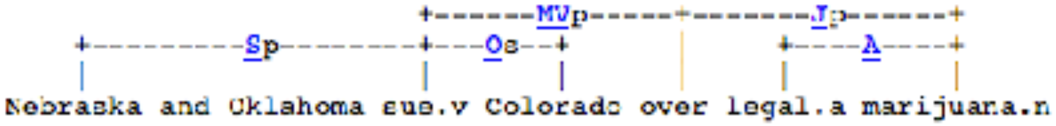
Introduction

The main question to ask is, who is consuming the title, summary and category labels? Is it a human or is it a machine? If the target is a machine, for say an application like retrieval, then extraction would suffice. If the target is a human abstraction would be needed and higher quality would be required. In other words the machine needs to match basic human intelligence related to language and linguistics.

Another question to ask is, should the approach be language independent or in other words, work for multiple languages. Most of the NLP resources are language specific and hence if the target consumer is a human being, one would have to implement the same solution differently for each language. On the other hand if the target consumer is a machine, many language independent statistical techniques can be followed.

Approach to solve the problem of title and summary generation

1. Assume that the story is well formatted. This means, it has good sentence delimiters, names of people and places start with an upper-case alphabet, spurious characters are removed, title is clearly marked etc…
2. break each story into sentences, sentences into phrases, phrases into words
3. select certain types of phrases, mostly noun phrases for title and noun phrases with a verb phrase for candidate summary sentence
4. exclude stop-words and get a probability distribution for words selected from these phrases
5. pick high probability words
6. identify the phrases having these high probability words and select them as a curated set
7. identify sentences containing these curated phrases and identify them as candidate sentences to be included in a a summary.
8. select phrases that are 5 to 10 words long as candidate title phrases to be used in the title
9. select ordered set of sentences as potential summary.
10. avoiding plagiarism. This is a tough problem as it would involve natural language generation. Some of the steps that can be followed are as below:
    - Once we have identified important sentences and phrases based on statistical properties, these sentences can be subjected to link grammar analysis.
    - For example the sentence "**Nebraska and Oklahoma sue Colorado over legal marijuana**" was identified as an important sentence by the statistics of words within the story. This sentence when parsed by link grammar gives



```
                        +------MVp-----+-------Jp------+
    +---------Sp--------+---Os--+      |     +----A----+
    |                   |       |      |     |         |
Nebraska and Oklahoma sue.v Colorado over legal.a marijuana.n
```

- Take the “Sp - Os” link and perform a tense change by reversing the link to “Os - Sp”. In the process change the tense of the associated verb with an additional conjunction to get *Colorado* ***sued by*** *Nebraska and Oklahoma*.
- Such reversals are not always possible. So one approach is to see if such phrases exist in the story and focus on them.

Typical quality improvement operations when preparing the data

- Use of specific stop word list, removal of short words, punctuations
- Use of lemmatizer to normalize words in a distribution to when representing as a feature vector

Typical additional features to extract

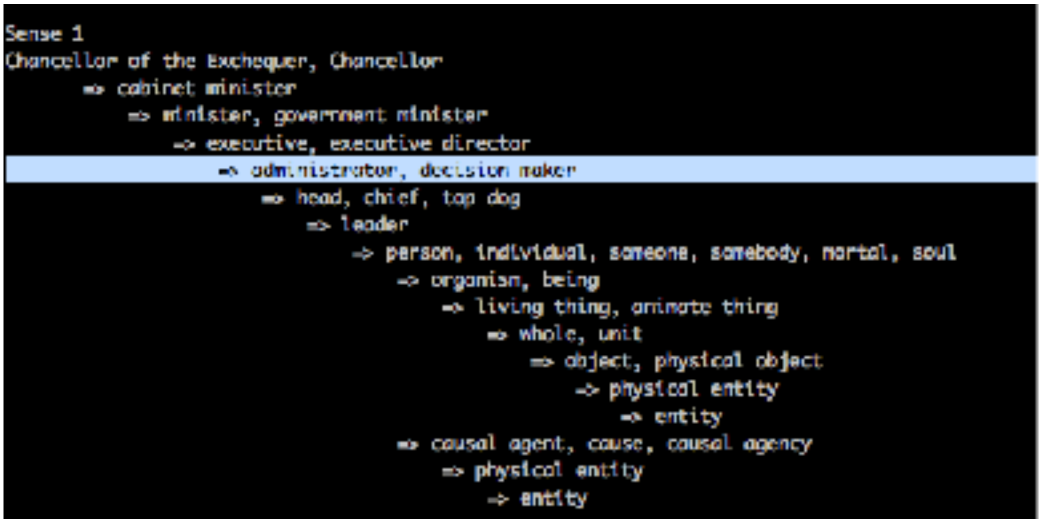
- use a named entity extractor to identify names of people and places separately
- use of word net hypernyms tree to represent words in terms of higher level concepts so that clustering can be on an abstracted concepts instead of the words themselves.
- anaphora resolution to replace words such as “their” with the actual subject

Evaluation and quantification of the quality

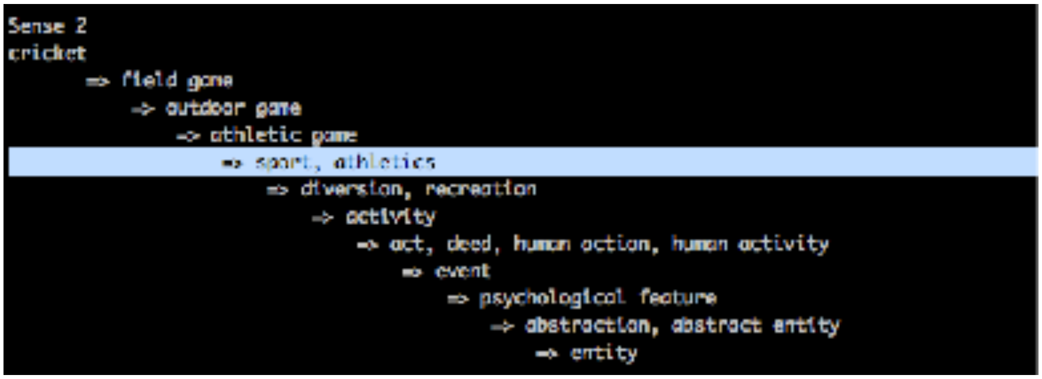
- This is subject to the end application. If we go by absolute human evaluation then, the problem of summarization is never solvable until we have AI implemented in a computer.
- Summarization has always been a subjective issue. A summary good for one person is not good for another. There are many quantitative approaches to evaluate the quality of a summary and they all involve a “gold standard”. Some of the evaluation techniques known to me are BLEU score, DeFUSE score.
- Another approach to evaluation is to constitute a jury of human evaluators and use a majority scoring scheme to arrive at a quality metric.
- Quality is measured in terms of continuity of flow, importance of content in summary compared to the entire story, how often phrases or content repeats in the summary
- Interestingly the DUC challenges have shown that for news story summarization, sophisticated techniques fail to beat a simple technique such as “select the first 30% of the news story as the summary”. This is because, the new editor packs the most important information within the first 30% of the “news real estate” that is available.

Categorization of articles (unsupervised)

- represent each story as a feature vector of phrase occurrence probability or probability of collocations of specific POS tags, such as a Noun-Noun or Noun-Verb word collocations. The biggest challenge is the massive sample space.
- reduce sample space based on frequency filtering
- build a word2vec model using the Noun-Noun and Noun-Verb collocations as the basic elements in a token sequence.
- replace each collocation with its best “synonym” as per the Word2Vec model
- recompute the probabilities now that collocations are replaced by their synonyms.
- Use the new feature vectors and perform clustering : try with SVD, Min-Hash, Hierarchical K-means (here we can create an agglomerative scheme that is optimal at each level of agglomeration). If we use SVD, we can get a list of eigen phrases that can be considered as cluster labels. If we use other methods, we can choose the most frequent phases in a cluster and multiply their frequency with the sum of their TFxIDF weights. So if a phrase has different TFxIDF weights in different documents, take the sum of the TFxIDF. Then also observe the frequency of this phrase in the generated cluster. Multiply the two together. Do the same for all the phrases in the cluster. Rank them by this score and pick a few.
- break each phrase or collocation into individual words and for each word perform a word net hypernyms lookup. get the hypernyms tree and traverse the tree from the root to the leaf. While traversing select the words/phrases from a point in the tree that is at mid-level depth or 2-4 levels from the root. use these as the representative cluster label. For example the word **chancellor** will be replaced with the word **administrator.** Word **cricket** is replaced by **sport, atheletics**



```
Sense 1
Chancellor of the Exchequer, Chancellor
       => cabinet minister
           => minister, government minister
               => executive, executive director
                   => administrator, decision maker
                       => head, chief, top dog
                           => leader
                               => person, individual, someone, somebody, mortal, soul
                                   => organism, being
                                       => living thing, animate thing
                                           => whole, unit
                                               => object, physical object
                                                   => physical entity
                                                       => entity
                                   => causal agent, cause, causal agency
                                       => physical entity
                                           => entity
```



```
Sense 2
cricket
       => field game
           => outdoor game
               => athletic game
                   => sport, athletics
                       => diversion, recreation
                           => activity
                               => act, deed, human action, human activity
                                   => event
                                       => psychological feature
                                           => abstraction, abstract entity
                                               => entity
```

Titles, Summaries generated by extraction using proof of concept code (Categorization not implemented for want of time)

Source code for the results given below
https://github.com/raviguntur/TextAnalytics

Story number:1

short title candidates

----- (The Kurdish offensive against IS forces besieging Mount Sinjar,9)

----- (Kurdish forces in northern Iraq,5)

longer title candidates

--- Masrur Barzani, Chancellor of the Kurdistan Region Security Council, said the operation had been to advance from Zumar - which Kurdish forces recaptured in October - to Mount Sinjar and to rescue the Yazidi people trapped there.

------------- Extract ----------

IS controls a swathe of Iraq and Syria, where it has declared a caliphate.

Meanwhile, the Pentagon's top officer says US air strikes have killed several high-ranking military leaders of IS in Iraq.

"It was a very big operation and thankfully it was concluded very successfully," he said.

Peshmerga commanders said they expected the evacuation of those trapped on the mountain to begin on Friday.

A statement from the Kurdish command said large numbers of militant fighters had fled westwards into Syria or eastwards towards Mosul, which they captured in June.

****

Story number:2

short title candidates

----- (the children, all aged between 18 months and 15 years,",10)

----- (the children, aged between 18 months and 15 years,,9)

longer title candidates

--- "During an examination of the residence police located the bodies of the children, all aged between 18 months and 15 years," said the statement.

------------- Extract ----------

Police said a 34-year-old woman had been taken to hospital with stab wounds but was stable.

Australian Prime Minister Tony Abbott said in a statement it was an "unspeakable crime", and that all parents would feel "gut-wrenching sadness at what has happened".

Bruno Asnicar of Cairns Police said that the scenes his officers had witnessed were

"extremely distressing"
The house in the Manoora suburb has been cordoned off and detectives are searching the yard.
Police have said it was a "tragic event" but there was no cause for public concern.
They have not made any arrests, but said the injured woman was assisting with their investigations.
****

Story number:3
short title candidates
	----- (North Korean leader Kim,4)

 longer title candidates
--- The Mystery of the Sony Pictures Hack
Screenings of Team America, which also mocks North Korea, have been cancelled at other cinemas
The Interview, made by Sony Pictures, features James Franco and Seth Rogen as two journalists who are granted an audience with North Korean leader Kim Jong-un.

 ------------- Extract ----------
Sony hack: White House views attack as security issue
The Interview stars Seth Rogen and James Franco as journalists enlisted to kill Kim Jong-un
Sony scraps The Interview worldwide Watch
A cyber attack on Sony Pictures that forced the cancellation of a major film release is being seen as a serious national security matter, the US says.
Sony withdrew The Interview, a new comedy film about North Korea's leader, after threats from hackers.
Hackers have already released sensitive information stored on Sony computers.
They later issued a warning to members of the public planning to see The Interview.
Referring to the 11 September 2001 terror attacks, they said "the world will be full of fear" if the film was screened.
****

Story number:4
short title candidates
	----- (Crimea targeted Pro-Russian rebels in eastern Ukraine,7)
	----- (a long-term strategy on Russia,5)

longer title candidates
--- EU needs 'long-term' Russia strategy, says Donald Tusk
Donald Tusk, the former Polish prime minister, is the new president of the European Council
The EU needs a long-term strategy on Russia instead of simply reacting to events, the new EU summit chairman Donald Tusk has said.

------------- Extract ----------
Mr Tusk said he had been "really moved" by the leaders' discussion of the Ukraine crisis.
"We need a plan for years we're not too optimistic, we have to be realistic," Mr Tusk said.
"Russia is our strategic problem, not Ukraine," said Mr Tusk, the former Polish Prime Minister.
Russia's behaviour needs a pragmatic solution, a united, common position."
Continuing clashes in the Donetsk region have undermined peace efforts.
****

Story number:5
short title candidates
----- (a review prompted by a White House security breach,9)
----- (the US White House security lapses,6)

longer title candidates
--- Insular US Secret Service needs external boss - report
The White House is meant to be one of the most secure sites in the US
White House security lapses detailed
The US Secret Service, which guards the US president, is too insular and must recruit its next head externally, a review prompted by a White House security breach says.

------------- Extract ----------
The report, a summary of which was released by the Homeland Security Department, said the agency needed more plainclothes and uniformed staff.
The review came after an intruder with a knife entered the building.
Omar Gonzalez, a former US soldier, was apprehended deep inside the presidential residence in September after he had scaled a fence around the building and evaded several guards - including one with an attack dog.
Julia Pearson, the Secret Service director at the time, resigned partly as a result of this breach.
The Secret Service is tasked with guarding the US president, as well as several senior government officials.
****

Story number:6
short title candidates

longer title candidates
--- Israel dismisses Palestinian peace deal plan as 'gimmick'
Palestinians want an end to Jewish settlements in the West Bank and East Jerusalem
Simmering city
Israel says a Palestinian effort to set a three-year deadline for it to end its occupation of Palestinian territories is a "gimmick".

------------- Extract ----------
A draft resolution, submitted by Jordan to the UN Security Council, also calls for a peace accord within a year.
Jordan has indicated it will not seek a quick vote, opening the way for further discussion.
US State Department spokeswoman Jen Psaki said they would not support any action that would prejudge the outcome of negotiations.
"We have seen the draft, it is not something we would support and we think others feel the same and are calling for further consultations," she said.
Mr Lieberman said the draft resolution would only deepen the conflict.
****

Story number:7
short title candidates
----- (the town of Chibok in Borno state in April,9)
----- (the town of Chibok in Borno state,7)

longer title candidates
--- The kidnapping of more than 200 schoolgirls from the town of Chibok in Borno state in April sparked international outrage.

------------- Extract ----------
He said that suspected Boko Haram militants had seized young men, women and children from Gumsuri village.
The attack happened on Sunday but news has only just emerged, after survivors reached the city of Maiduguri.

Meanwhile, Cameroon's army says it has killed 116 Nigerian militants who had attacked one of its bases, AFP reports.
Residents told the BBC that armed militants attacked the border town of Amchide on Wednesday, arriving in two vehicles with many others on foot.
They raided the market area, setting fire to shops and more than 50 houses.
****

Story number:8
short title candidates

longer title candidates
--- In exchange for Mr Gross, who was in poor health, and the unnamed intelligence officer, Washington released three members of the so-called "Cuban Five" who were serving lengthy sentences for espionage.

------------- Extract ----------
Florida Senator Marco Rubio promised on CNN to block the nomination of any US ambassador to Cuba.
Other anti-Castro legislators suggested Congress would remove funding for any normalised ties with the country.
US-Cuban ties have been frozen since the early 1960s - a policy of isolation Mr Obama condemned as a failure.
On Wednesday, the US president said it was time for a new approach.
As part of the deal, US contractor Alan Gross, 65, and an unnamed intelligence officer loyal to the US were released from Cuban prison in return for three Cubans held in the US.
****

Story number:9
short title candidates
----- (a number of conflicts - the assassination of senior insurgents.,10)
----- (The removal of senior Taliban leaders,6)

longer title candidates
--- The 2009 report analyses "high value targeting" in a number of conflicts - the assassination of senior insurgents.

------------- Extract ----------

It said the Taliban's ability to replace lost leaders has hampered the effectiveness of coalition operations against its leadership.

The CIA would not comment on the leaked documents.

As well as examining recent actions in Iraq and Afghanistan, the report assesses British action in Northern Ireland, Sri Lankan operations against Tamil Tigers and French efforts during the Algerian civil war, among others.

Benefits of HVT operations, according to the report, include "eroding insurgent effectiveness, weakening insurgent will and reducing the level of insurgent support".

Potential negatives include "strengthening an armed group's bond with the population" and "radicalising an insurgent group's remaining leaders".

****

Story number:10

short title candidates

----- (freedom of expression and association in South Korea.,8)

----- (South Korea's constitutional court,4)

longer title candidates

--- South Korea court bans 'pro-North' political party

UPP members, led by party leader Lee Jung-hee (centre), demonstrated against the decision on Friday

South Korea profile

South Korea has banned a political party for the first time in decades, with a court ordering a party accused of supporting the North to disband.

------------- Extract ----------

Some UPP members were previously arrested for plotting a rebellion.

Both UPP supporters and its opponents held demonstrations, shouting slogans and waving signs, reports said.

Eight out of nine judges agreed on Friday to accept the government's petition to disband the UPP, order it to forfeit its seats in parliament and ban an equivalent party from forming.

Chief Judge Park Han-chul said "there was an urgent need to remove the threat posed by the party to the basic order of democracy".

But the UPP has said it only wants greater reconciliation with the North.

****

Story number:11
short title candidates

----- (fears over Russian economic crisis What's behind rouble's,9)
----- (Russian economic crisis What's behind rouble's,7)

longer title candidates
--- Russia is on the verge of recession due to falling oil prices and sanctions over its role in the Ukraine crisis.

------------- Extract ----------
Watch
President Vladimir Putin has sought to ease fears over Russia's economy, insisting that the dramatic fall in the rouble will stabilise.
Speaking at his end-of-year news conference, which lasted over three hours, he blamed "outside factors" for the currency hitting an all-time low.
But he admitted Russia's central bank could have acted more swiftly.
However, the president denied pursuing an "aggressive" foreign policy and accused the US and EU of conspiring to weaken Russia.
'Growth inevitable'
Mr Putin accepted Russia had failed to diversify its economy for the past two decades and relied too heavily on its oil and gas exports.
****

Story number:12
short title candidates

----- (Carbon dioxide satellite mission returns first global maps,8)
----- (a quick picture of global carbon dioxide,",7)

longer title candidates
--- Carbon dioxide satellite mission returns first global maps
By Jonathan Amos Science correspondent, BBC News, San Francisco
The map contains 600,000 data points
Euro forest disturbances increasing
Nasa's Orbiting Carbon Observatory (OCO-2) has returned its first global maps of the greenhouse gas CO2.

------------- Extract ----------
This should help scientists better understand how human activities are influencing the climate.
The new maps contain only a few weeks of data in October and November, but demonstrate the promise of the mission.
Clearly evident within the charts is the banding effect that describes how emitted gases are mixed by winds along latitudes rather than across them.
Also apparent are the higher concentrations over South America and southern Africa.
These are likely the result of biomass burning in these regions.
****

Story number:13
short title candidates
----- (the most comprehensive survey of world's deepest place ever,9)
----- (New record depth for deepest fish,6)

longer title candidates
--- New record depth for deepest fish
By Rebecca Morelle Science Correspondent, BBC News, San Francisco
The fish moves in from the bottom-left of the image towards the baited lander
'Supergiant' found in deepest sea
A new record has been set for the world's deepest fish.

------------- Extract ----------
The bizarre-looking creature, which is new to science, was filmed 8,145m beneath the waves, beating the previous depth record by nearly 500m.
Several other new species of fish were also caught on camera, as well as huge crustaceans called supergiants.
The animals were discovered during an international expedition to the Mariana Trench, which lies almost 11km down in the Pacific Ocean.
The Hadal Ecosystem Studies (Hades) team deployed unmanned landers more than 90 times to depths that ranged between 5,000m and 10,600m.
They studied both steep walls of the undersea canyon.
****

Story number:14
short title candidates

longer title candidates
--- Man finds travel companion with same name as ex-girlfriend

18 December 2014
From the section Front Page
A Canadian man has found a woman with the same name as his ex-girlfriend to travel with after booking a three-week holiday before they split up.

------------- Extract ----------
Jordan Axani, 28, broke up with Elizabeth Gallagher in May but couldn't get a refund for the trip.
One was from student Elizabeth Quinn Gallagher from Nova Scotia.
He said he was impressed with her social conscience.
A friend of Elizabeth Gallagher's tweeted a photo of her passport to Jordan Axani in November
"It's totally platonic," Jordan Axani said from New York, where the trip starts on Sunday.
"Do I think we'll become friends?
****

Story number:15
short title candidates
----- (marijuana grown in Colorado coming into their states.,8)
----- (the sale of recreational marijuana Nebraska,6)

longer title candidates
--- Print
Nebraska and Oklahoma sue Colorado over legal marijuana
In 2012, Washington and Colorado became the first two US states to legalise the sale of recreational marijuana
Nebraska and Oklahoma have asked the US Supreme Court to nullify a 2012 law that made marijuana legal in the US state of Colorado.

------------- Extract ----------
The two states allege that Colorado's law is in violation of federal law.
Colorado's attorney general said their suit was without merit.
"Federal law undisputedly prohibits the production and sale of marijuana," said Nebraska attorney general Jon Bruning in a press release.
"Colorado has undermined the United States Constitution, and I hope the US Supreme Court will uphold our constitutional principles."
But he said he would vigorously defend Colorado's law as "it appears the plaintiffs' primary grievance stems from non-enforcement of federal laws regarding marijuana, as opposed to choices made by the voters of Colorado".
****

Story number:16
short title candidates
----- (one of the girls' backpacks, police,6)
----- (the girls' backpacks, police,4)

longer title candidates
--- Following their arrest they told investigators about their belief in paranormal figure Slenderman and their desire to become his "proxies" by killing to demonstrate their loyalty, police said.

------------ Extract ----------
Judge Michael Bohren ruled Morgan Geyser, 12 and Anissa Weier, 13, mentally competent during a hearing.
Doctors retained by the defence previously deemed Ms Geyser not mentally competent.
They allegedly stabbed another girl 19 times in "dedication" to Slenderman, a fictional website character.
The victim, also 12 at the time, was reportedly found by a cyclist on 31 May after crawling from the woods where she was attacked with stab wounds to her arms, legs and torso.
The unidentified victim thanked her well-wishers following the attack
She has recovered and since returned to school.

****

Story number:17
short title candidates
----- (federal police officers and members of the army,8)
----- (members of Mr Mora's vigilante,5)

longer title candidates
--- Mexico troops sent to La Ruana after vigilante shoot-out
Hundreds of federal police officers and members of the army have been deployed to La Ruana
More than 400 federal police officers and soldiers have been sent to a town in Mexico's western Michoacan state.

------------ Extract ----------
Ballistic tests showed all of those killed had fired their weapons in the two-hour gun battle in La Ruana.
Luis Antonio Torres (centre) says his men were fired on by gunmen shooting from rooftops
Hipolito Mora (foreground) says he will not give up his weapons
Mr Mora's 33-year-old son was also among those shot dead.
"We only felt the bullets raining down on us, so we defended ourselves," he said, describing how Mr Mora's supporters shot at them from surrounding rooftops.
"We weren't going to stand by with our arms crossed waiting to be killed."
Mr Mora did not give a description of the shooting but said he would request protection from the police.

****

Story number:18
short title candidates

----- (The peace talks in Cuba - which began in 2012,10)
----- (peace talks with the Colombian government.,6)

longer title candidates
--- The announcement was made in Cuba, where the Farc has been holding peace talks with the Colombian government.

------------ Extract ----------
The leftist rebels said the truce should become a formal armistice and would only end if they were attacked.
The rebels released the general unharmed in November in an effort to revive the talks.
But following Wednesday's announcement by the Farc, President Santos may come under renewed pressure now to match the rebel offer, BBC regional analyst Leonardo Rocha says.
More on This Story
31 OCTOBER 2014, LATIN AMERICA & CARIBBEAN
From other news sites
The battle for guns at university
Most Popular
Colombia Farc rebels declare indefinite unilateral truce
The conflict has left some 220,000 people dead since it began in the 1960s
Colombia general set to be freed
Colombia's Farc rebels have declared a unilateral ceasefire for an indefinite period, starting from Saturday.
****

Story number:19
short title candidates
----- (one of the most chaotic parliamentary sessions in Kenya's,9)
----- (one of the most chaotic parliamentary sessions,7)

longer title candidates
--- TV feed cut
The governing Jubilee Coalition MPs approved the changes despite howls of protest from the opposition in one of the most chaotic parliamentary sessions in Kenya's history, reports the BBC's Emmanuel Igunza from the parliament.

------------ Extract ----------
Opposition MPs shouted and ripped up copies of the bill, warning that Kenya was becoming a "police state".
Four lawmakers were assaulted and another two engaged in a fist-fight.
Parliamentary officials adjourned the debate twice, before the controversial changes were pushed through.
The government says it needs more powers to fight militant Islamists threatening Kenya's security.
The al-Qaeda-linked al-Shabab group has stepped up its military campaign in Kenya,

```
killing 64 people in two attacks in the north-eastern Mandera region since last month.
*******************************
--------------------------------------------
Story number:20
short title candidates
	----- (a search of the suspect's home in Uganda US authorities,10)
	----- (euros, rupees and various African currencies.,6)

 longer title candidates
--- American arrested in Uganda over $2m 'currency fraud'
A pile of counterfeit currency was found during a search of the suspect's home in Uganda
US authorities have arrested an American in Uganda for allegedly leading a large
international counterfeiting ring.

 ------------ Extract ----------
Ryan Gustafson, 27, was charged with conspiracy and counterfeiting outside the US after
the fake currency was used at multiple American businesses.
The suspect faces 25 years in prison.
"We will hold cyber criminals accountable and bring them to justice no matter where they
reside," US Attorney for the Western District of Pennsylvania David Hickton wrote in a
statement.
Fake US currency was discovered at Pittsburgh, Pennsylvania, retail stores and
businesses and traced to a postal box in the area.
Federal agents determined the currency was being sent from Uganda.
*******************************
QWVB:TextProcessing apple$
```