topic_modelling

July 2, 2018

0.0.1 Topic Modelling - and more - with Gensim!

This tutorial will attempt to walk you through the entire process of analysing your text - from pre-processing to creating your topic models and visualising them.

python offers a very rich suite of NLP and CL tools, and we will illustrate these to the best of our capabilities. Let's start by setting up our imports.

We will be needing:

- Gensim
- matplotlib
- spaCy
- pyLDAVis

```
In [1]: import matplotlib.pyplot as plt
        import gensim
        import numpy as np
        import spacy

        from gensim.models import CoherenceModel, LdaModel, LsiModel, HdpModel
        from gensim.models.wrappers import LdaMallet
        from gensim.corpora import Dictionary
        import pyLDAvis.gensim

        import os, re, operator, warnings
        warnings.filterwarnings('ignore')  # Let's not pay heed to them right now
        %matplotlib inline
```

For this tutorial, we will be using the Lee corpus which is a shortened version of the Lee Background Corpus. The shortened version consists of 300 documents selected from the Australian Broadcasting Corporation's news mail service. It consists of texts of headline stories from around the year 2000-2001.

We should keep in mind we can use pretty much any textual data-set and go ahead with what we will be doing.

```
In [25]: # since we're working in python 2.7 in this tutorial, we need to make sure to clean our
         data to make it unicode consistent
         def clean(text):
             return unicode(''.join([i if ord(i) < 128 else ' ' for i in text]))

         test_data_dir = '{}'.format(os.sep).join([gensim.__path__[0], 'test', 'test_data'])
         lee_train_file = test_data_dir + os.sep + 'lee_background.cor'
         text = open(lee_train_file).read()
```

0.0.2 Pre-processing data!

It's been often said in Machine Learning and NLP algorithms - garbage in, garbage out. We can't have state-of-the-art results without data which is aa good. Let's spend this section working on cleaning and understanding our data set. NTLK is usually a popular choice for pre-processing - but is a rather outdated and we will be checking out spaCy, an industry grade text-processing package.

```
In [3]: import spacy
        nlp = spacy.load("en")
```

For safe measure, let's add some stopwords. It's a newspaper corpus, so it is likely we will be coming across variations of 'said' and 'Mister' which will not really add any value to the topic models.

```
In [4]: my_stop_words = [u'say', u'\'s', u'Mr', u'be', u'said', u'says', u'saying']
        for stopword in my_stop_words:
            lexeme = nlp.vocab[stopword]
            lexeme.is_stop = True
```

```
In [5]: doc = nlp(clean(text))
```

Voila! With the `English` pipeline, all the heavy lifting has been done. Let's see what went on under the hood.

```
In [6]: doc
```

```
Out[6]: Hundreds of people have been forced to vacate their homes in the Southern Highlands of
        Indian security forces have shot dead eight suspected militants in a night-long encount
        The national road toll for the Christmas-New Year holiday period stands at 45, eight fe
        Argentina's political and economic crisis has deepened with the resignation of its inte
        Six midwives have been suspended at Wollongong Hospital, south of Sydney, for inappropr
        The Federal Government says it should be safe for Afghani asylum seekers in Australia t
        The United States team of Monica Seles and Jan-Michael Gambill scored a decisive victor
        Hundreds of canoeists are enjoying hard-earned New Years Eve celebrations following fiv
        There has been welcome relief for firefighters in New South Wales overnight with milder
        Some roads are closed because of dangerous conditions caused by bushfire smoke. Motoris
        Work is continuing this morning to restore power supplies to tens of thousands of homes
        Peru has entered two days of official mourning for the more than 220 people killed in a
        President General Pervez Musharraf says Pakistan wants to defuse the brewing crisis wit
        Talks between Afghan and British officials in Kabul have ended without final agreement
        The Israeli army has killed three Palestinian militants who attacked one of its armoure
        Only a protest will can now stop Bumblebee 5 from wining the overall handicap honours i
        South Africa is considering playing left arm spinner Nicky Boje in this week's third Te
        Spain has begun its Hopman Cup campaign in Perth with a 3-0 victory over Argentina. Ara
        New Zealand has rewarded Lord of the Rings director Peter Jackson in its New Years Hono
        The next few hours are crucial for firefighters on alert in the Blue Mountains. Firefig
        Argentine President Adolfo Rodriguez Saa has asked the country's banks to help re-estab
        The nation's road toll has risen to 37, after another death on New South Wales roads. A
        The Federal Government says new national fuel quality standards for petrol and diesel w
        Americans' fears about airplane security continue to increase, after a man made it thro
        A team of police is currently escorting two Swiss tourists back to the safety of a cent
```

The handicap winner of the Melbourne to Hobart yacht race is close to being decided. "S
Pakistan's President Pervez Musharraf says he is ready to meet Indian Prime Minister At
Swedish Round the World ocean racer Assa Abloy has taken line honours in the 57th Sydne
A United States federal magistrate has refused to free on bail a British man accused of
The Palestinian leadership is calling for US peace envoy Anthony Zinni to return to the
The Northern Territory Aids Council says it is not surprised the Territory's rate of HI
There is a one in 20 chance of a dramatic rise in world sea levels over the next centur
An earthquake measuring 4.1 on the Richter scale has shaken parts of Western Australia
New South Wales firefighters are hoping lighter winds will help ease their workload toc
Pakistan's Foreign Ministry has announced retaliatory sanctions against India, saying i
A spokesman for Afghanistan's Defence Ministry claims Osama bin Laden has fled to Pakis
New York Mayor Rudolph Giuliani today bade farewell to the city he has led for the past
Australia's quicks and opening batsmen have put the side in a dominant position going i
A rafter who raised the alarm after most of his party was swept into the Franklin River
A total of 14 yachts have now retired from the Sydney to Hobart Yacht Race, after three
Firefighters across New South Wales are gearing up for a wind change that may bring fur
The man accused to trying to blow up an American Airlines flight on Sunday could not ha
Virgin Blue has begun offering $5 flights from Melbourne to five major cities, in an ag
After a bad start to the holiday period on Australia's roads, there have been no fatali
A Palestinian man has been shot dead as Israeli forces charged into Palestinian-control
Sir Nigel Hawthorne, the British actor best known for his role as the scheming civil se
Seven yachts have been forced to retire from the Sydney to Hobart yacht race, after a s
Australia will be aiming to take early wickets on day two of the second cricket Test ag
Thousands of firefighters remain on the ground across New South Wales this morning, as
European monarchs have reflected on the impact of the September 11 terrorist attacks in
Afghan security forces have arrested a wounded Arab Al Qaeda fighter, but seven others
Russian authorities have sentenced Chechen warlord Salman Raduyev to life in prison for
Skippers are expecting a spectacular start to the 57th Sydney to Hobart yacht race toda
A project working on ways to reduce the debt of prisoners in Australian jails has run o
Police are interviewing a 21-year-old man for stealing a car with a child inside from t
Melbourne's weather is one of the question marks hanging over the second Test between A
An American Airlines flight from Paris to Miami has been diverted to Boston Airport und
Afghanistan's new interim government is to meet for the first time later today after an
Pakistan President Pervez Musharraf believes there is a strong possibility Osama bin La
Australian cricket selectors have made just one change to the squad that beat South Afr
Israel has rejected Palestinian leader Yasser Arafat's bid to make his annual visit to
A Victorian couple is seeking approval to have a baby that has been genetically matched
Japanese officials say their Coast Guard has sunk an unidentified boat after an exchang
Tight security is causing headaches for American travellers this Christmas. The Septemb
A high profile church leader says the Governor-General must clarify his statement defen
The Chief of the Army, Lieutenant-General Peter Cosgrove, has confirmed an Australian m
Argentina's Government has crumbled after at least 20 people were killed and hundreds i
A pay freeze dispute involving Qantas and its maintenance workers will remain unresolve
After months of delays, the company behind plans for a multi-billion dollar Timor Sea g
An Iraqi defector who has applied for residency in Australia claims he has information
The Queensland Premier says he accepts full responsibility for a mentally ill killer be
In the United States, Australian actress Nicole Kidman has been nominated for two Golde
Australian cricket selectors have made just one change to the squad that beat South Afr

The Prime Minister has thrown his full support behind the Governor-General, Dr Peter H
The United Nations Security Council has authorised a multinational force to help keep
US President George W Bush has marked the 100th day of the campaign against terrorism
The death toll in Argentina's food riots has risen to 20. Local media reports say four
The Woomera Detention Centre in outback South Australia has experienced its first quie
The private business sector has to comply with national privacy laws from today which
Dozens of people were injured, some seriously, and others were trapped after a roof co
Zimbabwe has been given five weeks to stop the political violence and invasions of whi
A rare calm in the Palestinian territories has been shattered with the death of a Pale
The owner of a nudist resort in South Australia's Riverland is expecting hundreds of p
The Opposition leader, Simon Crean, says a child abuse scandal in Brisbane has damaged
It has been confirmed two asylum seekers at the Woomera Detention Centre have mutilate
Hamas militants have fought gun battles with Palestinian security forces in the Gaza S
Argentina's Economy Minister Domingo Cavallo is reported to have resigned in the face
The Australian Transport Safety Bureau has called for pilots to be better trained on t
The coroner investigating the death of a race marshal at the 2001 Australian Formula O
After the torching of more than 20 buildings over the past three days, the situation a
Anti-child abuse groups are calling on Australia's Governor-General to resign or expla
The Flanders graveyard of thousands of Australian World War I soldiers in Belgium coul
Federal Labor MP Carmen Lawrence says there is a lot of momentum within the party for
A senior Hamas official has said the radical Palestinian movement has decided to stop
Foreign Minister Alexander Downer says the Commonwealth's democracy watchdog should pu
Legal abortion in Tasmania is one step closer with the lower house of Parliament votin
England batsman Michael Vaughan has become just the seventh player in the history of T
Australian authorities are to be granted access to David Hicks, arrested by the Northe
The Pentagon says the US military is continuing to search caves in the Tora Bora regio
The Civil Aviation Safety Authority (CASA) says it has already warned operators of pis
The Northern Territory's coroner has found that an Aboriginal boy who died in custody
Ansett's administrators are confident of paying out the entitlements of almost 5,000 w
The secretary general of the Law Council, Michael Lavarch, says the Government's propo
The HIH Royal Commission has heard evidence that there were doubts about the company's
Australian cricket captain Steve Waugh has supported fast bowler Brett Lee after criti
Fresh palls of smoke are billowing from the Woomera Detention Centre in South Australi
The Federal Government has called on Labor not to delay its plans to increase the Aust
The Pentagon says the US military is continuing to search caves in the Tora Bora regio
The International Monetary Fund (IMF) has described economic conditions in Australia a
Fire has damaged part of St John the Divine cathedral in New York, one of the world's
The radical Palestinian group Hamas has reportedly shifted the focus in its guerrilla
Joseph Gutnick, the saviour and former president of the Melbourne Football Club, has f
Australian cricket captain Steve Waugh has supported fast bowler Brett Lee after criti
The Immigration Department says overnight fires at the Woomera Detention Centre are pa
The Federal Cabinet has today endorsed a series of anti-terrorism measures at a meetin
Australia is continuing to negotiate with the United States Government in an effort to
Yasser Arafat has accused Israel of escalating violence by killing three Palestinians,
Unions representing Qantas maintenance workers have warned of escalating industrial ac
Australia has beaten South Africa by 246 runs in the first Test at the Adelaide Oval.
Australia is continuing to negotiate with the United States Government in an effort to
Unions representing Qantas maintenance workers have warned of escalating industrial ac

The latest business expectations survey is raising hopes of a solid economic start to
At least four people, including two policemen, have been killed during an attempted co
A new report has revealed there are fewer young people using homeless services than wi
The Federal Opposition wants tougher penalties for ships which spill oil after last we
The United States Space Shuttle Endeavour has touched down at Florida's Kennedy Space
Federal Science Minister Peter McGauran says he is confident security measures at the
US forces backed by their Afghan allies are pursuing hundreds of Al Qaeda militants wh
Qantas has moved to assure travellers there will be no disruption to flights over the
The Governor-General will issue a statement this week to answer allegations about his
The condition of former Indonesian dictator Suharto has improved, a day after the 80-y
The new Solomon Islands Prime Minister has told his people that there are tough times
Australia has picked up two wickets in South Africa's second innings late on day four
The hunt for Osama bin Laden has shifted to the forests around the cave complex of Tor
Israel has reacted with caution to a promise from Palestinian leader Yasser Arafat to
A dispute which could threaten air services returns to the Industrial Relations Commis
A new report suggests the costs of an aging Australian population have been exaggerate
Striking Latrobe Valley power workers will meet today to consider ongoing industrial a
The members of the newly-elected Solomon Islands Parliament meet today to choose a pri
Australia will be looking to score quickly today to set South Africa a challenging vic
Osama bin Laden admitted planning the September 11 terrorist attacks on the United Sta
United States air strikes on Al Qaeda fighters have intensified following the collapse
The Defence Minister, Robert Hill, says the Australian Government is still trying to i
Kashmiri militant groups denied involvement in Thursday's attack on the Indian Parliam
An investigation is underway into what procedures were followed by New South Wales hea
Israeli helicopter gunships and warplanes have swooped again on Palestinian cities aft
The Australian and South African sides for the first cricket Test starting at the Adel
The Federal Government is negotiating with the United States and other countries about
The Israeli Government has declared Palestinian leader Yasser Arafat irrelevant and ha
The Federal Opposition says the unemployment rate has fallen because many job seekers
Industrial action will affect three of Australia's biggest banks over the next two day
Senior Construction Forestry Mining and Energy Union (CFMEU) officials giving evidence
The mind games are continuing as Australia and South Africa have their final hit-outs
At least two helicopters have landed near Tora Bora mountain in eastern Afghanistan, i
Australia and the United Nations have openly clashed in Geneva, over how best to deal
The former managing director of One.Tel has denied claims he mislead the board while t
Industrial action will affect three of Australia's biggest banks over the next two day
A British man has been found guilty by a unanimous verdict of the kidnap and murder of
The AFL's leading goal kicker, Tony Lockett, will nominate for the pre-season draft af
The Pentagon believes it has finally confirmed the whereabouts of Osama bin Laden to a
A French Moroccan man has been charged in the United States with conspiracy in the ter
Australian families of the victims from the Interlaken tragedy have welcomed today's g
The Federal Agriculture Minister, Warren Truss, says he has not been able to win any c
The secret Australian budget for the boat people pacific solution is set at $400 to $5
Japanese car maker Mitsubishi, has confirmed that it has asked for more money from the
Socceroos coach Frank Farina says he could sign a new contract in "three or four days"
The Federal Government says ASIO and the Australian Federal Police have interviewed th
Turning grief into defiance, Americans have paused in remembrance, three months after
Six Swiss tour company staff have been found guilty of manslaughter over the deaths of

A United Nations panel of judges in East Timor has found 10 militia gang members guilty
The United States Federal Reserve has cut a key interest rate by a quarter-point to a 4
Drug education campaigns appear to be paying dividends with new figures showing a 10 pe
The number of adults and children being diagnosed as obese is on the increase. The find
United States peace envoy Anthony Zinni has told a meeting of Israeli and Palestinian s
Milestones in the history of radio will feature on all six ABC radio networks today whi
About 60,000 bank staff will walk off the job this week in what is thought to be the fi
Anti-Taliban fighters say they have captured key areas in the mountainous Tora Bora reg
Israeli helicopters have again attacked Palestinian targets in the Gaza Strip. It is th
A 31-year-old Middle Eastern woman is said to be responding well to treatment after bei
Qantas has unveiled Australia's latest airline today, launching its single-class "leisu
Australia has linked $10 million of aid to a new agreement with Nauru to accept an extr
The Australian Transport Safety Bureau is investigating the overnight crash of a Royal
The Australian cricket team has arrived in Adelaide to prepare for the first Test again
Most of the Tora Bora mountain complex in eastern Afghanistan, where Osama bin Laden's
The Israel Government has expressed regret and promised an investigation into how two o
Unions are already expressing their dissatisfaction with the royal commission set up to
United Nationals secretary-general Kofi Annan has accepted the 2001 Nobel Peace Prize i
Qantas maintenance workers will decide by secret ballot whether to accept the airline's
One person has died after a Royal Flying Doctor Service (RFDS) aircraft crashed near th
New statistics released by the Cancer Council reveal some alarming trends about lung ca
Conservationists have applauded the one-year jail sentence given to a man who logged pr
Ian Thorpe has emulated Kieren Perkins feat by being named Australian swimmer of the ye
The United States is intensifying its bombing of the mountains of eastern Afghanistan i
A new study shows that nearly one third of the Aboriginal and Torres Strait Islander po
The Royal Commission into collapsed insurance giant HIH has been told directors, includ
Unions representing Qantas maintenance workers have not ruled out disruptions to Christ
Olympic 400 metres champion Cathy Freeman will return to competition at the Melbourne T
The Middle East peace process is under new pressure after an ultimatum from the United
Authorities are trying to track down the crew of a vessel that landed undetected at Coc
The Royal Commission into the Building Industry has ended the first day of public heari
The United States says a video tape found inside Afghanistan proves beyond doubt Osama
Prime Minister Ariel Sharon said Israel might step up its military operations in the We
An International study has found thousands of Australians are involved in the child sex
Qantas management and unions representing the airline's maintenance workers will meet a
An Iraqi doctor, being held at Sydney's Villawood Detention Centre, claims he was preve
Australian's casinos generated a $3.1 billion income in the 2000-2001 financial year. T
The royal commission into the building industry will hold its first public hearings in
Geoff Huegill has continued his record-breaking ways at the World Cup short course swim
Israeli tanks and troops have launched two incursions in the Gaza Strip near the Palest
The Federal Government has confirmed there is a blowout in the Defence budget because o
The Australian Government is continuing to talk to Indian authorities about a man who h
A gunman has died after he went on a shooting rampage that left another person dead and
New Zealand's ambassador to Brazil, Denise Almao, said she had identified the body of m
A Swiss fireman has told a court how he snapped a photograph of a body being swept alon
The United States offered full and direct approval to Indonesia's 1975 invasion of East
Jason Stoltenberg will become the new coach of world tennis number one Lleyton Hewitt,
A senior Taliban official confirmed the Islamic militia would begin handing over its la

A suspect allegedly involved in planning terrorist attacks on Australia has been detai
Refugee support groups are strongly critical of Federal Government claims that the "Pa
Several people, believed to be as many as 35, have been shot at a northern Indiana fac
The armed wing of the radical Islamic movement, Hamas has threatened to attack offici
Reserve Bank Governor Ian Macfarlane says he is confident Australia will ride through
New laws requiring all packaged food products containing genetically modified (GM) cro
Indonesian troop re-enforcements have started arriving in central Sulawesi as the gove
America's Cup winner Sir Peter Blake, one of the most successful sailors in yachting h
The Federal Government says a man who has claimed to have been planning terrorism atta
The three US soldiers killed by a misguided US bomb in Afghanistan were from a US Army
A tense stand-off is continuing in Gaza City between hundreds of supporters of the Ham
The Immigration Minister, Philip Ruddock, says the so-called Pacific Solution is worki
The Federal Education Minister, Brendan Nelson, says he accepts the need for the Gover
Two Swiss guides who survived the 1999 Interlaken canyoning accident - in which 14 Aus
The US space shuttle Endeavour has blasted off from the Kennedy Space Centre en route
Australian swimmers have won nine of the 12 events contested at the skins swimming eve
Three US troops and five members of the Afghan opposition were killed by a stray US bo
The Foreign Minister, Alexander Downer, has expressed concern about a man who was arre
Israel has demanded the arrest of 36 Palestinian militants and given leader Yasser Ara
Two Swiss guides who survived the 1999 Interlaken canyoning accident - in which 14 Aus
High interest rates on credit cards have prompted a call for an inquiry by the Austral
Centrelink is urging people affected by job cuts at regional pay TV operator Austar an
The Department of Foreign Affairs and Trade (DFAT) has moved to clarify what it says a
Counting is proceeding very slowly in the Solomon Islands national elections, as offi
Australian swimmers have won nine of the 12 events contested at the skins swimming eve
There has been another suicide bomb attack in the Middle East, this time in Jerusalem.
Four Afghan factions have reached agreement on an interim cabinet during talks in Germa
The Reserve Bank has cut official interest rates again, still concerned about the slow
Federal Treasurer Peter Costello has warned continued economic growth in Australia is
The AFL's all-time leading goalkicker, Tony Lockett, will decide within the next week
The Royal Commission into HIH has been adjourned until Monday after interviewing of th
Darwin Aboriginal custodians will become property developers at Palmerston, after the
The Defence Minister, Robert Hill, has announced more Australian SAS troops have arriv
Israel launched massive air raids across the West Bank and Gaza Tuesday, piling pressu
Interest rates and economic growth take centre stage for Australian financial markets
The Labor Party is set to have a wide-ranging review of its structures, with frontbenc
Computer virus experts have warned of a new "goner" computer virus that can erase data
Defendants in the Interlaken canyoning trial in Switzerland are continuing to deny the
The New South Wales State Emergency Service (SES) says it has now received 5,000 calls
A survey of literacy and mathematical skills of 15-year-old Australian school students
Hundreds of fans stood vigil today for the immersion of George Harrison's ashes into t
Australia has escaped with a draw after a dramatic final day of the third Test against
Israeli forces have launched attacks on some of the key Palestinian symbols of autonom
Traveland's wholly-owned travel centres have ceased operating from today, leaving more
Afghan opposition leaders meeting in Germany have reached an agreement after seven day
At the royal commission inquiry into the collapse of insurance giant HIH, it has been
A director of a defunct Swiss company that organised a canyoning trip in 1999 that end
Widespread damage from yesterday's violent storms in New South Wales has forced the Go

```
The Federal Government is under fire from unions over a new departmental report which
Australian fast bowler Brett Lee has been fined $8,250 for yesterday's on-field outburs
Israeli Prime Minister Ariel Sharon has opened an emergency security Cabinet meeting a
Opposition forces claimed to have captured half of Kandahar airport after fierce fighti
The Prime Minister, John Howard, has revealed he will go to Indonesia for a summit meet
Businessmen Solomon Lew and Lindsay Fox have called on the Federal Government to help
A director of a defunct Swiss company that organised a canyoning trip in 1999 that ende
The storm clean-up in Sydney will resume in earnest this morning as fresh crews are bro
The royal commission into the collapse of insurance giant HIH will resume in Sydney thi
Around 1,000 people have braved the cold for a vigil in the hometown of former Beatle
Australian cricket coach John Buchanan says his team will be going into the final day
Defence Minister Robert Hill has confirmed Australian troops arrived in Afghanistan thi
Israeli soldiers have shot dead five Palestinians in two West Bank towns. An Israeli mi
The royal commission looking into the collapse of insurance giant HIH says the possible
Forward indicators of the Australian labour market are failing to improve, with a furth
The Greens have officially won their second Senate spot in Federal Parliament. The Sena
Eight people are to appear in a Swiss court tomorrow charged with the manslaughter of
The administrator of the financially troubled travel chain, Traveland, says he is confi
France is celebrating victory over Australia in the Davis Cup tennis final, after Nicho
Secretary of State Colin Powell says the United States believes Saudi fugitive Osama bi
A new economic report claims Australia's economy is strong enough to break its close li
Malaysian police have arrested a man believed to have smuggled thousands of boat people
A royal commission will begin this morning in Sydney into the collapse of insurance gia
Eight people are to appear in a Swiss court tomorrow charged with the manslaughter of
There is a renewed attempt to move the debate over choosing an Australian head of state
A third case of mad cow disease has been confirmed in Japan. A panel of experts at Japa
Unions and a major electricity producer will take part in government-sponsored talks th
Rival Afghan factions are deadlocked over the shape of a future government. The Norther
George Harrison the guitarist, songwriter and film producer was widely known as the "qu
Virgin Airline's first dawn flight between Launceston and Melbourne got away on time th
A team of Australian and Israeli scientists have conducted what they believe is success
Today is World Aids Day and the latest figures show that 40 million people are living w
The Federal National Party has rejected a possible merger with the Liberals' at this st
A University of Canberra academic's proposal for a republic will be one of five discuss
Australia will take on France in the doubles rubber of the Davis Cup tennis final today
```

It seems like nothing, right? But spaCy's internal data structure has done all the work for us. Let's see how we can create our corpus. You can check out what a gensim corpus looks like here.

```
In [28]: # we add some words to the stop word list
         texts, article, skl_texts = [], [], []
         for w in doc:
             # if it's not a stop word or punctuation mark, add it to our article!
             if w.text != '\n' and not w.is_stop and not w.is_punct and not w.like_num:
                 # we add the lematized version of the word
                 article.append(w.lemma_)
             # if it's a new line, it means we're onto our next document
             if w.text == '\n':
                 skl_texts.append(' '.join(article))
                 texts.append(article)
                 article = []
```

And this is the magic of spaCy - just like that, we've managed to get rid of stopwords, punctuation markers, and added the lemmatized word. There's lot more we can do with spaCy which I would really recommend checking out.

Sometimes topic models make more sense when 'New' and 'York' are treated as 'New_York' - we can do this by creating a bigram model and modifying our corpus accordingly.

```
In [7]: bigram = gensim.models.Phrases(texts)

In [8]: texts = [bigram[line] for line in texts]

In [9]: texts[1][0:10]
```

```
Out[9]: [u'indian',
         u'security_force',
         u'shoot_dead',
         u'suspect',
         u'militant',
         u'night',
         u'long',
         u'encounter',
         u'southern',
         u'kashmir']
```

```
In [10]: dictionary = Dictionary(texts)
         corpus = [dictionary.doc2bow(text) for text in texts]
```

We're now done with a very important part of any text analysis - the data cleaning and setting up of corpus. It must be kept in mind that we created the corpus the way we did because that's how gensim requires it - most algorithms still require one to clean the data set the way we did, by removing stop words and numbers, adding the lemmatized form of the word, and using bigrams.

```
In [11]: corpus[1][0:10]
```

```
Out[11]: [(51, 1),
          (53, 1),
          (95, 1),
          (108, 1),
          (109, 3),
          (110, 2),
          (111, 1),
          (112, 1),
          (113, 4),
          (114, 1)]
```

0.0.3 LSI

LSI stands for Latent Semantic Indeixing - it is a popular information retreival method which works by decomposing the original matrix of words to maintain key topics. Gensim's implementation uses an SVD.

```
In [12]: lsimodel = LsiModel(corpus=corpus, num_topics=10, id2word=dictionary)
```

```
In [13]: lsimodel.show_topics(num_topics=5)  # Showing only the top 5 topics
```

```
Out[13]: [(0,
          u'-0.216*"israeli" + -0.211*"palestinian" + -0.196*"arafat" + -0.181*"force" + -0.14
          (1,
          u'0.321*"palestinian" + 0.306*"israeli" + 0.299*"arafat" + -0.171*"australia" + -0.1
          (2,
          u'0.266*"afghanistan" + 0.243*"force" + 0.191*"al_qaeda" + -0.180*"fire" + 0.176*"bi
          (3,
          u'-0.373*"fire" + -0.271*"area" + -0.198*"sydney" + 0.191*"australia" + -0.175*"fire
          (4,
          u'0.239*"company" + 0.222*"union" + -0.200*"test" + 0.186*"qantas" + 0.153*"australi
```

0.0.4 HDP

HDP, the Hierarchical Dirichlet process is an unsupervised topic model which figures out the number of topics on it's own.

```
In [14]: hdpmodel = HdpModel(corpus=corpus, id2word=dictionary)
```

```
In [15]: hdpmodel.show_topics()
```

```
Out[15]: [(0,
          u'0.005*company + 0.002*powell + 0.002*afghanistan + 0.002*howard + 0.002*austar + 0
          (1,
          u'0.003*airport + 0.002*taliban + 0.002*night + 0.002*opposition + 0.002*kandahar +
          (2,
          u'0.003*report + 0.003*cent + 0.003*company + 0.003*government + 0.003*job + 0.002*p
          (3,
          u'0.004*israeli + 0.004*arafat + 0.003*sharon + 0.002*official + 0.002*palestinian +
          (4,
          u'0.002*damage + 0.002*north + 0.002*launceston + 0.002*hit + 0.002*virgin + 0.002*s
          (5,
          u'0.003*match + 0.003*israeli + 0.002*rafter + 0.002*play + 0.002*ask + 0.002*not +
          (6,
          u'0.003*palestinian + 0.003*group + 0.002*sharon + 0.002*government + 0.002*israeli
          (7,
          u'0.003*krishna + 0.003*ash + 0.003*hare + 0.002*millicent + 0.002*ganges + 0.002*ha
          (8,
          u'0.002*go + 0.002*people + 0.002*australian + 0.002*hiv + 0.002*read + 0.001*predec
          (9,
          u'0.003*harrison + 0.002*george + 0.002*beatle + 0.002*aim + 0.002*16-hour + 0.002*m
          (10,
          u'0.003*director + 0.003*friedli + 0.002*reply + 0.002*company + 0.002*know + 0.002*
          (11,
          u'0.003*arrest + 0.003*australian + 0.002*health + 0.002*people + 0.002*afp + 0.002*
          (12,
          u'0.004*storm + 0.003*tree + 0.002*work + 0.002*home + 0.002*ses + 0.002*sydney + 0
          (13,
```

```
 u'0.003*strong + 0.003*economy + 0.002*australia + 0.002*follow + 0.002*banksa + 0.0
(14,
 u'0.003*cow + 0.002*case + 0.002*disease + 0.002*japan + 0.002*australian + 0.001*u
(15,
 u'0.002*president + 0.002*rabbani + 0.002*bonn + 0.002*interim + 0.002*security_for
(16,
 u'0.002*commission + 0.002*find + 0.002*include + 0.002*day + 0.002*newcastle + 0.0
(17,
 u'0.002*quiet + 0.001*future + 0.001*robertson + 0.001*appoint + 0.001*juvenile + 0
(18,
 u'0.002*canyon + 0.002*adventure_world + 0.002*people + 0.002*interlaken + 0.002*gu
(19,
 u'0.003*arrest + 0.002*source + 0.002*palestinian + 0.002*israeli + 0.002*soldier +
```

0.0.5 LDA

LDA, or Latent Dirichlet Allocation is arguably the most famous topic modelling algorithm out there. Out here we create a simple topic model with 10 topics.

```
In [16]: ldamodel = LdaModel(corpus=corpus, num_topics=10, id2word=dictionary)
```

```
In [17]: ldamodel.show_topics()
```

```
Out[17]: [(0,
  u'0.006*"australia" + 0.005*"australian" + 0.004*"tell" + 0.004*"government" + 0.004
 (1,
  u'0.006*"people" + 0.006*"palestinian" + 0.006*"israeli" + 0.006*"australian" + 0.0
 (2,
  u'0.006*"year" + 0.006*"australian" + 0.006*"power" + 0.005*"government" + 0.005*"n
 (3,
  u'0.006*"force" + 0.004*"day" + 0.004*"sydney" + 0.004*"year" + 0.004*"people" + 0.
 (4,
  u'0.007*"fire" + 0.005*"people" + 0.005*"area" + 0.004*"force" + 0.004*"new" + 0.00
 (5,
  u'0.007*"australian" + 0.004*"day" + 0.004*"government" + 0.004*"year" + 0.004*"uni
 (6,
  u'0.006*"group" + 0.006*"palestinian" + 0.005*"hamas" + 0.005*"government" + 0.004*
 (7,
  u'0.007*"australian" + 0.006*"australia" + 0.004*"israeli" + 0.004*"year" + 0.004*"
 (8,
  u'0.007*"australia" + 0.005*"day" + 0.004*"united_states" + 0.004*"centre" + 0.004*
 (9,
  u'0.007*"australia" + 0.006*"people" + 0.005*"arafat" + 0.005*"year" + 0.004*"man" +
```

0.0.6 Topic Modelling with scikit-learn

Let us now use NMF and LDA which is available in sklearn to see how these topics work.

```
In [20]: import sklearn
         from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
         from sklearn.datasets import fetch_20newsgroups
         from sklearn.decomposition import NMF, LatentDirichletAllocation
```

```
In [21]: dataset = fetch_20newsgroups(shuffle=True, random_state=1, remove=('headers', 'footers',
         'quotes'))
         documents = dataset.data

In [22]: documents
```

```
Out[22]: [u"Well i'm not sure about the story nad it did seem biased. What\nI disagree with is
          u"\n\n\n\n\n\n\nYeah, do you expect people to read the FAQ, etc. and actually accept
          u"Although I realize that principle is not one of your strongest\npoints, I would sti
          u'Notwithstanding all the legitimate fuss about this proposal, how much\nof a change
          u"Well, I will have to change the scoring on my playoff pool.  Unfortunately\nI don't
          u" \n \nI read somewhere, I think in Morton Smith's _Jesus the Magician_, that\nold J
          u'\nOk.  I have a record that shows a IIsi with and without a 64KB cache.\nIt\'s smal
          u"\n\n\nSounds like wishful guessing.\n\n\n\n\n'So-called' ? What do you mean ? How w
          u" Nobody is saying that you shouldn't be allowed to use msg.  Just\ndon't force it o
          u"\n  I was wondering if anyone can shed any light on just how it is that these\nelec
          u'Archive-name: graphics/resources-list/part1\nLast-modified: 1993/04/17\n\n\nComput
```

u"Hi,\n\nI'm interested in writing a program to generate a SIRD picture, you know\nt
u'.Chris\n',
u"*Reminder*   Plan now for the Andrew Conference.\n*Date* The dates are as noted be
u'does anybody have any info on this monitor or the manufacturers?\n\nall help throu
u'!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!-*-!\n
u"AllMartin MccormickWhat's Exactly in a Flour\n\nMM>From: martin@datacomm.ucc.okstat
u"Here's a question that may be simple enough to answer, but has stumped\nmyself and
u'I\'ve been reading, with much confusion, about whether or not to use\nATManager. La
u' ',
u'\n\tahh, yes, this is a fun topic. No, once the name is incribed on the\ndisk, tha
u'The Bmw speedo is triggered by a reed switch\\magnet assembly in the differential.
u'Sounds liek what the FED has to do is sign a 50 or more year lease to use\ncertain
u"\nFor the purpose of a contest, I'd bet some things could be cut. Like fuel\nfor
u"\nHow about those toneau covers? I've been thinking of building one\nfrom chipboar
u'}Crazy question: "Anyone ever wonder how birds can drop a load on a car\n}going ov
u'\n\nThey exist. Even photosynthetic varieties. Not economical at this\ntime, thou
u"\nMaybe because baseball is the only business where those who are\nresponsible for
u": There are chips which perform the voice compression/expansion. They can't\n: be
u'You think that\'s bad? I\'m in Bowling Green, OH, and we get ABC from\nToledo. We
u'\n\nHmmmm... The prefix "peri-" is Greek, not Latin, so it\'s usually used\nwith t
u"\n\nThat still doesn't mean we should cheer their deaths. Policemen are also in \n
u"a global key G, plus one key U_C for each chip C. The user can choose a\nnew sessio
u'I want to get rid of alot of comics that I have. I am selling for 30% off\nthe Over
u'\n"return_place" is probably incorrect. It should be a pointer, not an\ninteger.
u"] \tAnyone who really believes that the Caps can beat\n\n\nLet's be honest. The Pe
u'\nYou\'re admitting a lot more than that. You are admitting that\nyour morals are
u'\nMedical info without a name/body attached is completely useless for\ntreatment.\n
u"\nIt depends on the bike. Once you've found a bike you're interested in, call\nsome
u'\nCam chain.\n\n -Mike',
u'I have a friend who has a MAC (LC or LC II I think), and her family has an\n"extra'
u"I basically agree, the Tigers are my favorite team. Actually, their\npitching migh
u'I have an Alesis HR-16 drum machine for sale. It includes velocity-sensitive\npads
u"\nOK, as one last attempt, I'll take a different tack.\n\nWe all seem to be in agre
u'We really should try to be as understanding as we can for Brad, because it\nappears
u'I have finally decided to update my SE :-)).\nI am planning on buying a Centris 610
u'The ATF agent interviewed on "Street Stories" reported that the raid was\nill plann
u'\n\n\nAre you your own master? Do you have any habits that you cannot break?\nFor
u" (Neil Williams) writes...\n\n# \n#As long as we're on the subject... Several years
u'\n\n   As a data point from Tennessee, a friend of mine and a police\nofficer e
u'\n\nTry this:\n\nchar *name=NULL;\nunsigned long value;\n\nif(XGetFontProperty(font
u"\nHad an '83 Alliance for a long time. It was a comfortable but sluggish\ncar. I
u'\n\n   None. You need to buy 2 80ns 256k VRAM SIMMs. They cost about $30\neach fro
u'I have one thing to say-- why does everyone say that spliting them up is\nsuch a ba
u'Greetings. I\'ve been seeing the word "storage" mentioned\n\taround oscilliscopes l
u"I've been following the Giants closely over the off-season -- newspapers,\nnotesgro
u': \n: >> Please enlighten me. How is omnipotence contradictory?\n: \n: >By definit
u'\n\n\n\n\n\n',
u'\n\n\n   To display Millions of colors on a 16" monitor you need 2MB of VRAM\nin t

```
u"\n\n    > be the site of major commercial activity. As far as we know it has no\n
u"[... stuff deleted]\n[more stuff deleted...]\n\nHow do you calculate that figure? I
u" \n \n   Hmmmm.  I think, with really large keyspaces like this, you need to\nalter
u'\ncjackson> I am very glad to know that none of you judgemental little shits has\no
u'Two follow up\'s to Mark\'s last posting:\n\n\t1.  As far as current investigations
u'\nCould be the (folk?) song "Clementine".  If memory serves, part of it goes:\n\n
u"\nFirst of all, the chip doesn't do that.  It runs at 16 megabits/second,\nwhich is
u"As I was created in the image of Gaea, therefore I must\nbe the pinnacle of creatio
u"For Sale:  1990 Pontiac Grand Prix SE\n\nWhite, White rims, Gray interior.\n58K mil
u'\nI\'ve been a very intent NREN spectator of the NREN for years.  As a \ncommercial
u'\nThe Fluke 87 beeps at you if you try to take it out of a current measuring\nrange
u"Can someone out there tell me how to switch Window's screen resolution\nquickly and
u"I posted about this a while ago but without code excerpts noone was\nable to help m
u"hi folks\ni have 2 hd first is an seagate 130mb\nthe second a cdc 340mb (with a fut
u"\n\nIt's my understanding that the freezing will start to occur because of the\ngro
u"I was wondering if anyone out in net-land have any opinions on MGs\nin general.  I
u"\nWell, if you want to pick on Morgan, why not attack its ash (wood)\nframe or its
u'\nI can wait \'cos I\'ve already got an accelerated card. It does 1280x1024 but\nno
u"Please satisfy my curiosity. I'm interested in finding out who is using the\ne-mail
u'Whoops!! Wrong group. Soooooooooooooooorry folks..\n',
u'} How does one read the betting spreads for baseball?  They tend to be something\n]
u'    >Now let me get this straight. After a nice, long rant about\n    >how people m
u'Hi all,\ncould someone please tell me if there are drivers for windows 3.1 for\nthe
u"\n\nNo, he's not missing anything.  You're right that some models of the 650 ship i
u'[... a bunch of well-meaning (maybe) cynnical text about screw-thread\n\tsizes, the
u"\n  I think you will find that the active Linux and 386BSD communities are\n  popul
u'Well folks, after some thought the answer struck me flat in the face:\n\n"Why would
u'\n\n   "A handful of anti-gun zealots are telling the public that their\nright to s
u"[..]\nReferring to the manual of my motherboard with AMI-BIOS, 10 beeps are a \n'CM
u"                  ^^^^^^^^^^^^^^^^^^^^^^^^\n    Don't be so sure, the Blues played
u"}Out of what hat did you pull this one?  dB is a ratio not an RBOC!        \n} [..
u'I bought a 386DX33 system a little over 2 years ago, and was satisfied with\neveryt
u'Could be due to the rear-end ratio also.  \n\nUsually automatics have different rea
u'\n\nWell, there is a fair amount of evidence floating around that indicates\nthat (
u'rvenkate@ux4.cso.uiuc.edu (Ravikuma Venkateswar) writes ...\n\nBenchmarks are for m
u'Help!!!!!!!!!!!!!!!!!!! My computer from Gateway is freezing up on me.\nGateway t
u'\nCatcher is their weakest position, with the possible exception of second base.\nT
u" >With that in mind...\n\nI just recently purchased the GCC BLP Elite and I really
u"\nChris Chelios was Montreal's co-captain with Guy Carbonneau when he was traded to
u"Hi everyone. I recently posted about how I received a bad vram chip for my\nnew LCI
u'Hi!\n\nI think VGA-Copy can do what you need. \nIf you create a new floppy for your
u'\nOne rule of thumb is that if a person is making the claim, they are\nwrong.  I wa
u'\nYes it is, as has been evidenced by the previous two stages\nof withdrawal from t
u'\nYes, I do.  A couple of years ago, I did a comparison of the two\nproducts.  Some
u"I am having something very unusual happen.  First \nsome background on my system.
u'\n\n\n:  How about Kirlian imaging ? I believe the FAQ for sci.skeptics (sp?)\n:  I
u'\nThe FDA, I believe.  Rules say no blood or blood products donations\nfrom anyone
u"Could someone explain the difference between Tom Gaskins' two books:\n\n  o  PEXLIF
```

u'Has anyone heard what game ESPN is showing tonight. They said they will\nshow what
u'\nDefinitely, J.R. "Bob" Dobbs, numero uno, top dog, not one can touch, not\none ca
u'\nBefore the S4 became the S4 it was called the 200 turbo quattro 20v.\nThis model
u"\n\nThousands? Tens of thousands? Do some arithmetic, please... Skipjack\nhas 2
u"\n\nNo, I don't watch that Bu**Sh*t.\n\n\nSo, does this mean the cop is at fault fo
u"/* Apologies for those who have read this before but no-one has solved this */\n/*
u'True.\n\nAlso read 2 Peter 3:16\n\nPeter warns that the scriptures are often hard t
u'\nThe following packages meet your criteria in that they are PD and\npresent an aes
u"I'm looking for a version of xterm which handles color and vt220 style status\nline
u'\n\nhenrik] The Armenians in Nagarno-Karabagh are simply DEFENDING their \nhenrik]
u"\nIt's always possible, but if this is the case, I think that there is some\nblatan
u'',
u"Is their a pd/freeware hard drive utility that can handle\na compressed IDE drive w
u'I need help binding some value to the HOME and END keys on my keyboard. I have an\
u"\n\t[a list of large-integer arithmetic packages elided]\n\nI thought I would note
u'Hello All,\n\nI have a PC Transporter for sale. It will work with either an Apple I
u'\n\n\nDid you have anyone in particular in mind there Jody?',
u"\nGee, I never knew Valentine made a comment about how Viola signing\nwith Boston w
u'\n[Excellent discussion of DC-X landing techniques by Henry deleted]\n\n\nThe DC-X
u"I'm about to buy a new car and finance some of it. Since I paid\ncash for the last
u"As the subject says. It has 70k and my brother-in-law wants $250. Please don't\nm
u'\n Lev 17:11: For the life of the flesh is in the blood, and I have given\n it to
u"\nHere in switzerland, the situation is exactly the same. The cable tv companies \
u"\nDuring the regular season, when the intensity is down, not many teams\nhave forwa
u"\nMoreover, if two riders are riding together at the same speed,\none might be ridi
u'Where could I find a description of the JPG file format? Specifically\nI need to k
u'I deleted much of the following article in order to discuss the \nspecific issue of
u'\nI guess your strength isn\'t in math. Clinton hasn\'t been president for\n6 month
u'\n\nHams can legally run up to 1500 watts. It is very unlikely, however,\nthat a k
u'To clarify: \n\nVC++ *is* considered an upgrade for C7. There will be no product\
u'\nAnimal Rights people have been know to do that to other\n"Bike riding dogs.cats a
u"\nTwo shots at it: (1) Check the tires again - if you can see the wear bars,\nyou'r
u': \n: Probably not. But then, I don\'t pack heavy weaponry with intent to use it.\
u"\nOne difference will _probably_ be the same difference as between OS2 and\nWindows
u"\nI would be scared of trying to fit the one piece. When I got my\ntwo piece, I go
u":>Hey, gang, it's not about duck hunting, or about dark alleys,\n:>it's about black
u'\nwrites a "Who woulda thunk it" article which is really the same piece\nevery time
u'\n\nno kidding...just ask the White Sox...\n\ntoo bad, really...\n\n-John Neuharth
u'Here we go again. Is this the same idiot who posted the Gretzky\ntrade to Toronto?'
u'\n[Text deleted, no value judgement implied]\n\n\nMore than shocking. What this sa
u'# 74S\tLater modification of 74 for even higher speed, at some cost in\n# \tpower c
u"\nSorry!! :-)\n\nCall the four points A, B, C and D. Any three of them must be\nnon
u'Davidian-babble:\n\n\nTurkish government on usenet? How long are you going to keep
u'While I cannot answer questions about running XDM over a DECnet, I can\nsay that th
u"I have one round-trip ticket good for travel between USA or Canada and\nEurope, Haw
u'Timeshare week for rent / must use before July / Best offer!!\n\n\nWeek can be "tra
u"I live at sea-level, and am called-upon to travel to high-altitude cities\nquite fr
u'\nReading this definition, I wonder: when should you recognize something\nas being

u'I believe that in order to get at the innards of the new mouse,\nyou must remove th
u'Well put, Jason. I am not from Wisconsin, but I have close relatives who\nlive in
u"Just heard on the news that Mike Keenan formerly of the Blackhawks, Flyers,\nand Ge
u"\nA very good modeling package I found is `irit' (look for irit.tar.Z).\nHowever th
u"Cup holders (driving is an importantant enough undertaking)\nCellular phones and mo
u"THE WHITE HOUSE\n\n   Office of the Press Secretary\n\n__________
u" I don't\n\nWell, no problem! But I get pretty annoyed when they swing at non-stri
u'\n\n   Money orders operate pretty much like checks, with both parties being\ns
u"I've borrowed the 1992-93 version of this book from a friend...holy\nmoley! What a
u'And one of my profs is the chief engineer for the project (Dr. Ron\nHumble, Univ. C
u'\n\n\nHank Greenberg would have to be the most famous, because his Jewish\nfaith ac
u'}Dillon has published a letter in the Blue Press telling people\n}"How to Bankrupt
u':Is erythromycin effective in treating pneumonia?\n:\n:-fm\n\n\nNot only is it effe
u'------------- cut here -----------------\nlimits of AZT\'s efficacy and now suggest

u"I know that the placebo effect is where a patient feels better or \neven gets bette
u'\n\n\nWhat is happening is this:\n1) You turn the TV on.\n2) The horizontal output
u'\n\nFirst of all as far as I know, only male homosexuality is explicitly\nmentioned
u'\n\nThat\'s why zoologists refer to you as a \'fecal shield\'. Colonel Semen \nM. I
u"\n   Are you trying to say that there were no massacres in Deir Yassin\n   or in Sa
u"X Window installation on a Sun4/470 with CG6 alone and with CG2 as\nscreen:0.0 and
u"Hello,\n   I am looking to add voice input capability to a user interface I am\nc
u'\n\nWhy must it be a US Government Space Launch Pad? Directly I mean..\nI know of a
u"You can't. But good luck trying.",
u'\nI think this guy is going to be just a little bit disappointed. Lemieux\ntwo, To
u'Back when I was building round tail light 2002s they were Bimmers. It was\nonly wh
u': Most, if not all, credit card companies offer to double the warranty up\n: to one
u'\n\n\nSource: A. Alper Gazigiray, "Osmanlilardan Gunumuze Kadar Vesikalarla \n
u'\nThose rules/regulations/laws would be subject to the same attack: that\nthey are
u'\n\n\n\nIf I remember rightly PKU syndrome in infants is about 1/1200 ? They lack\n
u"Does anyone know a program that will record keyboard sequences that I \ndo in a win
u"A while back someone had several equations which could be used for changing 3 f\nil
u"\n\n\nHow would you deal with Arabs who ALWAYS threaten to drive you into the sea o
u"\n\nhe ones I have seen are all fluorescent tubes. Maybe you could find a\nsmall tu
u'I bought the diamond stealth 24 a few months ago. it seems to be a\ngreat card esp
u'\n',
u"Deion Sanders hit a home run in his only AB today. Nixon was 1 for 4. Infield\nsi
u'I recently posted an article asking what kind of rates single, male\ndrivers under

u"\nBe sure a dietician is up to date on Crohn's and Ulcerative Colitis. \nPreviousl
u'Let me begin by saying I think this is the world\'s first religion to use\nthe net
u'\n[ANAS] A high rank Israeli officer was killed during a clash whith a Hamas\n[ANAS
u"\n\nTo put it mildly. As I watched the Flyers demolish Toronto last night, 4-0,\nl
u'\nPhilip,\nI think your ideas are well taken and constructive. Thanks for\narticul
u'\n\n\nCome on! Most if not all Arabs are sympathetic to the Palestinian war \naga
u"I would like to sell the following sci-fi books at Best Offer.\nIf you are interest
u'A listmember (D Andrew Killie, I think) wrote, in response to the\nsuggestion that

```
u'I have a 1982 Regal and I am interested in buying\na fiberglass hood, trunk, and bu
u'\n\tDown to 170-some odd lines.  We must be making progress!\nOn an ironic note, wh
u'To recapitulate a bit:\n\n- The essence of marriage is two people\'s commitment to
u"Why use a ground launch pad. It is entirely posible to launch from altitude.\nThis
u'Oddly, enough,  The smithsonian calls the lindbergh years\nthe golden age of flight
u'\nJust to make sure everyone is clear on this: "it never has" refers to \n"protects
u'\nSteve,\n\nIf the Israelis are stupid enough to "allow" a second "Palestinian"\nst
'',
u"\nPlease, leave heaven out of it.  For his own sake, I pray that Dan does\ntake it
u"\nUSUALLY...go enough places and you'll see stuff happen you didn't think did.\n",
u"\n   My understanding is that the 'expected errors' are basically\n   known bugs in
u'\n\nSo?  Kratz was there - does that mean that he\'s a gang member?\n\nEven in the
u'\n\nCircuit Cellar Inc.\n4 Park St.  Suite 20\nVernon, CT  06066\n(203)875-2751',
u"Hello there,\n\nI am looking out for good scanners (gray-scale only, no color) whic
u'\n\nSorry, but it is *virtually* impossible to win a division with "no talent"\nove
u'\nDepends on the nature of the "rounding."  X allows the user to do bit\narithmetic
u'\nHi Noel,\n\nI\'ve made some attempts to write a converter that reads Adobe Type 1
u'\n\nYes, "Clipper" is a trademark of Intergraph.  Its the RISC chip used\nin some c
u'\n\nShouldn\'t have.  But he may need to see the shrink about why he\nwanted to kil
u'Archive-name: space/new_probes\nLast-modified: $Date: 93/04/01 14:39:17 $\n\nUPCOMI
```

u'Could someone please post any info on these systems.\n\nThanks.\nBoB\n-- \n-------
u'\nBecause Greyhound has apparently gotten around to installing their\nRADAR collis
u"Hi,\n\nMy name is rahul and I am doing MS at USU, Logan\nMy query is:\n\tI have a
u"Hello.\n\n\nI just read my first newspaper in a while and noticed an article on a\
u'\n\nBut with cheaper fuel from space based sources it will be cheaper to \nreach m
u'\n\nYou realize, of course, that inevitably some anal retentive moron is\ngoing to
u'Hi, I have a trident TVGA-8900 video card and need the updated\ndrivers for Win3.1
u'\n\n\nWoops! This is rec.sport.hockey! Not rec.sport.golf! Hope you check the\nn
u'SSRT ROLLOUT\n\n   Speech Delivered by Col. Simon P. Worden,\n
u"Jammer !\n\n Dit is geen fantastische advertentie over nep-rolexen\n maar een even
u'Bill Burns was looking for a description of the differnces between the\nCatholic
u'\n\nF A Q !\n\nReference:\nNewsgroups: comp.sys.mac.apps,comp.sys.mac.misc,comp.sy
u"Well, my 14inch VGA 1024x758-interlacing 2.5 year old no brand monitor just\nbit t
u"\nI thought Bill James' latest book completely and totally sucked. I bought\nit,
u'\nThat\'s not all that Kratz doesn\'t know.\n\n\nNow we know that Kratz doesn\'t u
u'As nobody in the food industry has even bothered to address my previous\nquestion
u'[I am posting this for a friend whose news service is "fubared as usual".\n I will
u"\n\n\nIt would also be great for another reason - when not docked, it could serve\
u'YESSS! You make me proud to\nbe an \'Merkun. Good thing you\ndidn\'t get shot th
u'\nTry Parts Express in Dayton, Ohio also. They have a complete line of \nprofessi
u': Has anyone ever heard of a food product called "Space Food Sticks?" \n\nI remembe
u"Hi,\n   I'm currently in the process of writing a number of PD programs\nfor the s
u"\n\nYou might try the recorder and make a micro. I done that to do certain operat
'',
u"\nWelcome to the conservative judiciary.\n\n\nI think Scalia's point was that you
u"Archive-name: space/mnemonics\nLast-modified: $Date: 93/04/01 14:39:14 $\n\nASTRON
u'If I have one thing to say about "No Fault" it would be\n"It isn\'t"',
u"\n: In any case, I think Viola would have made a better signing. Why?\n: Viola is
u"\nThe Fujitsu 2322 uses what is known as an 'SMD' interface (Storage Module\nDevic
u'',
u' ajteel@dendrite.cs.Colorado.EDU (A.J. Teel) writes...\n\n\n\nFine. If you think
u"Sorry I missed you Raymond, I was just out in Dahlgren last month...\n\nI'm the Vi
u"1) Complete 80386Dx25Mhz System for sale\n   SVGA card/w color Tatung VGA Monitor\
u'\nI don\'t meant to defend Eric Molas- I find it somewhat annoying when\nsomeone p
u'\n\n\n\nFoolish me. And here I thought it had something to do with the \nfact tha
u"\nI think this is a little inaccurate, based on Feynman's account of the\nsoftware
u"[...]\n\nI don't claim to be a crypto analyist... there isn't a whole lot of good\
u'\nIt was a test of the first reusable tool.\n\n\nPointy so they can find them or s
u"Off and on over the last several months, threads about RBIs and\nrelated topics ha
u'\n\n\n\nThis statement is just so blatantly disgusting and free of any implicit\nn
u'\nFor a good discussion of cryptographically "good" random number\ngenerators, che
u'\n\nNo, the 6551A is able to operate in a 2 MHz system, the 6551 can only\ntake 1
u'\nHello,\n\n\tI understand this philosophy. The bears are a national\ntreasure, t
u'\n\n\nJust as a matter of interest, a self-promo computer graphics sequence \nthat
u'\nI may be an anarchist nutcase, but I wouldn\'t have frothed overmuch\nhad the go
u"\n\tActually, this started as a great idea. Before steering-column\nlocks became
u'\nI have already called senators, legislators and the Governor demanding\nthat the
u'-- \n73, Tom\n==

```
u"\n\nAw, just take a moment to digest it and I'm sure you'll see the humour...",
u"\nAnother source:  There's a poly blitter for mode y (mode x in 320x200)\nat sunee
u'Is there a 768x1024 Trident driver for windows anywere.  \nThis mode is supported
u'Archive-name: space/math\nLast-modified: $Date: 93/04/01 14:39:12 $\n\nPERFORMING
```

```
u'My wife and I are in the process of selecting a pediatrician for our\nfirst child
u"(Amir Y Rosenblatt) writes\n   > Sam Zbib Writes\n   >>No one in his right mind wou
u'\n\nThe next time you go to church, you can check the better creed, that is,\nhave
u"\nDMorf (Dave's Morph, I think is what it means) and DTax (Dave's \nTGA Assembler)
u"Just as the title suggest, is it okay to do that?\nI havne't got DOS6 yet, but I he
u"The latest driver release is 59 and can be found at ftp.cica.indiana.edu\nin the pu
u'\nI have a new doctor who gave me a prescription today for something called \nSeptr
'',
'',
u"OK, you asked for it!\n\nI guess that doesn't bode well for the Cubs then does it?\
u"\nYes, I did punch in the wrong numbers (working too many late nites).  I\nintended
u"\nI got a male Mallard duck in the chest once.\n\nIt was like being kicked by my ka
u"\nPerhaps instead of this silly argument about what backup lights\nare for, couldn
u"I think you can add former A's first baseman Mike Epstein (no relation) to \nthe li
u"ON the subject of how many competing RC orders there are, let me point out the\nGol
u'Dear friend,\n  The RISC means "reduced instruction set computer". The RISC usually
u'Showing a meaningless (relatively) baseball game over the overtime of\ngame that wa
u'\n: 3DO is still a concept.\n: The software is what sells and what will determine i
u"\n\n\nNot to worry.  The Masons have been demonized and harrassed by almost \nevery
u'It sounds like a MAGNAVOX with a sick flyback on its way out!',
u"What's the deal?  c.s.h. has nothing on it yet.  Is it in OT, is it over,\nwhat?  I
u'   [snip]\n   In the first place the death of three soldiers on a patrol in occupie
u':P\n:P>My favorite reply to the "you are being too literal-minded" complaint is\n:I
u'\n[ These two paragraphs are from two different posts.  In splicing them \n  togetI
u"\n\nIf you only do read/print then there is no reason for the joystick stuff\nnot t
u'================================================================================\n',
u'\n  But, think of the *mystique* you are buying into for that extra $7k or\nmore!!
u"   I was recently thumbing through the 1993 Lemon-Aid New\nCar Guide.  What I found
u"Hi,\n\nI don't know much about Bible. Could you tell me the relations of\nChristian
u'True rumor.  Fact!  A big three way deal!\n\nEric Lindros going to Ottawa Senators
u'I wanted to create a postcript file with Win#.1, to print it on a\nlaserwriter II.
u"\tAre people here stupid or what??? It is a tie breaker, of cause they\nhave to hav
u"\n\nAbout 25 is correct for Numminen and Lumme.\n\n\n\nNo, Kurri's points are too l
u"\n\tI loved the ABC coverage. The production was excellent. The appearance\nwas exc
u'I have a Conner-disk model CP30061G (200Mb ??) with no info at all. The only thing
u'\nThe existence of repeated earth lives and destiny (karma) does not\nmean that eve
u'\nThat\'s the craziest thing I ever heard.   Are you serious?\n\n\t"it doesn\'t tal
u'Two LH Research SM11-1 power supplies (series SM10).\n\n1000W, 5V, 200A (currently
u'Can anyone tell me where to find a MPEG viewer (either DOS or\nWindows).\n\nThanks
u'\nRick Tocchet was captain of the Flyers for several years before he was\ntraded to
u'My news feed is broken and I haven\'t received any new news in 243 hours\n(more tha
```

u"\n\nWell, that's the question, isn't it? The goals are probably not all that\nobvi
u"Well, I was told that my last message came through without anything\nin it, so I'll
u'I will again *repeat* my statement: 1) I *do not* condone these \n*indiscriminate*
u"\nWho compared Quayle to Gore? Mark said he had never heard of any incident\nin wh
u"Does anyone have any Russian Contacts (Space or other) or contacts in the old\nUSSR
u'\n\nNot so. If you are thrown into a cage with a tiger and get mauled, do you\nbla
u'I am looking for a company that can make custom keys. For instance we need\na key
u'\nWhile shopping for a passenger helmet, I noticed that in many cases the\nexternal
u"If anyone's still interested, I have ONE Mattel electronic game left for sale\nor t
u': Consumer Reports once wrote about the S-10 Blazer that it "shook and rattled\n: I
u"\nAbout as good as Mussina's. better than Sutcliffe's and McDonald's\n\nHe's in the
u"\n\n\n\nI'm not sure, but it almost sounds like they can't figure out where the \n
u'\nIf I have a habit that I really want to break, and I am willing to\nmake whatever
u'\nPerhaps you\'re using the wrong brand! (Sorry all HP fans, but I have\na hard tim
u"\nIf anyone is interested in the history of AMORC, I do think Spencer\nLewis publis
u'There was a recent discussion of Dungeons and Dragons and other role\nplaying games
u"\nWell, this particular thread of vituperation slopped its venom over\ninto alt.atl
u'Here is the tollfree hotline for the Epilepsy Foundation\nof America - 1-800-EFA-10
u'\nProbably we would have much the same problems with only a slight shift in\nemphas
u'I am selling a one way ticket from Washington DC to Champaign, IL (the\nhome of th
u'According to a Software engineering professor here, what was actually rated\nlevel
u'note: i am not the original poster, i am just answering because i\nthink this is in
u'\nTo compute this, and many other astronomical things, go and get (x)ephem written\
u'\n\nEven the 68000 can fetch two bytes at a time.\n\nThe new instruction in the 680
u'\nAccording to the (seen several times) postings from Dale Adams of Apple\nComputer
u"I don't know the exact coverage in the states. In Canada it is covered\nby TSN, so
u'\nStop! Hold it! You have a few problems here. Official history says that \nthe fir
u"I was having a look through a couple of components catalogues when I\ncame across a
u'\n\tI think this is a misnomer.\n\n\n\tBut, this just shows then that painful execu
u'sandiego and graig nettles\n\n\n',
u'\nBut the impressive performance of the Graphite was\nnot its Winmark, it was its W
u'{Jason Haines} said\n   "what to do with old 256k"\n   to <All> on 04-15-93 04
u'1. Large padded Cordura bag (maker unknown) nge exterior, black\n\t straps and inte
u'I am trying to get a copy of the _official_ rules of baseball.\nSomeone once sent m
u"Is there a QIC-80 format tape drive that comes\nwith an EISA controller ?\nColorado
u"Hi, Anybody interested in buying my Labtec speaker?\n\n\tLabtec SS-200\n\tAmplified
u'He who overcomes will inherit all this, and I will be his God and he will\nbe my so
u"Can someone tell me in 25 words or less how to compile posix\nmessage catalogs so t
u"I'm writting letters to my Congresscritters and was wondering if\n there is any rea
u'\n\n\tCannot? Try, will not.\n\n---\n\n   "One thing that relates is among Nav
'',
u'\n\nperhaps you can quote just a bit of her argument?',
u'==\nI am posting thi
u"\nYes, but as has been mentioned many times before, the Islanders play at\nthe tale
u'Question for those familiar with Quadra VRAM:\n\nI put 2 256K VRAM SIMMs in my Quad
u'If the heading is true, Mr. Frank should be ashamed of himself.\n\nNothing makes me
u'I am looking for a 8 meg 72-pin SIMM for my Centris 610. Where is the\nbest place
u'\nThe shell is waiting for the window-manager to respond to its\npositioning reques

```
u"I just won an IBM Wheelwriter 6 typewriter in a raffle here on campus. \nSince I ha
u'This is as bad as the "Did You Know"  Japan bashing of 2 weeks ago.  After\nfinding
u"\nIf you specify the rootwindow when you are creating your GC. You may \nuse Xlib t
u"There's\n\nThe sound.exe is actually a self extracting script which includes the .c
u'Hi folks,\n\nCan anyone give me some information, the location of some\ninformatior
u'Hi All!\n\n   I would like to know what are the popular ICs of the type, their capa
u"\nI second the motion.\n\nAll in favor?\n\nBTW>> a few days ago, Charles Fee <CXF1i
u'\nthe question is by going East or West from the misisipi. on either choice\nyou wo
u"\n\n\nKilled by handgun, or killed?  If I'm dead, I don't much care if it\nwas by h
u"Ok, then where is the info for the Licensing kept?  Which file?  In the\norganizati
u"I have a few minor problems with the article posted as proof of \nChrist's resurrec
u'\n\nThis is ok in my opinion as long as the stuff *returns to earth*.\n\n\nIf this
u'\nBut they can make you piss in a jar, and possibly provide DNA, semen,\nand hair s
u"\nFor better worse, the source on this on is Michael Barnsley. His article\nin The
u"\n\n\tAnd the Commonwealth of Virginia has not exactly butted\nagainst the issue or
u"\tThis is a very good point.  AT&T obviously knew and participated in the\ndevelopm
u"What a great day! Got back home last night from some fantastic skiing\nin Colorado
u"I'm looking for a decent Windows news reader. I've given up on winvn 0.76\nsince it
u'\nRef: Encyclopedia of Religion, Mircea Eliade\n\nMAGI: \n\n[Sneak Preview: Later s
u'The SDIO has "contracted" with the NRL (Naval Research Laboratory) to fly the Clemc
u'\nHi Adda,\n\nMost Bible scholars agree that there was one copy of each book at a c
u"\nOr John Edgar Hoover's USA.\n\n\nyet.\n\n\nyet.\n\n\nso far.\n\n\nWhat harm is th
u'The stragegy of the government is interesting.  The real fear comes from\nthem doin
u'\nCandida albicans can cause severe life-threatening infections, usually\nin people
u"\n     ah c'mon, give the guy three days and see what comes up.\n\n     LEO",
u'\nI\'d hardly call that "giving up his chance to be Vice President of the US";\nthe
u"A few points about Mary's being taken into heaven at the end of her life on\nearth
'',
u'--\n--------------------------------------------------------------------------------\n
u'Dear Netters\nI want to send EMG-signals from a running person to a computer. \nEac
u'\n\n\n\n   Hmm ... Turks sure know how to keep track of deaths, but they seem to\nl
u'[insert deletion of unnecessary quote]\n\n\nFirst of all, God does not take any sor
u'Have you tried re-installing the software?   Otherwise I would be dubious about\nsi
u'stuff deleted...',
u"\nYou can change it. As part of a continuously downsizing Government\norganization
u'\n\tUnfortunately, I am one of the "negative-impaired".  The\n\tabove sentence says
u"\n\nWell said, Michael!\n\nThe Catholic traditon has a list of behaviours called th
u"\nNon-smoking, normal law student needs furnished place to live in Memphis\nthis su
u"\nduh, why not just chop out the .sig?\n\nbob vesterman.\n\nps: hey kids, take all
u"I was laughing about the law part.\n\nI've driven thru SOHO...manahattan, _I_ know
u'Do you recall which issue this was in? I posted a message related to this a\nwhile
u"es:\n.C\n\nExcuse me for sticking my nose in, but any parent/parents who do not all
u'Hi! I was wondering if anyone out there could point me to where I can\nget the VESA
u'\n\n\nHey tough guy, freedom necessitates responsibility, and\nno freedom is absolu
u" \n   Not necessarily.  I've been thinking about this, and if this chip/scheme\nis
u"\nLack of build quality was the thing  I notced on the first 2 LH's I\nsaw months l
u'Panasonic KX-T3000H, Combo black cordless & speaker phone all in one.\n new- $160,
u'Accounts of Anti-Armenian Human Right Violations in Azerbaijan #011\n
```

```
u"\nPerhaps it is because witnesses who have left the compound have all\ntestified th
u'I produced an error last week about CHORION:\n\n\nI abhor publishing trash (I abhor
u"\n\nSounds more like an Opel GT.  Neat cars, fun to drive.  Sold through\nBuick fro
u'\n\n[ ... ]\n\n\n\nTalk about adding insult to injury ... \n\nI, for one, believe t
u'/ hpcc01:rec.motorcycles / groverc@gold.gvg.tek.com (Grover Cleveland) /  9:07 am
u"\nSeems he didn't understand anything about realities, liar, lunatic\nor the real t
u'Your posting provoked me into checking my save file for memorable\nposts. The first
u"I'm considering the purchase of a 486DX-33 VLB system to run linux.\nThe system has
u'\n\nOh, and us with the big degrees don\'t got imagination, huh?\n\nThe alleged dic
u'\n\nI recommend the book "Adams _v_ Texas", the story of a man (Adams) who\nwas ser
u"\n  Hello, I know that this has been discussed before.  But at the time\nI didn't r
u"Hi.  According to my always questionable newspaper, there's a silver\ndollar being
u'Being a baseball fan and a fan of the above mentioned band I was\nwondering if anyo
u"I sent a response to the White House at\n\n\t0005895485@MCIMAIL.COM (White House)\n
u'Hi! Everyone,\n\nSince some people quickly solved the problem of determining a sphe
u'                                          ^^^^^^^^^^\n \tnit: Is he unable to type
u"\n  And why are we in Somalia?   When right across the Gulf of Aden are\n  some of
u"\n: I guess this changes my opinion of them and I thought I would warn\n: any prosp
u'\n\t\t\t   ^^^^\n\nJust what are these "scientific facts"?  I have never heard of s
u'\nWetteland comes off the DL on April 23rd, and will be evaluated on the 24th.\nHe
u"\n\nThis brings up an interesting subject that has not been discussed much,\nand pr
u"\n  The commercial uses of a transportation system between already-settled-\nand-ci
u"\nOh, darn.\nOkay, okay, let's stop slamming Ipser, and get on with making fun of o
u'\nCompartment syndrome occurs when swelling happens in a "compartment"\nbounded by
u'Peir-Yuan Yeh asks:\n\nsame >or part of them are the same?  How about Torah? Are th
u"Has anyone found a fix for the following problem?\n\nClient Software:\tSunOs 4.1.1
u'\nThis sounds like another of Ali AbuTaha\'s 57 different "real causes" of\nthe Cha
u"\nAllen, sometimes I think you're OK.  And sometimes you tend to rashly leap into\n
u'The story I related is one of the seven apparitions\napproved by our Church as wort
u"Could someone please send instructions for installing simms and vram to \njmk13@po
u"From: Center for Policy Research <cpr>\nSubject: Ten questions about Israel\n\n\nTe
u"\n\nYes, this is certainly one of the traditional ideas about the Mithraic\ncult (a
u"Does anyone know of bigger raster fonts?  I'm using a Mag 15H monitor with\na Diamo
u"\nStandard colormaps were spec'd with the intention that window managers\nwould mak
u"\nIs there such a thing as the new '94 Eagle Talon? I heard from a freind that\nthe
u"\nWell, actually, the Davidians (Koreshians?) started the fire themselves, \nthe la
u"I have a chance to buy a used PS/2 Appletalk card to create a network with\nmy home
u'\nI gues it is Keesler. The others do not ring the bell but they might be \ninvolve
u"In <1r1om5$c5m@slab.mtholyoke.edu> jbotz@mtholyoke.edu (Jurgen Botz)\n\n\n\n\nThen
u'\n\n\n\n\n\nA lot of us out here in the hinderland will trust the Japanese\nbefore
u'\n\n\n\n\n\n\nBeleive it or not:  NY state once considered eliminating tolls for mo
u'\n   {Description of "External Tank" option for SSF redesign deleted}\n\n\nYo Ken,
u"Pascal Perret, in article <1993Apr21.125750.263@eicn.etna.ch>, wrote\n\nFunny thing
u'\n\n\nI think George is referring to switch.zip in the ~ftp/pub/pc/win3/drivers/vid
u'Is there evidence independent of the FBI that indicates that the Branch\nDavidians
'',
u"Well, there *is* a difference.\n\nI don't happen to have my SAE manual handy, but c
u"SP> From: paulson@tab00.larc.nasa.gov (Sharon Paulson)\nSP> to describe here.  I ha
```

```
u'Hello again folks!\n \nBeen a while since I last sold thangs, but the last time wer
u"\n\nWhile I agree with most of Jon says (I deleted those parts, of course), I \nhav
u'CB>From: behanna@syl.nj.nec.com (Chris BeHanna)\n\nCB>>|>\nCB>>|>  Grf. Dropped my
u"\nThe oclock widget was written using the SHAPE extension.\nYou can do the same in
u'would there be any problems with hooking up a Toshiba 3401 external CD-ROM\ndrive t
u"\nWhat, did you leave the room each of the 100 or so times they said that\nthere WI
u'Does anyone have any experience using XRunner, CAPBAK/X, or PreVueX\nas an automate
u"I have a few things to sell. All items are in great condition. All\nprices include
u"\nThe average amerikan today seems to think that the government should be\nable to
u"Just finished reading Burton Mack's new book, _The Lost Gospel, Q and Christian\nOr
u'\nCould you expand on your definition of knowing? It seems a bit monolithic\nhere,
u"So long as we think that good things are what we *have* to do rather than\nwhat we
u'\n       If you happen to know a political position which does not\nhave people adv
u'\nI Have Version 3.5.1 which I believe was needed for a 040 machine.\nYou should be
u"ites:\n Yeah, and the cop couldn't catch me...",
u'\nRight. So all the cops will be buying antique muscle cars for chase cars;\notherw
u"We will be holding a bake and craft sale at Communiversity in Princeton on  \nNassa
u"\nThe screws are Torx screws and the tool isn't to hard to find. It's a\nmatter of
u'\nI really don\'t understand all this!  I watched on satellite network feeds as  \n
u'I have a 1986 Acura Integra 5 speed with 95,000 miles on it. It is positively\nthe
u'on Date: Sat, 3 Apr 1993 12:38:50 GMT, Paul Dietz <dietz@cs.rochester.edu>\n\n/in c
u"\nThe General Chairman is Paul Bialla, who is some official of General\nDynamics.\n
u"Hello,\n\n\tHas anyone built cxterm (X11R5) on a MIPS platform.  If you have,\nplea
u"\nI saw a mask once that had drawings of band-aids, presumably for every puck\ntha
u"]I'm going to buy a BMW just to cast a vote for Groucho.\n\nI thought you were gonn
u"For Sale:\n\nOS/2 2.0 Extended Services -\n\n        * Extended Database support\n
u'Greetings fellow motorcycle roadracing enthusiasts!\n\n BACKGROUND\n ----------\n\n
u'Can somebody tell me what all the letter spesifications on motorcycle models \nreal
u'\n: >DUCATI3.UUE\n: >QUUNCD Ver. 1.4, by Theodore A. Kaldis.\n: >BEGIN--cut here--C
u'I want to start of list for Syclone and Typhoon owners.  If you are interested\nin
u'Hi:\n\nI am digitizing a NTSC signal and displaying on a PC video monitor.\nIt is l
u'\n  > \n  > I am looking for an algorithm to determine if a given point is bound by
u'Guess the subject says it all.\n  I would like references to any introductory mater
u"Acorn Software, Inc. has 3 tape drives (currently used on a VMS\nsystem) for sale.
u'\nWell, actually now that you mention it, a few weeks ago the CBC ran a\ndocumentar
u"\n\n\n\nSyria had been bombing Israeli settlements from the Golan and sending\nterr
u'] >     I think the dialogue would go better if (at least some) gays\n] >showed awa
u'In  <1qvos8$r78@cl.msu.>, vergolin@euler.lbs.msu.edu (David Vergolini) writes...\n\
u"\n\nFollowing precedent in other areas, the government is likely to put a tax on\ne
u"\nWhat gives the US the right to keep New York?  It is the home of the\nUnited Nati
u"\nHave you tried the library?\nSince you go to WPI (so do I), go to AK and look on
u'Hello All,\n\n\n        Goucher College will soon be retiring a MicroVax II, world
u"\nI haven't seen any speculation about it. But, the Salyut KB (Design Bureau) \nwas
u"\n\nWhat, a dog weighs 150lb maybe, at max? You can't handle it?\n\nYou have, I pre
u'By Dave Luecking Of The Post-Dispatch Staff\n\nAt 9:11 Thursday night, the scoreboa
```

u'\n\n\n\tThe quick answer: Revelation 12:7-9\n\n\t"And there was war in heaven. M
u'\n\n   >But is it any worse than the current unsecure system? It becomes much\n   >
u'For Sale\n\nDining Table (wooden) with 6 chairs $ 125\nDining Table Scandinavian s
u'\n\n\nIt clearly depends on the type of questions you are asking but in many\ncases
u'...\n...\n\nSome other owners on the ford-probe@world.std.com mailing list have co
u'\n\n\nShades of the Edsel! They had pushbuttons in the steering wheel hub\nthat co
u'I am looking for a rat cell line of adrenal gland / cortical cell -type. I \nhave
u'\nUnless I\'ve got my notes mixed up, 939 F.2d 499 comes close to this.\nRegular gu
u' > repeated lives on earth.',
u"This notice will be posted weekly in sci.space, sci.astro, and\nsci.space.shuttle.
u"\nNot if they are unwilling to go through a public marriage ceremony,\nnor if they
u"'>First off, with all these huge software packages and files that\n'>they produce,
u'\n\n\n# I hope I gave you a fairly solid answer to this one: I simply don\'t agree
u"\nJews won't agree with you, Malcolm.\n\nCheers,\nKent",
u"It seems to me that you are the one who is supposed to do some reading. I think\nt
u':>It looks like Dorothy Denning\'s wrong-headed ideas have gotten to the\n:>Adminis
u'Hello everyone. I\'m new to motorcycles so no flames please. I don\'t\nhave my bi
u'\n\n\n',
u'\n[KK] Bugunlerde "jewish jokes" muhabbetlerinden esinlenerek sunu\n[KK] yazayim de
u'\nIt helps to have some idea of the source of the distortion - or at least\na reaso
u"I have a Laserwriter IIg that has disappeared completely from the\nNetwork, i.e. it
u'I want to get a car alarm and I am thinking about getting an Ungo Box.\n   Does an
u'Does anyone have any experience using LCIII with MIDI? Do they get along OK?\nI ha
u'\nDevallano went earlier and more extensively to the Russian strategy\nthan anyone
u"\nIf I'm not mistaken, this is the usual sort of precaution against loss of\ncommun
u'I am sorry to once again bother those of you on this newsgroup. \nIf you have any s
u"[posted for a friend]\n\nOkay, I looked through the FAQs and didn't see this, but I
u'(reference line trimmed)\n\n\n\nWell, I\'d say that a murderer is one who intentio
u"kevinh, on the Tue, 20 Apr 1993 13:23:01 GMT wibbled:\n\n: |> >>Rolls-Royce owned h
u'I am looking for a package that implements standard\nimage processing functions (re
u'The most current orbital elements from the NORAD two-line element sets are\ncarried
u"I edited a few newsgroup from that line (don't like to crosspost THAT\nmuch). I ca
u"\n \tTake a look at ftp.cica.indiana.edu at pub/pc/win3/(util?misc?)\nfor a program
u' So-Called Cool-Hot boxes have been advetised for several years. I recall\nDamark
u'What happened in Waco is not the fault of the BATF. If they would of\nhad the prope
u'\nThis a "tried and true" method utilized by guerilla and terrorists groups:\nto co
u'This is the third and final call for votes for the creation of the\nnewsgroup misc

```
u'Thanks to all those people who recommended Workspace managers for\nWindows 3.1.  I
'',
u'************************************************************\n*
u"I can only comment on the Kings, but the most obvious candidate for pleasant\nsurpr
u"\n: The Bible says there is a God; if that is true then our atheism is\n: mistaken
u'\nJust an comment:  I don\'t like it when people decide what\'s good for me...\nIf
u"\n\n\nARGHHHHHHHHHh\n\nREAD THE MEMOS!!!!\n\nI said that I PERSONALLY had other pe
u'\n\nSo where was she?  And would she consider staying there?\n-- ',
u"From article <C4zLJ8.Bun@queernet.org>, by rogerk@queernet.org (Roger B.A. Klorese)
u"\nAmusing, isn't it?  Seems only the SDCNs realize how much baseball is\na *team* g
u'\n\n\n\n\n\n\n\n\n\nGood I had a bad feeling about this problem because of a specia
u'\nWas THAT your argument.  Well, you didn\'t make it very well.  You started \nfrom
u"Can someone recommend how to ship a motorcycle from San Francisco\nto Seattle?  And
u'My Apple 13" RGB monitor has over the past few months gone brighter and\nbrighter a
u"\n\n\nI hope you realize how trivial it is to manufacture these compounds.  Given\n
u"\n\n\n\n\n\n\n\nI've been told by our local computer guru that you can't do this un
u'PSA 145:9  The LORD is good to all: and his  tender  mercies  are\nover all his wor
u"\nYou should be able to pick up an ADB cable at any computer wiring store...\nI'd g
u'\nI missed the presentations given in the morning session (when Shea gave\nhis "ran
u"The M code stream  might be independently attacked based on knowledge of\nclipper c
u'I have an external hard drive I wish to use as startup disk.\nproblem is, when I sw
u'\n\nWell, these are Armenian and Jewish scholars, not me. Denying the obvious?\n\n\
u"\n\nYup.  We're using both and they work just fine.  Hopefully, someday WFWG\nwill
u'Hi there,\n\nI\'ve made a VGA mode 13h graphics library available via FTP.  I origi
u'Help!\nI am working on a project that involves using text with foreign language\nch
u"Hi,\n\nxtici worked for my system.\nI'm using X11R5 pl 17 clientside only on a DEC
u"\n\nWRONG:  I don't think even PKP claims this one.  It is not unlawful to\ndistrib
u'\n\n\nGood thing i stuck in a couple of question marks up there.\n\nI seem to recal
u'\n\n\n468DX2/66 EISA/VESA. Royal, among others, is celling such a system for $2010
u'\n    The immunization program is just a "useful first step". Among other\n    thin
u': >Is there an ftp site where I can get the MS speaker sound driver?  There\'s\n: >
u"Well, that would depend on how much we wanted the US and how much\nwe wanted the $1
u"The 68070 _does_ exist. It's number was licensed to Philips to make their\nown vari
u"\n\nSorry to disappoint you but the Red Wings earned the victory...easily.\n\nI wat
u'\nStimulation of the vagus nerve slows the heart and drops the blood\npressure.\n\n
u' \n\nMore like Philadelphia making dust out of dead Leafs.\nRoger why do you contin
u'Played in Scandinavium, Gothenburg, April 15 1993:\n ==============================
```

u'THere is a defect in the 13" hi-res monitors, bring it to a dealer and \nthey will
u"\nI would hardly consider the BD's to be Christian. They were acting in\ndirect co
u"\n\tClose Roger, but no banana, er avocado or is it artichoke ?!?\n\n\tGeracie in t
u"Here are some recent observations taken by the Hubble Space Telescope:\n\n   o TH
u'   What does anyone think that Judge Wopner would do if Karadzic was\n   on t
u"Ack! Sorry for the repeat posts: I thought I was posting to the newsgroup\non which
u'# #Slavery makes economic sense (it NEVER makes MORAL sense) when human\n# #muscle
u'\nTo which Mark Ira Kaufman responds:\n\nMark,\nWere you one of the millions of Ame
u"Don't knock Vaughn for being a spring training .400 hitter\nbut a .250 regular seas
u'\n\n\n\nMy argument is mainly a proposal of what I think is a plausible argument\na
u'[most of post deleted]\nThere is an easy way out...\nPost the flyers on the stall o
u"I have this used equipment for sale, everything is negotiable!\n\n1200 Baud Compuad
u'# # "labor" is a tough one. Labor is defined, economically, as the efforts,\n# # bo
u'\n\nCNN just claimed he bought 104 "semi-automatic assault rifles". And\nthey say
u'\nOur panel of judges has deliberated the question, and the answer is:\n\nSend the
u'My brother is in the market for a high-performance video card that supports\nVESA l
u"\n\tNot to pick on Mr. May in particular, of course, but isn't this\nkind of the do
u'\n\n\tCan you say, "I get more background radiation from living in\nDenver or havin
u'Does anybody know if there is a mailing list or newsgroup for\nPower Systems and re
u"\nThat would be neat, but nowhere in the Bible does it say\nthat one who has the gi
u'OK, here\'s something for all of those people who think cops are always\nmore respo
u'   \n\n\nThis is not good advice. A couple of yea
u"\n\n[...stiff deleted...]\n\n\n[...stiff deleted...]\n\nSpeed is a quantifiable mea
u'\nI am not an expert in the cryptography science, but some basic things\nseem evide

```
u' > >(specifically the terrorists and drug dealers who proponents of key escrow\n >
u" \n\nNaw, the owners of WordPerfect are Mormons, and by Tony Rose's\nand Robert Wei
u"Distribution: usa\nReply-To: lihan@ccwf.cc.utexas.edu\n\nWhile I'm on the net buggi
u'My Honda Accord just hit the magic 100,000 mile mark and now\nall sorts of things a
u'\n\n\nSince we are in the subject, I have one more question. I have a Trident\n890
u"In reference to the limits of acceleration with guns launching solid\nrockets as pa
u'\n  No. The christians were leary of having an atheist spokesman\n  (seems so clanc
'',
u"\n\nI know it doesn't make sense, but since when is 'Napoleon' about\nsense, anyway
```

```
u"Greetings!\n\nSteve Summers and the Chief were on 48 Hours last night shmoozing\nsp
u'\n\nSpecifically, which changes are you talking about?  Are you arguing\nthat the r
u'\n\n\n  >[I\'m sort of mystified about how a Christian might respond to this.]\n\n
u"\n\nNo, IMO, Mr. Stowell missed the point.\n\n\nMr. Stowell seems to have jumped ra
u"...\n...\n\nYeah...I've seen you're grand mother...I bet she could.",
u'\n',
u'\n\nAnd the \'Turkish Karabag\' is next. As for \'Cyprus\', In 1974, Turkiye \nstep
u'\n',
u"\n\tI'm definitely going to write my Congressman, and nobody's ever\ngoing to make
u'Hi, everybody:\n    I guess my subject has said it all.  It is getting boring\nlook
u"I used the information provided in the recent resource listings and\ntried to ftp t
u': In article <C4u3x5.Fw7@magpie.linknet.com> manes@magpie.linknet.com (Steve  \n:
u"-- \nHi netland,\n\nI thought that I once read about the existance of a virtual mwr
u'[Note, Ohio legislation unlike Federal legislation, shows the entire\nlaw as it wou
```

```
u'\n\n\nAnd organized religion is a religion built from organized values.\nAnd Ford
u"\n\nOh, you foolish person.  I do know what the fuck I'm talking about\nand will g
u"\n  The same Bill James? Why do you say that? It sounds like you're suggesting \nB
u"\n\nWith yet another tax being floated by the Clinton administration to\npay for n
u'VHS movie for sale.\n\nDance with Wovies\t($12.00)\n\nThe tape is new and just ope
u"\n\n\n >   most of their leaders are stupid, and/or not independent, and/or\n\n\n\
u'I was following an example of the LH the other day, and noticed the fit\nbetween t
u"\nI'm not sure about Juha, but another top center, Rauli Raitanen([ss{t)\nis draft
u'\n[...]\n\n   In the September 1992 issue of THE TUFTS UNIVERSITY DIET AND NUTRITI
u"\nNow there's a good idea ! All you need is 20 amps DC for a few minutes, and\na g
u'for a\nidentifies\nthat\n\nIs this software available either commercially or publi
u"I need definitions of the SPEC and Dhrystone benchmarks. Any background\nmaterial
u'\nThat will make it easy for a car thief.\nSaves him/her the trouble of popping yo
u'\nFor the first Move incident (no bomb, several members killed in\ngunfire, circa
u'\nStep 1) Join the AMA (American Motorcycling Association).  Call 1-800-AMA-JOIN.\
u'\n\nI notice you did not offer an alternative number.  Try this one on for\nsize..
u'\n\nOn my LC (RZ to any ex-colonists) I replaced the bolt at the bottom of the bar
u"There is another useful method based on Least Sqyares Estimation of the sphere equ
u"Boy, hats off to any Cubs fan who can actually muster up the courage to put\ndown
u"\n\n\nI second that suggestion.  Although I don't own the HP Portable Deskjet,\nI
u'',
u"\n\nHow many NuBus slots do you have?\n\nApplied Engineering has something called
u'\n\nNo, there\'s no evidence that would convince any but the most credulous.\n\nTh
u'Is there a FAQ on Cyrix 486DLC? Could anyone please repost it or\nemail to me, if
u'\n\n\tThe word that is missing in this whole discourse is not the "B"\nword, or th
u"While playing around with my Gateway 2000 local-bus machine last\nnight, it became
u'Is there a Wyse 60 Terminal Emulator or a comms toolbox kit available on the\nnet
u"\n\tI suppose ALL media want something to happen, otherwise what would\n\tthey rep
u"\nOh... I forgot... Art Shamsky, former Red and Mets player.  Batted .301\nbetween
u"I don't mean to be disrespectful to your concerns, but it seems to me \nthat you'r
u"Hello networld,\n\nI'm looking for an X mailreader. Is there a Xelm?\n\nAndreas\n\
u"\n\n\nIf by that you mean anything on the GD approach, there was an article on\nit
u"Any more news on Steve's status since he lost the starting job\nwould be appreciat
u'NUT CASE PANICS!!!!JUMPS THE GUN ON THE NET BEFORE GETTING FACTS STRAIGHT!!!!\n',
u"\n\nThere is a rite like this described in Joseph Campbell's\n_Occidental_Mythology
u"[reply to geb@cs.pitt.edu (Gordon Banks)]\n \n \n\nI made a decision a while back
u"I'm looking for some Game Boy games.  Please e-mail me with your list and offers!
u'The Chevrolet brothers were respected racers & test drivers for the\nBuick Co. whe
u"\n\tThere is a free program called 'xkernel' which does just that.\nIt is by Seth
u'\nThis is a very good point.  One that I have held for sometime.  We do not\nallow
u':    Indeed, if NSA really designed the algorithm to be secure, it\'s very likely\
u"\n\nI am told (by the person who I care a lot about and who I am worried\nis going
u"I posted this to the apps group and didn't get any response, so\nI'll try here. I
u'}>More like those who use their backs instead of their minds to make\n}>their livi
u'\nOne should be aware that foreign doctors admitted for training\nare ineligible t
...]
```

In [30]:
```
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
```

```
            print "Topic %d:" % (topic_idx)
            print " ".join([feature_names[i]
                            for i in topic.argsort()[:-no_top_words - 1:-1]])

    # dataset = fetch_20newsgroups(shuffle=True, random_state=1, remove=('headers',
    'footers', 'quotes'))
    # documents = dataset.data

    no_features = 1000

    # NMF is able to use tf-idf
    tfidf_vectorizer = TfidfVectorizer(max_df=0.95, min_df=2, max_features=no_features,
    stop_words='english')
    tfidf = tfidf_vectorizer.fit_transform(skl_texts)
    tfidf_feature_names = tfidf_vectorizer.get_feature_names()

    # LDA can only use raw term counts for LDA because it is a probabilistic graphical model
    tf_vectorizer = CountVectorizer(max_df=0.95, min_df=2, max_features=no_features,
    stop_words='english')
    tf = tf_vectorizer.fit_transform(skl_texts)
    tf_feature_names = tf_vectorizer.get_feature_names()

    no_topics = 10

    # Run NMF
    nmf = NMF(n_components=no_topics, random_state=1, alpha=.1, l1_ratio=.5,
    init='nndsvd').fit(tfidf)

    # Run LDA
    lda = LatentDirichletAllocation(n_topics=no_topics, max_iter=5,
    learning_method='online', learning_offset=50.,random_state=0).fit(tf)

    no_top_words = 10
    display_topics(nmf, tfidf_feature_names, no_top_words)
    display_topics(lda, tf_feature_names, no_top_words)
```

```
Topic 0:
afghanistan bin laden qaeda al force taliban tora bora afghan
Topic 1:
palestinian arafat israeli israel hamas gaza attack suicide sharon militant
Topic 2:
qantas union worker industrial maintenance dispute wage freeze action relations
Topic 3:
test africa south match day waugh bowler wicket cricket lee
Topic 4:
river guide adventure canyon court trip interlaken australians swiss accident
Topic 5:
detainee centre woomera detention facility department damage overnight visa night
Topic 6:
hollingworth dr governor abuse general anglican child school allegation statement
Topic 7:
new year australia south government people sydney australian wales state
Topic 8:
harrison beatle cancer george krishna lord lung know ceremony life
Topic 9:
commission hih royal collapse hearing company report union martin evidence
Topic 0:
space station shuttle endeavour russian crew ice vaughan centre launch
Topic 1:
test south day australia match lee africa wicket waugh cricket
Topic 2:
afghanistan force taliban government laden bin president australian united al
Topic 3:
russian people christmas authority security cause economy drop america kilometre
```

```
Topic 4:
union qantas worker industrial action company maintenance dispute pay relations
Topic 5:
palestinian israeli arafat attack hamas suicide gaza sharon israel kill
Topic 6:
win metre good year race event world new australia australian
Topic 7:
year company commission people australian report world director royal child
Topic 8:
new australia south people government sydney state australian storm year
Topic 9:
flight virgin disease airline melbourne blue tell second ansett japan
```

```
In [39]: tfidf_feature_names[2], tf_feature_names[2]
```

```
Out[39]: (u'absolutely', u'absolutely')
```

```
In [34]: for line in tf:
             print line
             break
```

```
  (0, 634)      2
  (0, 353)      3
  (0, 415)      3
  (0, 824)      2
  (0, 592)      8
  (0, 823)      6
  (0, 955)      4
  (0, 851)      1
  (0, 979)      2
  (0, 904)      1
  (0, 688)      1
  (0, 424)      1
  (0, 912)      1
  (0, 409)      3
  (0, 115)      2
  (0, 586)      1
  (0, 973)      2
  (0, 873)      3
  (0, 407)      2
  (0, 0)        1
  (0, 24)       1
  (0, 969)      2
  (0, 846)      1
  (0, 287)      1
  (0, 117)      1
  :     :
  (0, 679)      2
  (0, 992)      1
  (0, 223)      1
  (0, 91)       1
  (0, 835)      1
  (0, 183)      1
  (0, 521)      1
  (0, 898)      1
  (0, 69)       2
  (0, 700)      3
  (0, 325)      3
```

```
(0, 429)      1
(0, 597)      1
(0, 187)      1
(0, 139)      1
(0, 516)      1
(0, 838)      1
(0, 641)      1
(0, 807)      1
(0, 327)      1
(0, 198)      1
(0, 321)      1
(0, 671)      1
(0, 292)      1
(0, 344)      1
```

0.0.7 pyLDAvis

Thanks to pyLDAvis, we can visualise our topic models in a really handy way. All we need to do is enable our notebook and prepare the object.

```
In [22]: pyLDAvis.enable_notebook()
         pyLDAvis.gensim.prepare(ldamodel, corpus, dictionary)
```

```
Out[22]: PreparedData(topic_coordinates=            Freq  cluster  topics         x         y
         topic
         5      20.629337        1       1 -0.031791 -0.027891
         0      12.905266        1       2  0.004317  0.010642
         3      10.173109        1       3 -0.014312  0.011161
         6       9.588046        1       4  0.043393 -0.003117
         2       9.421682        1       5  0.029660 -0.027211
         7       7.752333        1       6 -0.045081  0.006402
         4       7.615487        1       7  0.037983  0.063091
         8       7.539375        1       8  0.039796 -0.047300
         1       7.404629        1       9 -0.041340 -0.014208
         9       6.970736        1      10 -0.022625  0.028433, topic_info=     Category
         term
         4617  Default  121.000000     australian  121.000000  30.0000  30.0000
         633   Default   38.000000          world   38.000000  29.0000  29.0000
         4761  Default   38.000000       pakistan   38.000000  28.0000  28.0000
         1873  Default   98.000000           year   98.000000  27.0000  27.0000
         926   Default   74.000000           fire   74.000000  26.0000  26.0000
         1489  Default   29.000000          india   29.000000  25.0000  25.0000
         4901  Default   27.000000         qantas   27.000000  24.0000  24.0000
         232   Default   50.000000          union   50.000000  23.0000  23.0000
         972   Default  122.000000      australia  122.000000  22.0000  22.0000
         3239  Default   69.000000            man   69.000000  21.0000  21.0000
         3225  Default   47.000000          claim   47.000000  20.0000  20.0000
         1092  Default   22.000000         indian   22.000000  19.0000  19.0000
         3006  Default   27.000000     commission   27.000000  18.0000  18.0000
         1826  Default   22.000000        economy   22.000000  17.0000  17.0000
         759   Default   60.000000        company   60.000000  16.0000  16.0000
```

```
298   Default  113.000000         people  113.000000  15.0000  15.0000
5280  Default   24.000000          metre   24.000000  14.0000  14.0000
1344  Default   99.000000     government   99.000000  13.0000  13.0000
858   Default    7.000000          virus    7.000000  12.0000  12.0000
5410  Default   37.000000         centre   37.000000  11.0000  11.0000
414   Default   20.000000         wicket   20.000000  10.0000  10.0000
4470  Default   21.000000       detainee   21.000000   9.0000   9.0000
2548  Default   75.000000         attack   75.000000   8.0000   8.0000
2644  Default   43.000000           cent   43.000000   7.0000   7.0000
3615  Default   59.000000           good   59.000000   6.0000   6.0000
1341  Default   50.000000         arrest   50.000000   5.0000   5.0000
1575  Default   19.000000           rate   19.000000   4.0000   4.0000
4011  Default   48.000000          south   48.000000   3.0000   3.0000
551   Default   13.000000        dispute   13.000000   2.0000   2.0000
4454  Default   18.000000          catch   18.000000   1.0000   1.0000
...       ...         ...            ...         ...      ...      ...
3225  Topic10    6.818107          claim   47.636185   0.7194  -5.6895
3422  Topic10    5.279711            pay   33.005504   0.8306  -5.9452
1277  Topic10    2.412867      explosive    9.586679   1.2839  -6.7282
239   Topic10    4.531026         worker   26.657778   0.8913  -6.0981
298   Topic10   10.952928         people  113.955664   0.3212  -5.2154
2817  Topic10    7.659687           tell   68.750885   0.4689  -5.5731
5098  Topic10    3.055805            hih   14.389699   1.1140  -6.4920
1054  Topic10    4.870582        fighter   32.538250   0.7642  -6.0258
3215  Topic10    2.617848     east_timor   11.217835   1.2083  -6.6467
972   Topic10    9.924347      australia  122.743143   0.1483  -5.3141
540   Topic10    4.357381        federal   28.541665   0.7840  -6.1372
3057  Topic10    6.720653  united_states   62.865899   0.4276  -5.7039
4188  Topic10    4.689575      bin_laden   34.033117   0.6815  -6.0637
5330  Topic10    2.424794        tension    9.999275   1.2467  -6.7233
759   Topic10    5.860094        company   60.959096   0.3214  -5.8409
1237  Topic10    3.519638           face   21.936677   0.8336  -6.3507
2092  Topic10    5.143894          month   51.682718   0.3561  -5.9712
2833  Topic10    5.158802           give   52.043946   0.3521  -5.9683
5256  Topic10    4.966436        include   47.878084   0.3975  -6.0063
1873  Topic10    6.404653           year   98.994679  -0.0746  -5.7520
2548  Topic10    5.673876         attack   75.720259   0.0723  -5.8732
3661  Topic10    5.013351        believe   53.273132   0.3001  -5.9969
1399  Topic10    5.428054       official   72.789889   0.0675  -5.9175
3981  Topic10    5.941918            day  101.159947  -0.1712  -5.8270
2783  Topic10    5.695442          force  103.558411  -0.2370  -5.8694
1344  Topic10    5.578603     government   99.616515  -0.2189  -5.8901
358   Topic10    4.951457           time   65.934225   0.0745  -6.0094
3180  Topic10    4.856833    afghanistan   69.817909  -0.0021  -6.0287
531   Topic10    4.762327           take   56.801343   0.1846  -6.0483
4753  Topic10    4.757573           come   59.989265   0.1290  -6.0493

[810 rows x 6 columns], token_table=      Topic      Freq       Term
```

```
term
2097      1  0.293088
2097      2  0.104674
2097      3  0.062805
2097      4  0.125609
2097      5  0.062805
2097      6  0.062805
2097      7  0.041870
2097      8  0.062805
2097      9  0.083739
2097     10  0.083739
2355      1  0.155569          $
2355      2  0.186682          $
2355      3  0.062227          $
2355      4  0.155569          $
2355      5  0.093341          $
2355      6  0.093341          $
2355      7  0.155569          $
2355      8  0.062227          $
2355      9  0.031114          $
2355     10  0.031114          $
4375      6  0.652501       10th
1761      1  0.723908       12th
2686      7  0.620177     1:00am
4768      2  0.553180         60
5201      6  0.658028     8:00am
1293      8  0.432622          >
2566      1  0.139190       able
2566      2  0.139190       able
2566      3  0.069595       able
2566      4  0.069595       able
...     ...       ...        ...
1873      1  0.131320       year
1873      2  0.212133       year
1873      3  0.060609       year
1873      4  0.191930       year
1873      5  0.070711       year
1873      6  0.040406       year
1873      7  0.121219       year
1873      8  0.080812       year
1873      9  0.040406       year
1873     10  0.060609       year
3762      1  0.195240  yesterday
3762      2  0.073215  yesterday
3762      3  0.097620  yesterday
3762      4  0.122025  yesterday
3762      5  0.073215  yesterday
3762      6  0.073215  yesterday
```

```
3762      7  0.122025  yesterday
3762      8  0.097620  yesterday
3762      9  0.073215  yesterday
3762     10  0.073215  yesterday
1347      1  0.704619   zimbabwe
1347      6  0.100660   zimbabwe
1347      9  0.100660   zimbabwe
1347     10  0.100660   zimbabwe
3628      1  0.533608      zinni
3628      3  0.088935      zinni
3628      5  0.088935      zinni
3628      6  0.088935      zinni
3628      8  0.088935      zinni
3628      9  0.088935      zinni

[2477 rows x 3 columns], R=30, lambda_step=0.01, plot_opts={'xlab': 'PC1', 'ylab': 'PC
```

0.0.8 Round-up

Okay - so what have we learned so far? By using spaCy, we cleaned up our data super fast. It's worth noting that by running our doc through the pipeline we also know about every single words POS-tag and NER-tag. This is useful information and we can do some funky things with it! I would highly recommend going through this repository to see examples of hands-on spaCy usage.

As for gensim and topic modelling, it's pretty easy to see how well we could create our topic models. Now the obvious next question is - how do we use these topic models? The news classification notebook in the Gensim notebooks directory is a good example of how we can use topic models in a practical scenario.

We will continue this tutorial by demonstrating a newer topic modelling features of gensim - in particular, Topic Coherence.

0.0.9 Topic Coherence

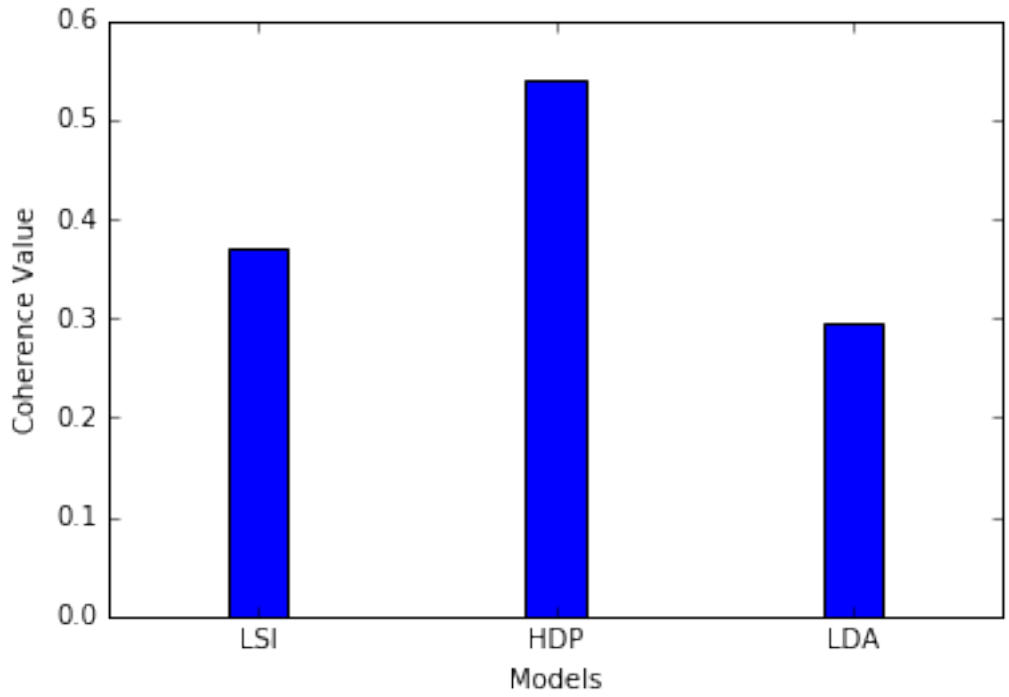
Topic Coherence is a new gensim functionality where we can identify which topic model is 'better'. By returning a score, we can compare between different topic models of the same. We use the same example from the news classification notebook to plot a graph between the topic models we have created.

```
In [18]: lsitopics = [[word for word, prob in topic] for topicid, topic in
         lsimodel.show_topics(formatted=False)]

         hdptopics = [[word for word, prob in topic] for topicid, topic in
         hdpmodel.show_topics(formatted=False)]

         ldatopics = [[word for word, prob in topic] for topicid, topic in
         ldamodel.show_topics(formatted=False)]

In [19]: lsi_coherence = CoherenceModel(topics=lsitopics[:10], texts=texts,
         dictionary=dictionary, window_size=10).get_coherence()

         hdp_coherence = CoherenceModel(topics=hdptopics[:10], texts=texts,
         dictionary=dictionary, window_size=10).get_coherence()
```

```
        lda_coherence = CoherenceModel(topics=ldatopics, texts=texts, dictionary=dictionary,
        window_size=10).get_coherence()
```

```
In [20]: def evaluate_bar_graph(coherences, indices):
             """
             Function to plot bar graph.

             coherences: list of coherence values
             indices: Indices to be used to mark bars. Length of this and coherences should be
         equal.
             """
             assert len(coherences) == len(indices)
             n = len(coherences)
             x = np.arange(n)
             plt.bar(x, coherences, width=0.2, tick_label=indices, align='center')
             plt.xlabel('Models')
             plt.ylabel('Coherence Value')
```

```
In [21]: evaluate_bar_graph([lsi_coherence, hdp_coherence, lda_coherence],
                            ['LSI', 'HDP', 'LDA'])
```



We can see that topic coherence helped us get past manually inspecting our topic models - we can now keep fine tuning our models and compare between them to see which has the best performance.

This also brings us to the end of the runnable part of this tutorial - we will continue however by briefly going over two more Jupyter notebooks I have previously worked on - mainly, Dynamic Topic Modelling and Document Word Coloring.