists 'killed' in Somalia

Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Storm: Tree crushes Amsterdam woman

2 Top Islamists 'killed' in Somalia

3 Why is broadband more expensive in the US?

4 Accused ex-BBC driver found dead

5 Woman flown to Grenada, not Granada

6 US: Spying 'constraints needed'

Streaming setting

28 October 2013 Last updated at 10:06 ET

Burger King profits up as costs fall

Burger King sales outside North America jumped

Fast food vendor Burger King has reported higher profits as sales increased outside the US.

The US restaurant chain said its better-than-expected profits for the July-to-September period of this year were due to cost reductions and more franchising.

Burger King's net profit rose to $68.2m (£42.2m), as against $6.6m in the same period a year earlier.

However, sales at its North American branches fell 0.3%.

The company has recently introduced Satisfries, a lower-fat variety of chips, in the US and Canada, in an effort to offer a healthier alternative after criticisms of the high fat content of its food,

Costs cut

The firm, which is the third-largest burger chain in the US, behind rivals McDonalds and Wendy's, said it has succeeded in reducing costs by 90%, through greater use of franchising instead of owning its restaurants.

In the past year, it refranchised more than 500 restaurants.

Burger King saw sales rise 2.4% in Europe, Africa and the Middle East for the three months to 30 September, with a successful online coupon-driven marketing push in Germany and Spain.

In the Asia Pacific region, Burger King sales rose 3.7%.

Related Stories

Burger King launches lower-fat chips
McDonald's: Goodbye to Heinz ketchup
US fast-food workers on strike

Top Stories

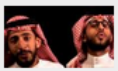

US admits spying constraints needed

Thirteen die as storm crosses Europe
Libyan van robbers snatch $54m
Top Islamists 'killed' in Somalia
Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Thirteen die as storm crosses Europe
2 Woman flown to Grenada, not Granada
3 Trending: No Woman, No Drive
4 US admits spying constraints needed
5 Why is broadband more expensive in the US?
6 Four die as storm hits southern UK

Top Stories


US admits spying constraints needed

Thirteen die as storm crosses Europe
Libyan van robbers snatch $54m
Top Islamists 'killed' in Somalia
Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Thirteen die as storm crosses Europe
2 Woman flown to Grenada, not Granada
3 Trending: No Woman, No Drive
4 US admits spying constraints needed
5 Why is broadband more expensive in the US?
6 Four die as storm hits southern UK

Top Stories

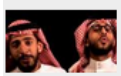

US admits spying constraints needed

Thirteen die as storm crosses Europe
Libyan van robbers snatch $54m
Top Islamists 'killed' in Somalia
Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

Shared | Read | Video/Audio

1 Storm: Tree crushes Amsterdam woman
2 Top Islamists 'killed' in Somalia
3 Why is broadband more expensive in the US?
4 Accused ex-BBC driver found dead
5 Woman flown to Grenada, not Granada
6 US: Spying 'constraints needed'

Top Stories

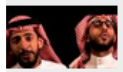

US admits spying constraints needed

Thirteen die as storm crosses Europe
Libyan van robbers snatch $54m
Top Islamists 'killed' in Somalia
Sahara migrants 'die of thirst'

Features


$90 a month
Why do Americans pay so much for home broadband?


Macho mayor
A Russian politician at war with drugs - and ready to break the law


No Woman, No Drive
Bob Marley's hit reworked as Saudi women defy driving ban


Day in pictures
Twenty-four hours of news photographs from around the world

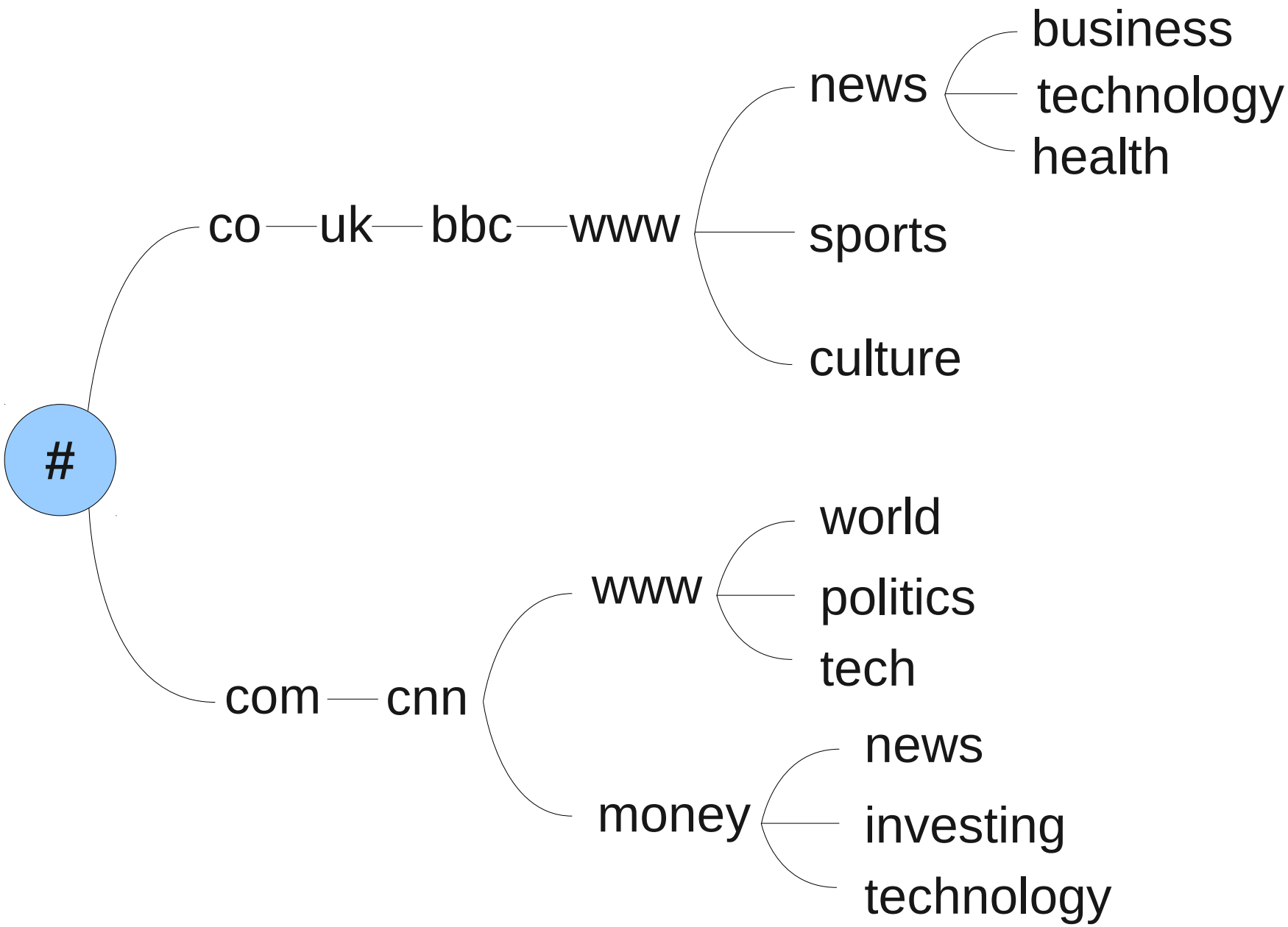
Shared | Read | Video/Audio

1 Storm: Tree crushes Amsterdam woman
2 Top Islamists 'killed' in Somalia
3 Why is broadband more expensive in the US?
4 Accused ex-BBC driver found dead
5 Woman flown to Grenada, not Granada
6 US: Spying 'constraints needed'

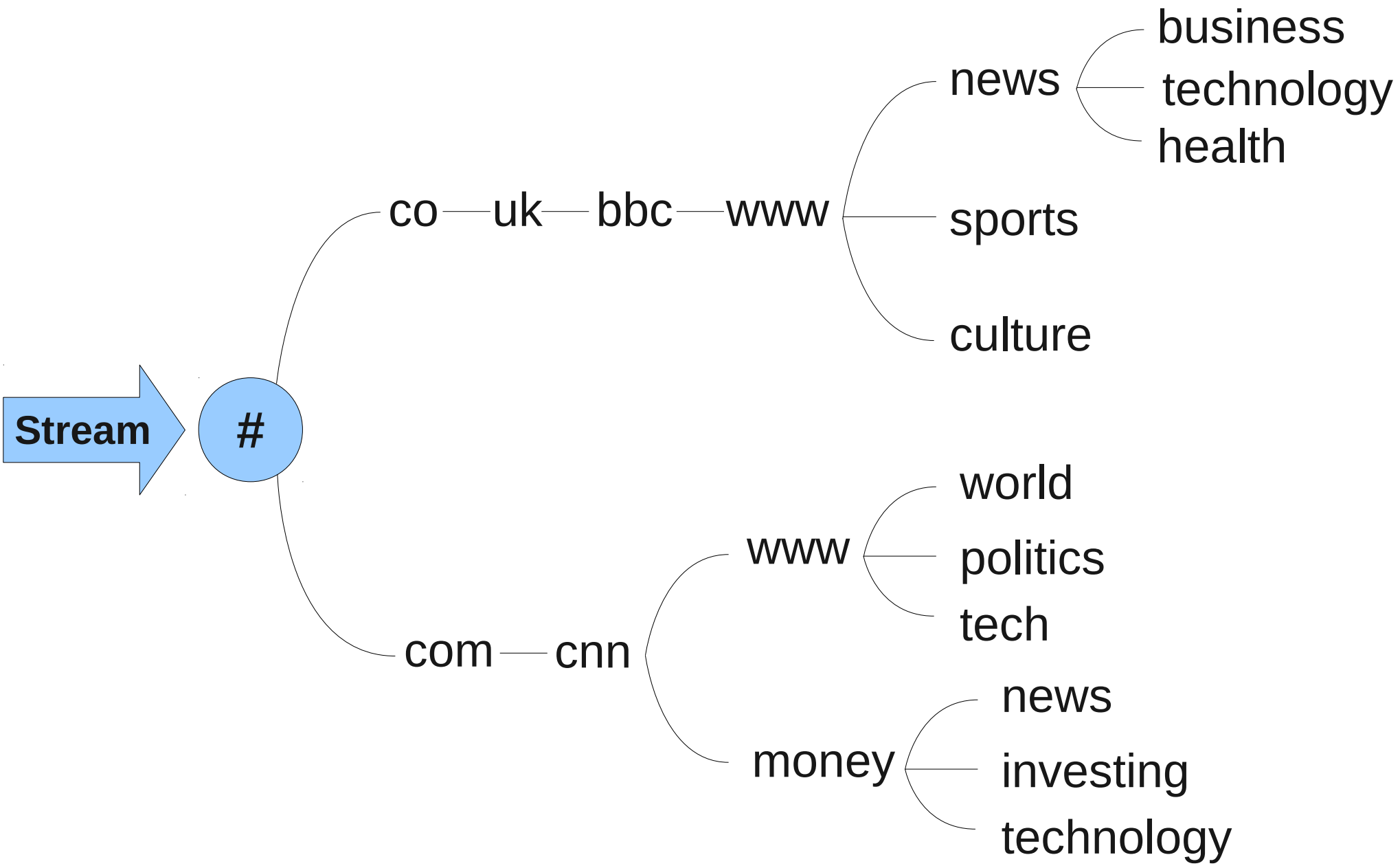
Assumption

Web pages at a similar web address share common boilerplate, whereas main content appears uniquely.*

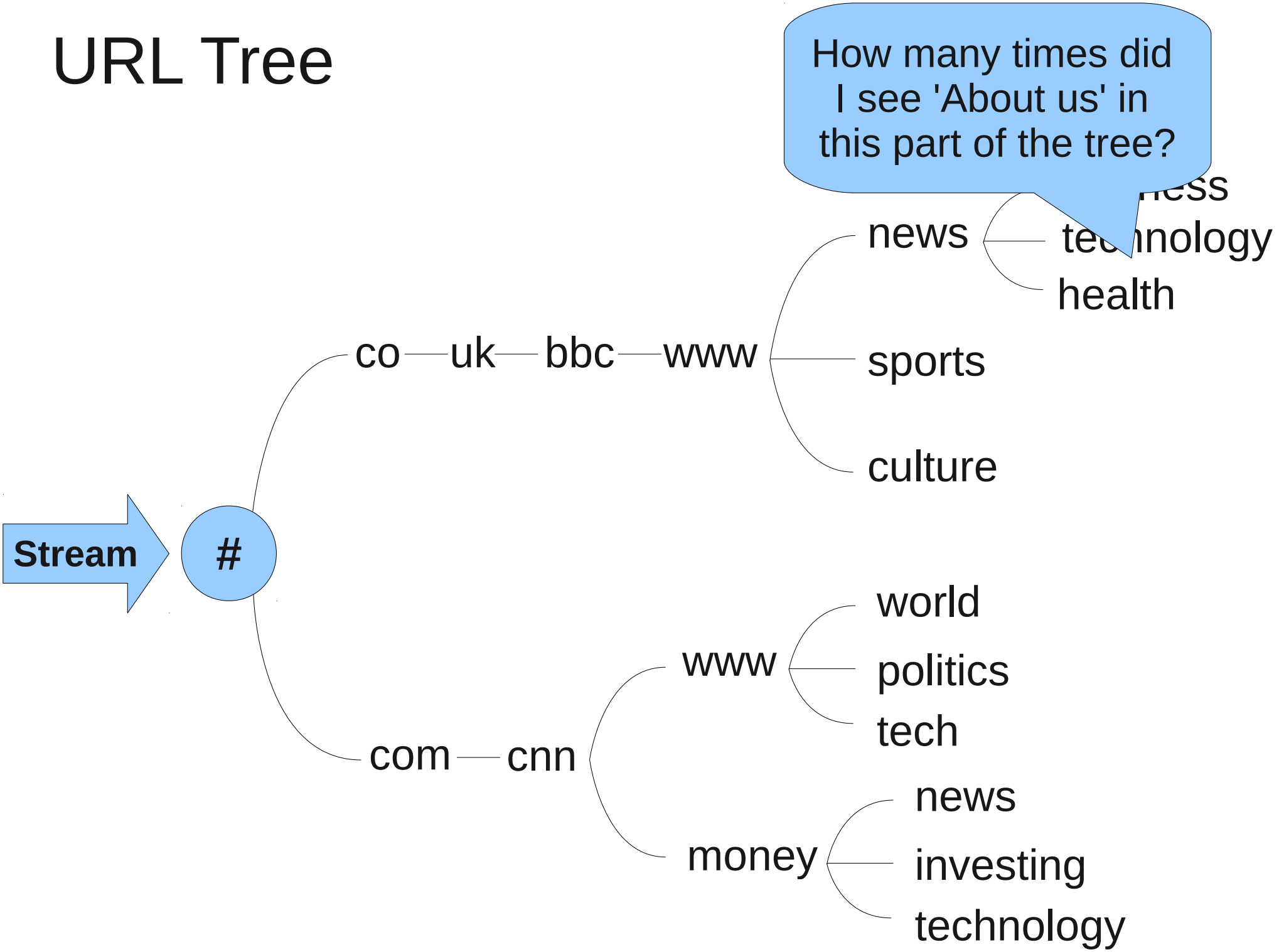
URL Tree


#
co
uk
bbc
www
news
business
technology
health
sports
culture
com
cnn
www
world
politics
tech
money
news
investing
technology

URL Tree


Stream
#
co
uk
bbc
www
news
business
technology
health
sports
culture
com
cnn
www
world
politics
tech
money
news
investing
technology

URL Tree


How many times did I see 'About us' in this part of the tree?
Stream
#
co
uk
bbc
www
news
technology
health
sports
culture
com
cnn
www
world
politics
tech
money
news
investing
technology

URL Normalization for duplicate removal

URL Normalization for duplicate removal

Why duplicates:

- Same documents served over multiple RSS feeds
- News aggregators

URL Normalization for duplicate removal

Why duplicates:

- Same documents served over multiple RSS feeds
- News aggregators

Normalize URLs to ***URL keys***:

1. resolve redirections
2. normalize the encoding and order of query params
3. normalize the query part according to the query normalization rule list, or drop it if no rule applies

URL Normalization example

Request URL :

```
http://news.google.com/news/url?
sa=t&fd=R&usg=AFQjCNHEIIAoeLGPfbkX6IdaQ2xoYptq
-w&url=http://abcnews.go.com/kabc/story?
section%3Dnews/local/los_angeles%26id3D8691010
```

URL Normalization example

Request URL :

```
http://news.google.com/news/url?
sa=t&fd=R&usg=AFQjCNHEIIAoeLGPfbkX6IdaQ2xoYptq
-w&url=http://abcnews.go.com/kabc/story?
section%3Dnews/local/los_angeles%26id3D8691010
```

- Step 1:

```
http://abcnews.go.com/kabc/story?
section=news/local/los_angeles&id=8691010
```

URL Normalization example

Request URL :

```
http://news.google.com/news/url?
sa=t&fd=R&usg=AFQjCNHEIIAoeLGPfbkX6IdaQ2xoYptq
-w&url=http://abcnews.go.com/kabc/story?
section%3Dnews/local/los_angeles%26id3D8691010
```

- Step 1:

```
http://abcnews.go.com/kabc/story?
section=news/local/los_angeles&id=8691010
```

- Step 2:

```
http://abcnews.go.com/kabc/story?
id=8691010&section=news/local/los_angeles
```

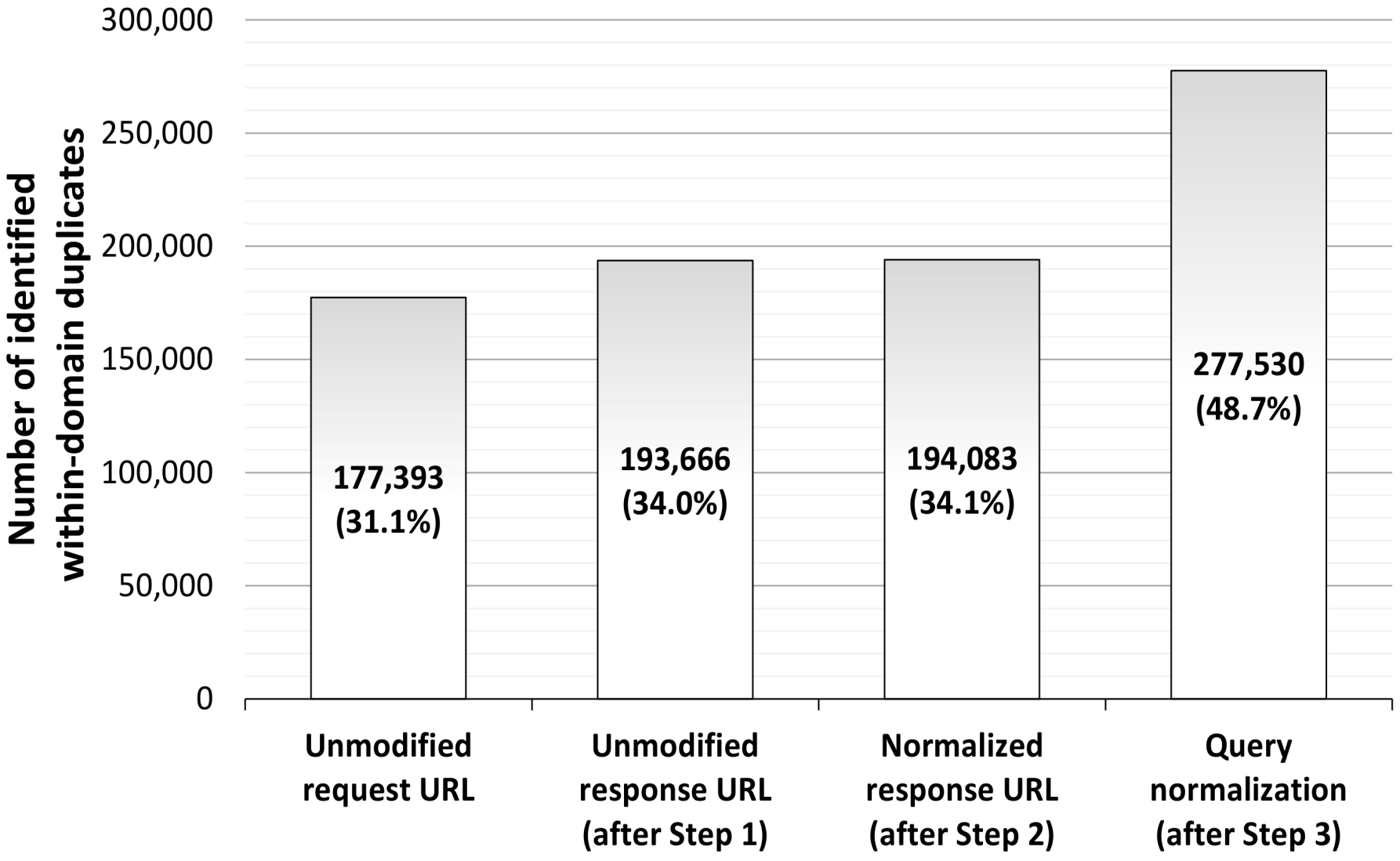
URL Normalization example

Request URL :

`http://news.google.com/news/url?sa=t&fd=R&usg=AFQjCNHEIIAoeLGPfbkX6IdaQ2xoYptq-w&url=http://abcnews.go.com/kabc/story?section%3Dnews/local/los_angeles%26id3D8691010`

- Step 1:
`http://abcnews.go.com/kabc/story?section=news/local/los_angeles&id=8691010`
- Step 2:
`http://abcnews.go.com/kabc/story?id=8691010&section=news/local/los_angeles`
- Step 3:
`http://abcnews.go.com/kabc/story?id=8691010`

URL Normalization Results



Dataset

- 569,583 timestamped documents – the stream

Dataset

- 569,583 timestamped documents – the stream
- From Oct 24 - Dec 19, 2011

Dataset

- 569,583 timestamped documents – the stream
- From Oct 24 - Dec 19, 2011
- 31 Web sites

Dataset

- 569,583 timestamped documents – the stream
- From Oct 24 - Dec 19, 2011
- 31 Web sites
- Sampled from the data acquired during the European project FIRST (http://www.project-first.eu)

Dataset

- 569,583 timestamped documents – the stream
- From Oct 24 - Dec 19, 2011
- 31 Web sites
- Sampled from the data acquired during the European project FIRST (http://www.project-first.eu)
- 292,053 documents after URL normalization

Dataset

- 569,583 timestamped documents – the stream
- From Oct 24 - Dec 19, 2011
- 31 Web sites
- Sampled from the data acquired during the European project FIRST (http://www.project-first.eu)
- 292,053 documents after URL normalization
- **Evaluation set:**
  56,436 documents annotated with manually designed regular expressions tailored for specific Web site templates

Content extraction with URL Tree

Content extraction with URL Tree

- Performance measured by cumulative *F*-score:

$$\overline{F_1} = \frac{1}{j} \sum_{i=1}^{j} F_{1i}$$

- Compared to 10 algorithms from 4 open source projects:
  - Boilerpipe (5 variants)
  - jusText
  - NCleaner (2 variants)
  - Readability (2 implementations)

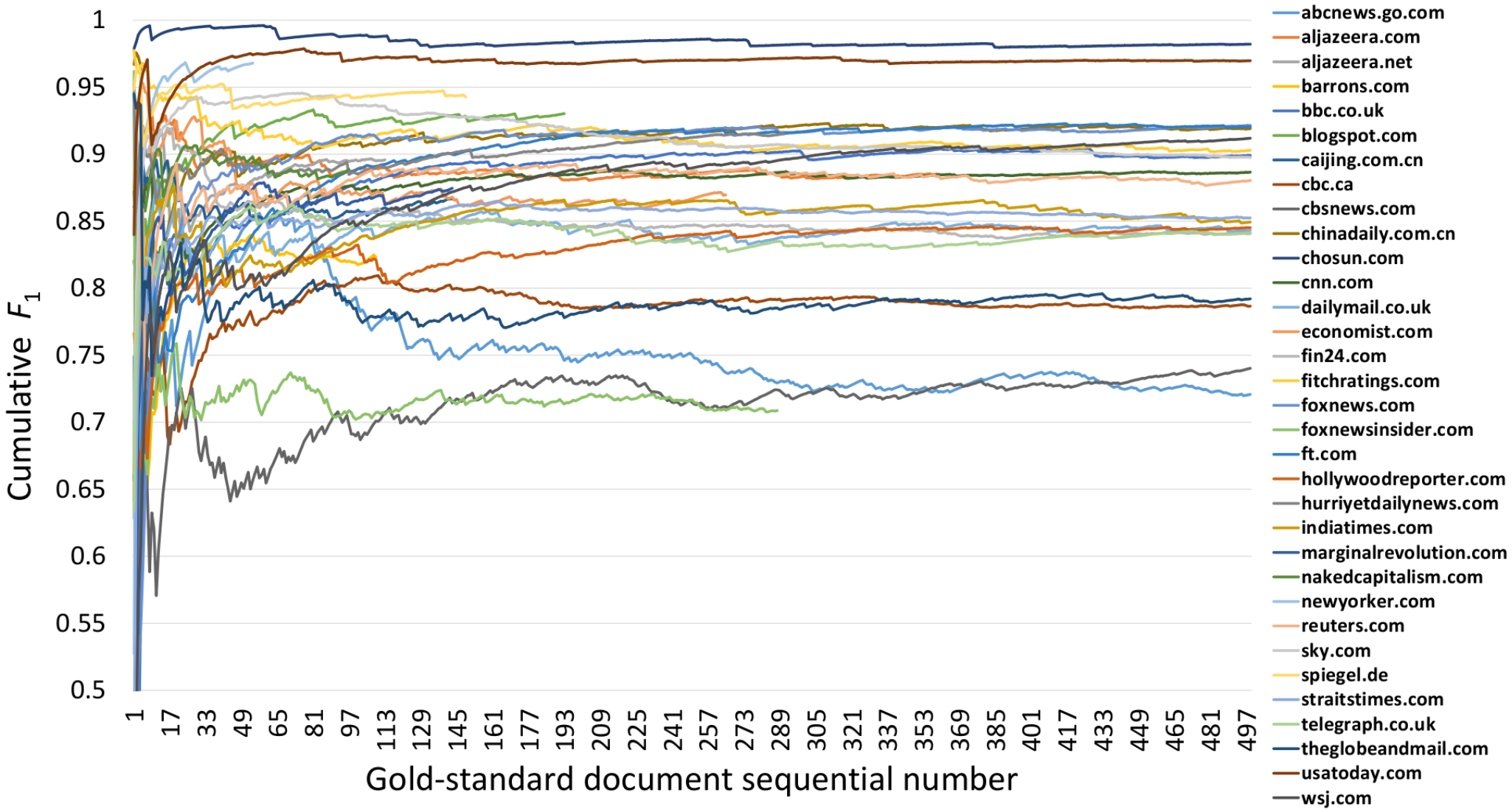
Boilerpipe: Kohlschütter et al. (2010) – Boilerplate detection using shallow text features.

jusText: Pomikalek (2011) – Removing Boilerplate and Duplicate Content from Web Corpora.

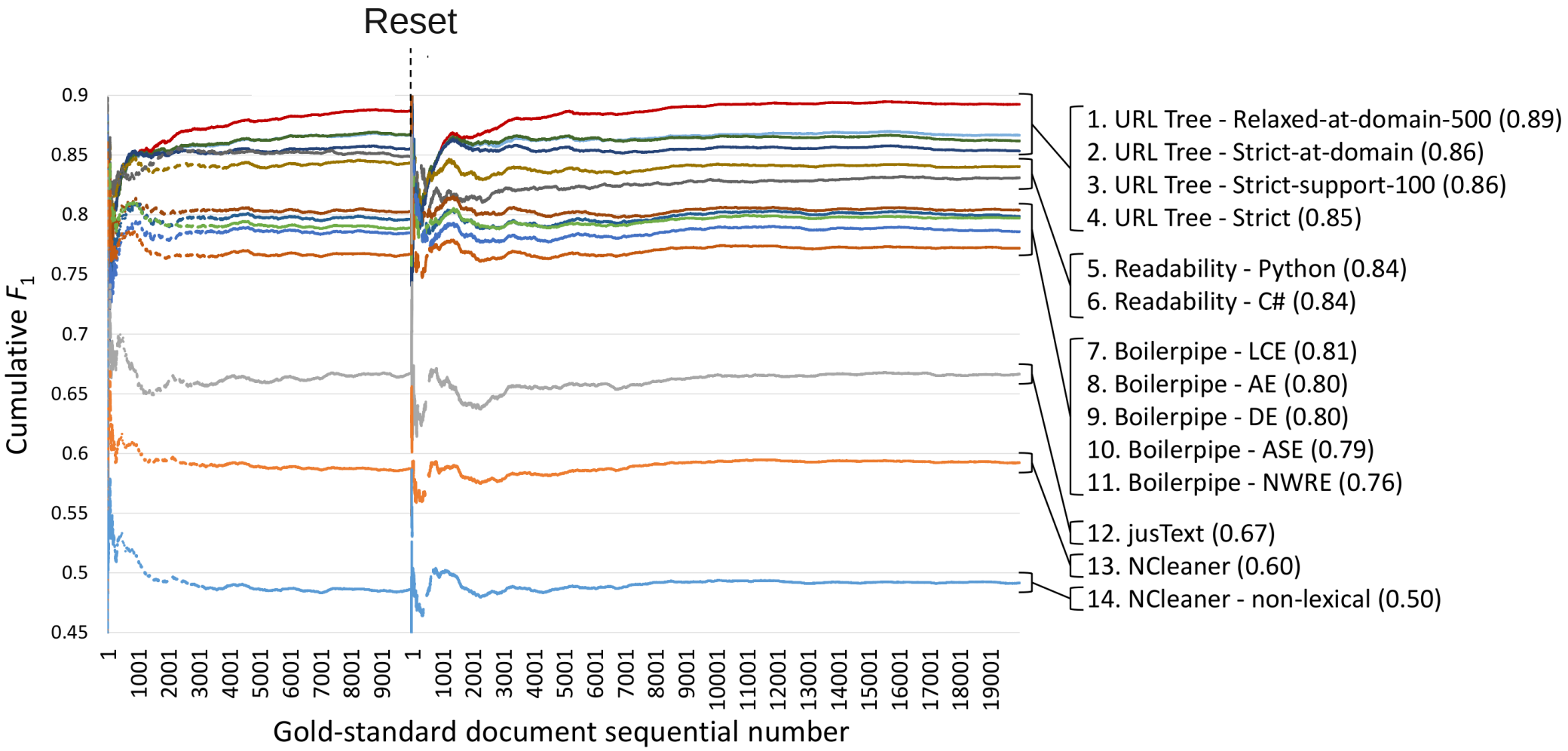
NCleaner: Evert (2008) – A lightweight and efficient tool for cleaning web pages.

Readability: http://www.readability.com, http://code.google.com/p/nreadability,https://github.com/buriy/python-readability

URL Tree results per domain



Aggregated results - comparison



Conclusions

Conclusions

- Proposed a new approah to content extraction from streams of HTML documents, which is:
  - online
  - unsupervised
  - language independent

Conclusions

- Proposed a new approah to content extraction from streams of HTML documents, which is:
  - online
  - unsupervised
  - language independent
- URL Tree outperforms the other evaluated open source algorithms in the RSS data acquisition setting

Conclusions

- Proposed a new approah to content extraction from streams of HTML documents, which is:
  - online
  - unsupervised
  - language independent
- URL Tree outperforms the other evaluated open source algorithms in the RSS data acquisition setting
- URL normalization for detection of duplicates

Conclusions

- Proposed a new approah to content extraction from streams of HTML documents, which is:
  - online
  - unsupervised
  - language independent
- URL Tree outperforms the other evaluated open source algorithms in the RSS data acquisition setting
- URL normalization for detection of duplicates
- Annotated dataset available for download and preview

http://first.ijs.si/urltreedataset

FIRST Content Extraction Dataset

Dataset Info

This dataset is a sample of the dataset that was acquired during the EU project FIRST through RSS feeds of mainly news and blog sites. It is primarily intended for evaluating and training content extraction (boilerplate removal) algorithms. It was semi-automatically annotated by employing regular expressions and performing manual revisions.

Number of documents:	56,436
Number of web sites:	31
Number of domains:	33

Acquisition time:

From	Until
2011-10-23 22:26:14	2011-10-31 23:59:05
2011-11-10 00:00:55	2011-11-30 23:58:18
2011-12-09 23:59:58	2011-12-19 19:22:09

TextBlock Annotations*:

- Headline
- Supplement
- Fulltext
- Comment

* Unannotated *TextBlocks* are considered to be boilerplate. Note that *TextBlocks* annotated with *Supplement* and *Comment* can be also considered as boilerplate.

Dataset Download

The "FIRST Content Extraction Dataset" is available for download as a .zip archive (~350 MB).

Download the dataset.

Cite

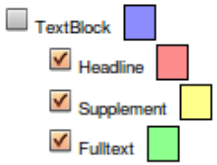
If you use this dataset, please include the following reference:

B. Sluban and M. Grčar (2013) - URL Tree: Efficient Unsupervised Content Extraction from Streams of Web Documents. http://dx.doi.org/10.1145/2505515.2505654

View Annotated Document

Domain name	Documents	Site type
abcnews.go.com	4,878	news site
aljazeera.com	360	news site
aljazeera.net	113	news site
barrons.com	109	news site
bbc.co.uk	6,687	news site
blogspot.com	193	blog
caijing.com.cn	143	news site
cbc.ca	4,231	news site
cbsnews.com	3,493	news site
chinadaily.com.cn	1,545	news site
chosun.com	597	news site
cnn.com	2,108	news site
dailymail.co.uk	2,306	news site
economist.com	265	news site
fin24.com	569	news aggregator
fitchratings.com	789	analysts' reports
foxnews.com	1,631	news site
foxnewsinsider.com	288	blog
ft.com	561	news site
hollywoodreporter.com	1,896	news site
hurriyetdailynews.com	291	news site
indiatimes.com	1,455	news site
marginalrevolution.com	143	blog
nakedcapitalism.com	97	blog
newyorker.com	54	news site
reuters.com	9,908	news site
sky.com	873	news site
spiegel.de	149	news site
straitstimes.com	1,558	news site
telegraph.co.uk	2,860	news site
theglobeandmail.com	791	news site
usatoday.com	981	news site
wsj.com	4,514	news site

Annotation tree



Features

guid = e0837a12fe08c178a2adbfc653aad49d

pubDate = 2011-10-24 00:09:13 +02:00

time = 2011-10-24 00:16:58 +02:00

link = http://rss.cnn.com/~r/rss/edition_americas
/~3/dPZA2yWaiBw/index.html

responseUrl = http://edition.cnn.com/2011/10/23/world
/americas/tropical-weather/index.html?eref=edition_americas&
utm_source=feedburner&utm_medium=feed&
utm_campaign=Feed:+rss/edition_americas+(RSS:+Americas)

domainName = cnn.com

Tropical depression forms near Honduras, Nicaragua

Tropical depression forms near Honduras, Nicaragua - CNN.com
EDITION: INTERNATIONAL
U.S.
MÉXICO
ARABIC
Set edition preference
Sign up
Log in
Home
Video
World
U.S.
Africa
Asia
Europe
Latin America
Middle East
Business
World Sport
Entertainment
Tech
Travel
iReport
Share this on:
Facebook Twitter Digg delicious reddit MySpace StumbleUpon LinkedIn Viadeo
Tropical depression forms near Honduras, Nicaragua
By the CNN Wire Staff
October 23, 2011 -- Updated 2209 GMT (0609 HKT)
STORY HIGHLIGHTS
The tropical depression is the 18th of the hurricane season
It is forecast to become Tropical Storm Rina late Sunday or Monday
A tropical storm watch is issued for portions of the Honduran coast
(CNN) -- A tropical depression formed in the western Caribbean Sea on Sunday, and is likely to become Tropical Storm Rina as it skirts the Honduran coastline, the National Hurricane Center said.
The depression -- the 18th of the 2011 Atlantic hurricane season -- was located off the Nicaragua/Honduras border late Sunday afternoon, forecasters said. The government of Honduras issued a tro
eastward to the Nicaraguan border. A tropical storm watch means tropical storm conditions, including sustained winds of at least 39 mph, are possible within 48 hours.
The depression was moving northwest at near 12 mph, and was expected to turn west-northwest Monday.
"On the forecast track, the center of the depression is expected to pass north of the northeastern coast of Honduras during the next couple of days," the center said.
The depression's maximum sustained winds were at 35 mph, just shy of tropical storm strength, forecasters said.
"Gradual strengthening is expected during the next day or two and the depression is forecast to become a tropical storm tonight or Monday," the center said.
The depression is expected to dump 2 to 4 inches of rain over eastern Honduras, with isolated maximum amounts of 7 inches possible in some spots, according to the center.
The forecast track shows the system brushing Belize as a tropical storm and making landfall near Chetumal, Mexico, on Friday. However, such long-range forecasts are subject to change.
Share this on:
Facebook Twitter Digg delicious reddit MySpace StumbleUpon LinkedIn Viadeo
Most Popular
Today's five most popular stories
Powerful earthquake strikes Turkey, killing at least 65
Gadhafi's autopsy reveals he was shot in head
Ugandan woman recalls harrowing tale of captivity
Why computer voices are mostly female
Gabrielle Giffords to undergo 'intensive' therapy in North Carolina
Loading weather data ...
Home | Video | World | U.S. | Africa | Asia | Europe | Latin America | Middle East | Business | World Sport | Entertainment | Tech | Travel | iReport
Tools & Widgets | RSS | Podcasts | Blogs | CNN Mobile | My Profile | E-mail Alerts | CNN Radio | CNN Shop | Site map | CNN Partner Hotels
CNN en ESPAÑOL | CNN Chile | CNN México | العربية | 한국어 | 日本語 | Türkçe

Terms of service | Privacy guidelines | Ad choices
| Advertise with us | License our content | About us | Contact us | Work for us | Help
CNN TV | HLN | Transcripts

Thank You

http://first.ijs.si/urltreedataset