Topic Modeling

Praful Gupta
2018CSB1112

Indian Institute Of Technology
Ropar , Punjab

Abstract

This report includes analysis using Topic Modeling on given State of the Union and AP wire Stories Dataset using Lsi and Lda gensim models.

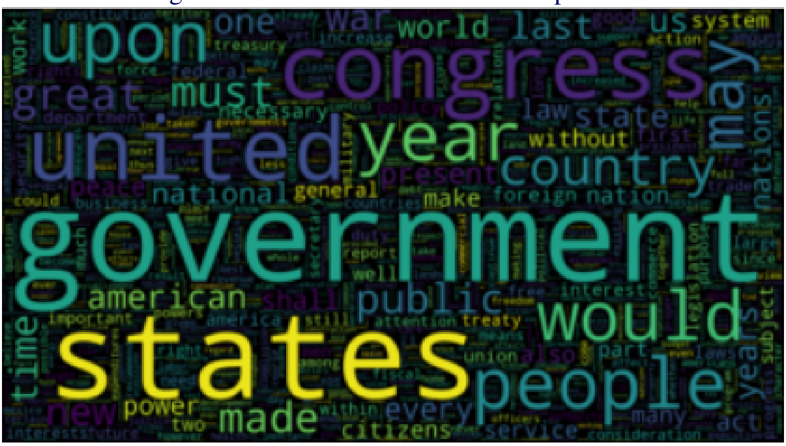
1 Topic Modelling and on State of the Union Dataset

The State of the Union is an annual address by the President of the United States before a joint session of congress. In it, the President reviews the previous year and lays out his legislative agenda for the coming year.
The data-set contains about 240 speeches between years 1790 and 2012

1.1 Preprocessing The Speeches

- Punctuation's and numbers were removed from the speeches and then using nltk library. English stop-words were removed.
- Frequency table was generated and words occurring less than 5 times were removed. (Figure one shows the Word Cloud of the frequency table)

Figure 1: Word Cloud for Processed Speeches



1.2 Creating Tfidf Vectors

- A dictionary was generated using gensim's corpora Dictionary method. Using that dictionary corpus was generated using dictionary's doc2bow method.

- Using gensim Tfidf model vectors were created for the precious corpus.

1.3 Finding Appropriate number of Topics for Given Data-Set

- We check the Coherence values using Lsi Models to find appropriate number of topics.

- We first check with topic values between 5-100 with a step of 5.(Shown in Figure 8)

- As we see the graph is increasing with steep increase between 2-25. Then we check with topic values between 2-25 again with step 1.(Shown in Figure 9)

- Lda Models Coherence values can be found in Figure 10.

- Using Lsi model the number of topics Predicted are about 8 and using the Lda model Number of topics predicted are about 16. (We use the Lsi model for further analysis so we take number of topics to be 8)

Figure 2: Coherence Values using Lsi model for Number of Topics Between 5 - 100 with step 5

```
Coherence Value for num_topics= 5  is  0.31982842763070246
Coherence Value for num_topics= 10  is  0.3889334274113749
Coherence Value for num_topics= 15  is  0.4050154215048631
Coherence Value for num_topics= 20  is  0.4027924823182823
Coherence Value for num_topics= 25  is  0.4134955314784509
Coherence Value for num_topics= 30  is  0.3989856491767648
Coherence Value for num_topics= 35  is  0.41416328155308835
Coherence Value for num_topics= 40  is  0.4289670875741831
Coherence Value for num_topics= 45  is  0.43727391978864083
Coherence Value for num_topics= 50  is  0.45894244346744256
Coherence Value for num_topics= 55  is  0.45142624509859647
Coherence Value for num_topics= 60  is  0.47514707209932044
Coherence Value for num_topics= 65  is  0.46121871138318693
Coherence Value for num_topics= 70  is  0.46792504157949333
Coherence Value for num_topics= 75  is  0.47553952630366175
Coherence Value for num_topics= 80  is  0.48062290414006786
Coherence Value for num_topics= 85  is  0.48438149282307824
Coherence Value for num_topics= 90  is  0.48375698043515986
Coherence Value for num_topics= 95  is  0.4868354943475704
Coherence Value for num_topics= 100  is  0.5002498313192807
```

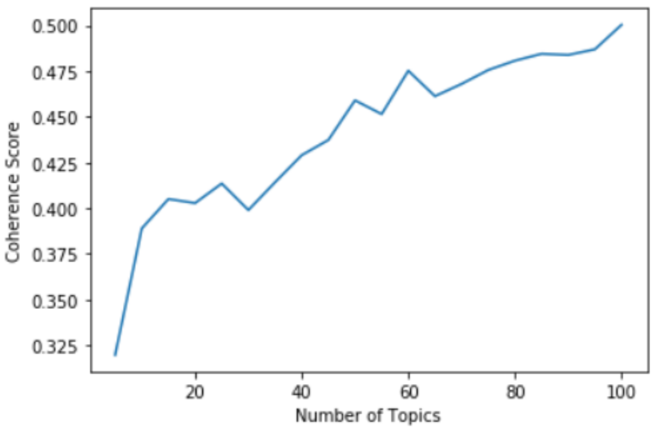


Figure 3: Coherence Values using Lsi model for Number of Topics Between 2 - 25 with step 1

```
Coherence Value for num_topics= 2  is  0.3014703376327279
Coherence Value for num_topics= 3  is  0.3420900059137872
Coherence Value for num_topics= 4  is  0.3372034402374564
Coherence Value for num_topics= 5  is  0.3672042426089052
Coherence Value for num_topics= 6  is  0.39771533084978145
Coherence Value for num_topics= 7  is  0.3899146955134409
Coherence Value for num_topics= 8  is  0.4663146689776875
Coherence Value for num_topics= 9  is  0.46042846666925613
Coherence Value for num_topics= 10  is  0.4050226842374256
Coherence Value for num_topics= 11  is  0.41086255449914044
Coherence Value for num_topics= 12  is  0.3908663792504066
Coherence Value for num_topics= 13  is  0.41652316125550815
Coherence Value for num_topics= 14  is  0.39916884648018286
Coherence Value for num_topics= 15  is  0.38722065864939315
Coherence Value for num_topics= 16  is  0.40095126411397664
Coherence Value for num_topics= 17  is  0.4106221881953701
Coherence Value for num_topics= 18  is  0.39707820704113217
Coherence Value for num_topics= 19  is  0.42311134256328414
Coherence Value for num_topics= 20  is  0.39361852654048646
Coherence Value for num_topics= 21  is  0.4101434812257851
Coherence Value for num_topics= 22  is  0.40414345616968
Coherence Value for num_topics= 23  is  0.40869477760637607
Coherence Value for num_topics= 24  is  0.42743185093822667
```

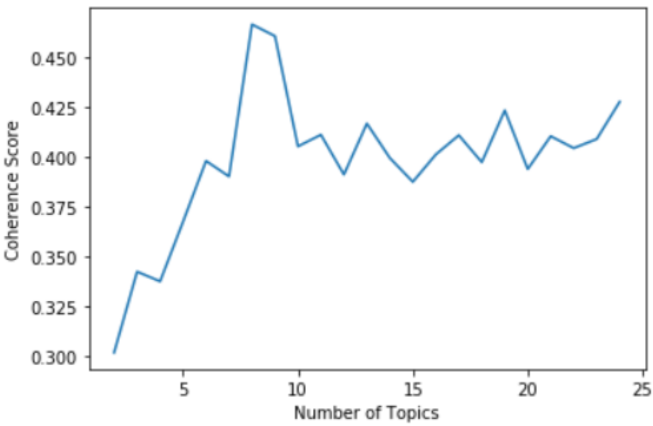
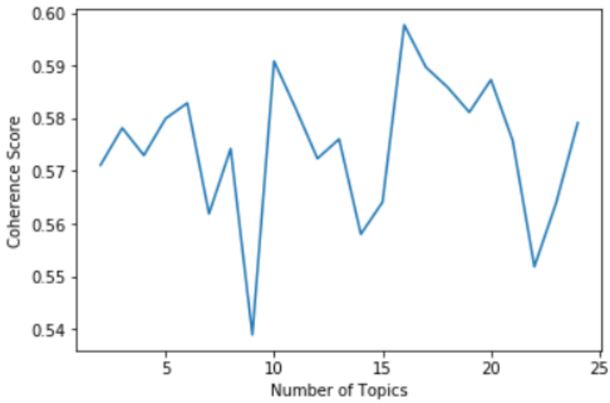


Figure 4: Coherence Values using Lda model for Number of Topics Between 2 - 25 with step 1

```
Coherence Value for num_topics= 2  is  0.5711203979983086
Coherence Value for num_topics= 3  is  0.5781589238323497
Coherence Value for num_topics= 4  is  0.5729682382244139
Coherence Value for num_topics= 5  is  0.5799524569678575
Coherence Value for num_topics= 6  is  0.5828971945465499
Coherence Value for num_topics= 7  is  0.5618661543465319
Coherence Value for num_topics= 8  is  0.5742326121347834
Coherence Value for num_topics= 9  is  0.5388853152971835
Coherence Value for num_topics= 10  is  0.5908510826563929
Coherence Value for num_topics= 11  is  0.5818407136169895
Coherence Value for num_topics= 12  is  0.5723679287866267
Coherence Value for num_topics= 13  is  0.5760580651548849
Coherence Value for num_topics= 14  is  0.5580051923009381
Coherence Value for num_topics= 15  is  0.5640914304266981
Coherence Value for num_topics= 16  is  0.5977378504737618
Coherence Value for num_topics= 17  is  0.589699562344598
Coherence Value for num_topics= 18  is  0.5859051093252605
Coherence Value for num_topics= 19  is  0.5811595677798136
Coherence Value for num_topics= 20  is  0.5873239795094488
Coherence Value for num_topics= 21  is  0.5757644892029903
Coherence Value for num_topics= 22  is  0.551829449486992
Coherence Value for num_topics= 23  is  0.5638844227267573
Coherence Value for num_topics= 24  is  0.5791446661942
```



1.4 Topics Detail

- Following are the Given Topics predicted by the Lsi model. [
  Lsi Model Topics Are

- (0, [('tribes', 0.017462194042481453),
  ('communicated', 0.017446540061313756),
  ('extensive', 0.017364412694110102),
  ('interesting', 0.017360463553673596),
  ('objects', 0.01734684005778459),
  ('expedient', 0.017297663131170368),
  ('fellowcitizens', 0.01728799457199079),
  ('manufactures', 0.017243535794947352),
  ('article', 0.017207347732291243),
  ('enlightened', 0.01720644628271886)]),

- (1, [('jobs', 0.0446811153518045),
  ('tonight', 0.04464966128740647),
  ('nuclear', 0.04367920185400383),
  ('commitment', 0.04283577046925831),
  ('job', 0.04269302676930842),
  ('percent', 0.042491333075452296),
  ('programs', 0.041762189496397796),
  ('challenges', 0.041445959115161496),
  ('technology', 0.041317682478339585),
  ('billion', 0.0410715497000257)]),

- (2, [('barbary', 0.04113349740148824),
  ('weve', 0.04002907350074656),
  ('tonight', 0.03927077752768368),
  ('ensuing', 0.03888665530781813),
  ('cooperative', -0.03872279791260953),
  ('method', -0.037570205966162l4),
  ('bless', 0.037493722199719054),
  ('interstate', -0.03738176752960894),
  ('militia', 0.037310256138892196),
  ('administrative', -0.03663080723046625)]),

- (3, [('objectives', 0.048062239675209994),
  ('mediterranean', 0.045397546771088565),
  ('collective', 0.042763424020835106),
  ('posture', 0.04273954769540204),
  ('aggression', 0.042011191700680545),

('atomic', 0.04145760972923314),
('barbary', 0.0409577406851493 8),
('labormanagement', 0.04042785302758908),
('objective', 0.03927854345342361),
('cooperative', 0.03832007498253019)]),

- (4, [('interstate', -0.04928635607795380 7),
('muscle', -0.04565934186297095),
('ensuing', -0.04438500148877331),
('marketing', -0.04370296454147274),
('th', -0.04359832428192144 5),
('shoals', -0.0423843514523368),
('proofs', -0.04186666249023194 4),
('depending', -0.04054682280910659),
('moderation', -0.03998601521988546),
('militia', -0.03965340460958656 5)]),

- (5, [('woodrow', -0.05712623881097163),
('fivetwenties', 0.04351123057467303),
('wilson', -0.04339928555670558),
('buildings', 0.04233144949224068 6),
('venezuela', 0.04219836596582226),
('centennial', 0.04203547801720103),
('arthur', 0.04188424735914639),
('etc', 0.04154856883657003 6),
('repayment', 0.0413616676567107),
('utter', -0.04134164899598925)]),

- (6, [('shoals', 0.05776010109102927),
('muscle', 0.05552080543617285 4),
('gun', -0.04442041997020465 5),
('regulatory', 0.04064400839828259),
('conciliation', 0.03997242115785415 4),
('edicts', -0.03903950732468335),
('retaliation', -0.03876044723170706),
('pacification', -0.03857105153502072),
('marketing', 0.0385655225656633 4),
('philippines', -0.0384195411188173 1)]),

- (7, [('cheney', 0.06437683939370846),
('coalition', 0.0612846953571354),

('ghent', 0.0574543982627838),
('iraqis', 0.0569008096814775l),
('faithbased', 0.0559697298548594 64),
('iraqi', 0.0557094360868182),
('al', 0.0544525229543597 26),
('homeland', 0.0538059121541499 3),
('w', 0.0535727583193513 56),
('terrorist', 0.0521115899075307 2)])]

- Short Description about the given topics are- Topic 0 - About Country Upliftment and Expences
  Topic 1 - About Job Opportunity and Technology Advancement
  Topic 2 - About War and Millitary
  Topic 3 - About Planning and War
  Topic 4 - Doesnt Make any Sense
  Topic 5 - About Areas
  Topic 6 - About Foreign Regulation and War
  Topic 7 - About War and Terrirism

- Following are the Given Topics predicted by the Lda model. [
  Lsi Model Topics Are

- (0, [('crimes', 0.00020110916),
  ('color', 0.00018630698),
  ('louisiana', 0.00018451946),
  ('conformity', 0.00018433567),
  ('examples', 0.00018275737),
  ('examine', 0.00018161422),
  ('organizing', 0.00018000664),
  ('fulfillment', 0.00017958351),
  ('fortunately', 0.00017843739),
  ('enduring', 0.00017725879)]),

- (1, [('leadership', 0.00018087168),
  ('misery', 0.00018026207),
  ('play', 0.00018018816),
  ('lift', 0.00017945761),
  ('poor', 0.00017942072),
  ('goals', 0.00017842274),
  ('incentives', 0.00017793893),
  ('space', 0.00017773628),
  ('speaker', 0.00017699544),
  ('inflation', 0.00017660524)]),

- (2, [('experienced', 0.00019583029),
  ('entirely', 0.00019479048),
  ('sincere', 0.00019370114),
  ('combined', 0.00019201568),
  ('statement', 0.00018981003),
  ('contest', 0.00018763327),
  ('patriotic', 0.00018703837),
  ('slave', 0.00018644273),
  ('reimbursement', 0.00018544069),
  ('injustice', 0.0001849274)]),

- (3, [('seize', 0.00021218021),
  ('rural', 0.00021184223),
  ('double', 0.00021132233),
  ('grow', 0.00021087268),
  ('tempting', 0.00020901489),
  ('programs', 0.00020614667),

('ahead', 0.00020091562),
('play', 0.00020011407),
('proud', 0.00019915277),
('around', 0.0001989628)]),

- (4, [('ascertained', 0.00025869),
('tranquillity', 0.00024015976),
('effectual', 0.00023589794),
('providence', 0.00022716665),
('engage', 0.00022139522),
('belligerent', 0.00022122938),
('supplied', 0.00021870481),
('constituents', 0.00021684852),
('considerations', 0.00021378897),
('exertions', 0.00021173165)]),

- (5, [('solve', 0.00019653529),
('outside', 0.00018867577),
('length', 0.00018592199),
('broken', 0.00018425663),
('air', 0.00018366489),
('warfare', 0.00018297728),
('marked', 0.00018200712),
('respects', 0.00017984795),
('seeking', 0.00017776419),
('settle', 0.00017693912)]),

- (6, [('quickly', 0.00021429753),
('patriotism', 0.00020209927),
('council', 0.00019675095),
('pledge', 0.00019588406),
('accommodation', 0.00019524702),
('belongs', 0.0001950628),
('happiness', 0.00019464531),
('soviet', 0.00019387303),
('eyes', 0.00019274918),
('ensure', 0.00019188011)]),

- (7, [('expectation', 0.00020352917),
('decay', 0.0001955889),
('spanish', 0.00019000907),

('presidential', 0.0001874696),
('brave', 0.0001866724),
('cruisers', 0.00018486429),
('paris', 0.0001846007),
('funding', 0.00018336992),
('infractions', 0.00018326132),
('congressional', 0.0001827047)])]

- Short Description about the given topics are- Topic 0 - About Country Upliftment and Crimes
  Topic 1 - About Future Goals and Economy
  Topic 2 - About Slavery and Justice
  Topic 3 - Doesnt Make any Sense
  Topic 4 - About Country and War
  Topic 5 - Doesnt Make any Sense
  Topic 6 - About Pledges
  Topic 7 - About Economy of Country

- In Lda only Positive values are assigned to the weight of bag of words in a topic whereas Lsi have negative values to strongly distinguish distant keywords within a topics.

- Lsi Topics make more sence together than Lda model.

- Training Lsi is much Faster than Lda Model

- Lsi prediced about half the number of topics (8) than predicted by using Lda model (16).

- Lda Models's Topic are more appropriate and resembles a specific topic better than Lsi Model.

1.5 Summary

- We can See there Were Specific Range of Years when a specific Topic was actively being used
- Topic 6 Which is centered around World War 2 contains strongly war related words.
- Topic 3's have many occurrence around the Civil Right moment (1940 - 1960) and contains strong words like labourmanagement , collective , cooperative etc related to that moment.
- Topic 2 is around the Barbary War and contains related words to that topic.
- Topic 1 is Currenly lastes grossing topics and contains topic about Technology advancement.

Figure 5: Usage Of Speech Spanning Over Years (Topic on Y axis , Years on X axis)

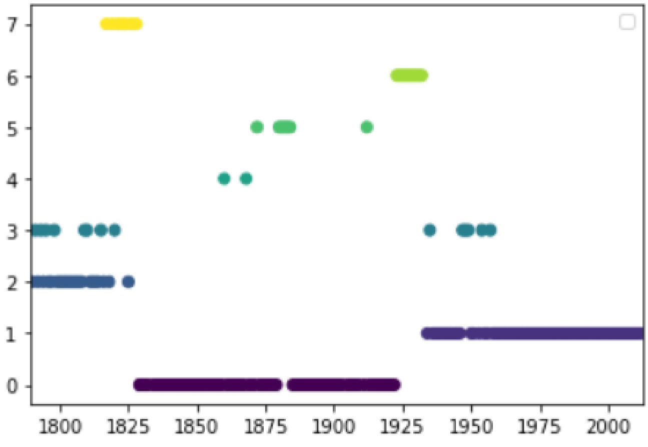
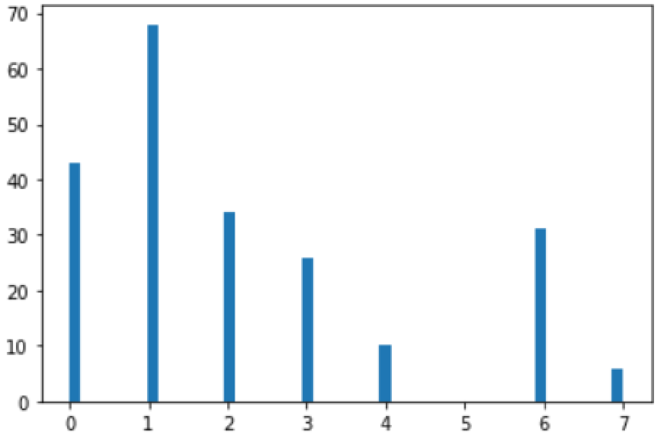


Figure 6: Usage Of Speech Spanning Over Years (Topic on X axis , Number Of Speech on Y axis)



2 Topic Modelling and on AP Wire Stories Dataset

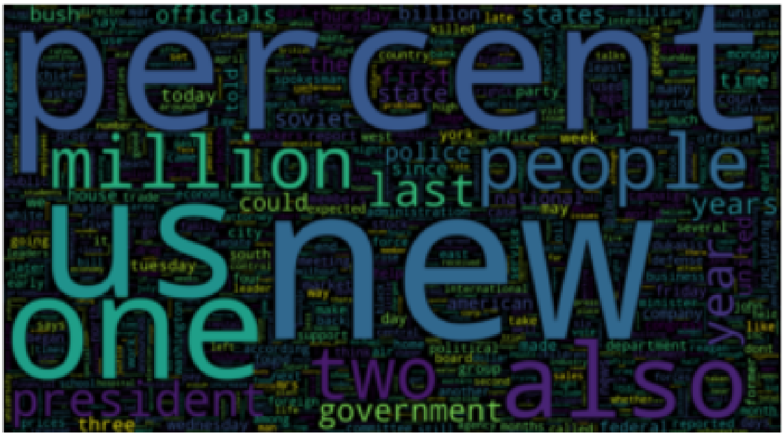
The Associated Press' data team uses data.world to share data to help news organizations tell stories in their cities, counties and states.
The data-set contains about 2250 speeches between years 1790 and 2012

2.1 Preprocessing The Speeches

- Punctuation's and numbers were removed from the speeches and then using nltk library. English stop-words were removed.

- Frequency table was generated and words occurring less than 5 times were removed. (Figure 7 shows the Word Cloud of the frequency table)

Figure 7: Word Cloud for Processed Speeches



2.2 Creating Tfidf Vectors

- A dictionary was generated using gensim's corpora Dictionary method. Using that dictionary corpus was generated using dictionary's doc2bow method.

- Using gensim Tfidf model vectors were created for the precious corpus.

2.3 Finding Appropriate number of Topics for Given Data-Set

- We check the Coherence values using Lsi Models to find appropriate number of topics.

- We first check with topic values between 5-100 with a step of 5.(Shown in Figure 2)

- As we see the graph is increasing with steep increase between 2-25. Then we check with topic values between 2-25 again with step 1.(Shown in Figure 3)

- Lda Models Coherence values can be found in Figure 4.

- Using Lsi model the number of topics Predicted are about 4 and using the Lda model Number of topics predicted are about 9. (We use the Lsi model for further analysis so we take number of topics to be 4)

Figure 8: Coherence Values using Lsi model for Number of Topics Between 5 - 100 with step 5

```
Coherence Value for num_topics= 5  is  0.5719247406544795
Coherence Value for num_topics= 10  is  0.4572103358009647
Coherence Value for num_topics= 15  is  0.36873208505818766
Coherence Value for num_topics= 20  is  0.361044077694194
Coherence Value for num_topics= 25  is  0.43077278508262296
Coherence Value for num_topics= 30  is  0.35442473384218776
Coherence Value for num_topics= 35  is  0.3730123877156479
Coherence Value for num_topics= 40  is  0.35957265845823494
Coherence Value for num_topics= 45  is  0.36501098189358094
Coherence Value for num_topics= 50  is  0.3666762758575321
Coherence Value for num_topics= 55  is  0.36569310226311363
Coherence Value for num_topics= 60  is  0.379950828606449
Coherence Value for num_topics= 65  is  0.37076280967584785
Coherence Value for num_topics= 70  is  0.3860968234104301
Coherence Value for num_topics= 75  is  0.3739891052075726
Coherence Value for num_topics= 80  is  0.3868253552113491
Coherence Value for num_topics= 85  is  0.3893246284356489
Coherence Value for num_topics= 90  is  0.40388659966083257
Coherence Value for num_topics= 95  is  0.3949055861472612
Coherence Value for num_topics= 100  is  0.40068108731237934
Coherence Value for num_topics= 105  is  0.40492811931640765
```

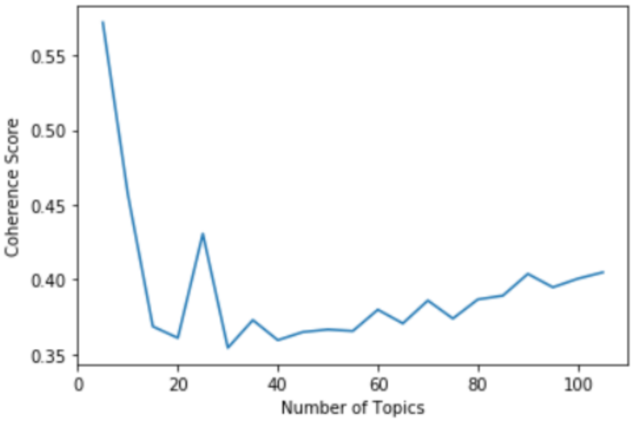


Figure 9: Coherence Values using Lsi model for Number of Topics Between 2 - 25 with step 1

```
Coherence Value for num_topics= 2  is  0.30954809188433074
Coherence Value for num_topics= 3  is  0.5035114300835511
Coherence Value for num_topics= 4  is  0.6576310269360467
Coherence Value for num_topics= 5  is  0.5866897473804772
Coherence Value for num_topics= 6  is  0.6156969440041531
Coherence Value for num_topics= 7  is  0.6234472588599816
Coherence Value for num_topics= 8  is  0.5853462502455834
Coherence Value for num_topics= 9  is  0.5301943165068084
Coherence Value for num_topics= 10  is  0.4566058058054357
Coherence Value for num_topics= 11  is  0.435803981408405
Coherence Value for num_topics= 12  is  0.4204643975421818
Coherence Value for num_topics= 13  is  0.43983124652938127
Coherence Value for num_topics= 14  is  0.41077939790658297
Coherence Value for num_topics= 15  is  0.4186125143181322
Coherence Value for num_topics= 16  is  0.4227140356705027
Coherence Value for num_topics= 17  is  0.38258856201311325
Coherence Value for num_topics= 18  is  0.41017166952553946
Coherence Value for num_topics= 19  is  0.4142424786735709
Coherence Value for num_topics= 20  is  0.41302510778611073
Coherence Value for num_topics= 21  is  0.40782043051363914
Coherence Value for num_topics= 22  is  0.36546250968717325
Coherence Value for num_topics= 23  is  0.39748170836581875
Coherence Value for num_topics= 24  is  0.3689446407374713
```

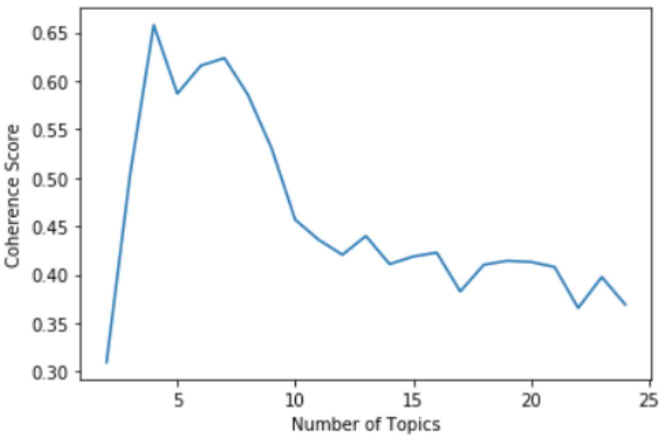
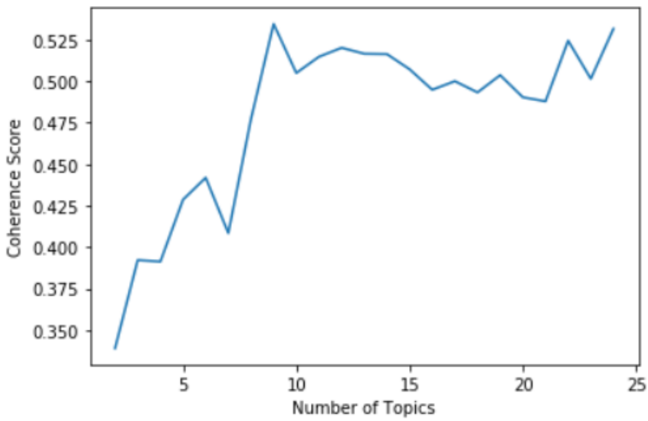


Figure 10: Coherence Values using Lda model for Number of Topics Between 2 - 25 with step 1

```
Coherence Value for num_topics= 2  is  0.33915843661998285
Coherence Value for num_topics= 3  is  0.3922485299523905
Coherence Value for num_topics= 4  is  0.3913434641878085
Coherence Value for num_topics= 5  is  0.42864404223565444
Coherence Value for num_topics= 6  is  0.441911287286613
Coherence Value for num_topics= 7  is  0.4085015513581954
Coherence Value for num_topics= 8  is  0.4772870130989449
Coherence Value for num_topics= 9  is  0.5345472955957518
Coherence Value for num_topics= 10  is  0.5049833850728815
Coherence Value for num_topics= 11  is  0.5147662140300829
Coherence Value for num_topics= 12  is  0.5201733832739449
Coherence Value for num_topics= 13  is  0.5165908198171524
Coherence Value for num_topics= 14  is  0.516375263528495
Coherence Value for num_topics= 15  is  0.5072406924381679
Coherence Value for num_topics= 16  is  0.49488258784938666
Coherence Value for num_topics= 17  is  0.5000490757205694
Coherence Value for num_topics= 18  is  0.49328248318577445
Coherence Value for num_topics= 19  is  0.503689731549836
Coherence Value for num_topics= 20  is  0.49036210161200283
Coherence Value for num_topics= 21  is  0.48790621485587643
Coherence Value for num_topics= 22  is  0.5244750965306029
Coherence Value for num_topics= 23  is  0.5015069826736096
Coherence Value for num_topics= 24  is  0.5316507946279079
```



2.4 Topics Detail

- Following are the Given Topics predicted by the Lsi model. [
  Lsi Model Topics Are

- (0, [('percent', 0.13872200190343137),
  ('bush', 0.11205348322242695),
  ('soviet', 0.10378761214331984),
  ('million', 0.08949234034358147),
  ('police', 0.0880543260633868),
  ('government', 0.0806330951764572),
  ('dukakis', 0.08023166970173935),
  ('us', 0.07929644914143005),
  ('stock', 0.07909691690883816),
  ('billion', 0.07696482886989682)]),

- (1, [('stock', 0.2626253587849622),
  ('index', 0.23089997498795925),
  ('yen', 0.19841949271129256),
  ('dollar', 0.18321128542482346),
  ('points', 0.1772781162939468),
  ('market', 0.17270840928987452),
  ('exchange', 0.16171242786563694),
  ('trading', 0.14381775525277837),
  ('shares', 0.14165541133492887),
  ('prices', 0.1400594603603592)]),

- (2, [('dukakis', -0.6070187772679716),
  ('bush', -0.28644268980108484),
  ('jackson', -0.24921161982986179),
  ('campaign', -0.1583087827861026),
  ('convention', -0.14553469456499682),
  ('bentsen', -0.13728773982258036),
  ('democratic', -0.1249864973245429),
  ('poll', -0.1064937166305671),
  ('republican', -0.10124721952381714),
  ('presidential', -0.10123917466693842)]),

- (3, [('yen', -0.37760893819493846),
  ('dollar', -0.34616325903778355),
  ('percent', 0.19390846614989396),
  ('ounce', -0.18476685775121798),

('gold', -0.1844652106935693),
('bid', -0.17037110934385236),
('late', -0.16722149003710005),
('london', -0.16136818499846328),
('francs', -0.14753438670313337),
('german', -0.11369087943198944)])]

- Short Description about the given topics are- Topic 0 - About Government and Country
  Topic 1 - About Economy of Country
  Topic 2 - About Election and Politics
  Topic 3 - About Prices and Global Economy

- Following are the Given Topics predicted by the Lda model. [
  Lsi Model Topics Are

- (0, [('iraq', 0.0037951125),
  ('kuwait', 0.002370425),
  ('iraqi', 0.0020041554),
  ('saddam', 0.0016193549),
  ('gulf', 0.0015876857),
  ('baghdad', 0.0015569825),
  ('bush', 0.001406033),
  ('saudi', 0.0013336844),
  ('percent', 0.0013016361),
  ('arabia', 0.0011976055)]),

- (1, [('million', 0.0012143492),
  ('percent', 0.0011862465),
  ('october', 0.0011232881),
  ('party', 0.0009929037),
  ('getz', 0.0009647185),
  ('hubbert', 0.00085083046),
  ('opposition', 0.00084403506),
  ('ershad', 0.000840004),
  ('new', 0.0008056607),
  ('mrs', 0.00080137025)]),

- (2, [('percent', 0.0033409519),
  ('million', 0.0016201037),
  ('index', 0.0014306721),
  ('points', 0.001370489),
  ('prices', 0.0013213306),
  ('milken', 0.0012625147),
  ('soviet', 0.0012499173),
  ('stock', 0.0012304388),
  ('shares', 0.0011757393),
  ('financial', 0.0011637734)]),

- (3, [('police', 0.0021370954),
  ('students', 0.0013386154),
  ('nosair', 0.0012829889),
  ('karpov', 0.0011463232),
  ('kasparov', 0.0011288534),
  ('court', 0.0010971264),

('game', 0.0010214546),
('united', 0.0009927868),
('soviet', 0.00094452273),
('reported', 0.00092499267)])]

- Short Description about the given topics are- Topic 0 - About Global News (Mostly Arab States)
  Topic 1 - Doesnt Make any Sense
  Topic 2 - About Economy and Finance
  Topic 3 - About Crime and Protests

- In Lda only Positive values are assigned to the weight of bag of words in a topic whereas Lsi have negative values to strongly distinguish distant keywords within a topics.

- Lsi Topics make more sence together than Lda model.

- Training Lsi is much Faster than Lda Model

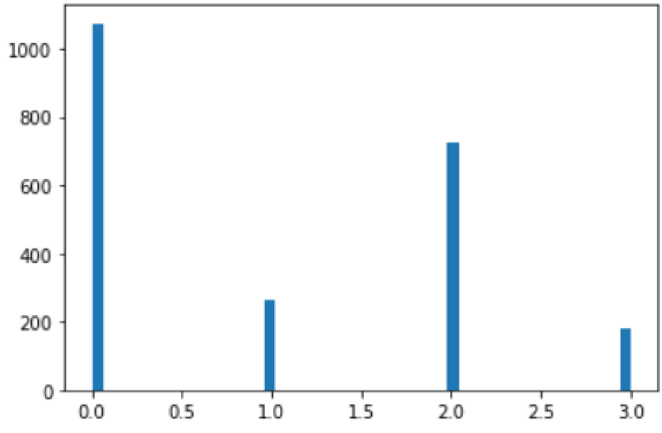
- Lsi prediced about half the number of topics (4) than predicted by using Lda model (9).

- Lsi Models's Topic are more appropriate and resembles a specific topic better than Lda Model.

2.5 Summary

- There we see the distribution of the topics in the histogram in Figure 11

Figure 11: Usage Of Speech Spanning Over Years (Topic on X axis , Number Of Speech on Y axis)



3 Observation and Summary

We can see from both the topic modelling that the AP Wire Dataset has less number of topics that the State of the Union dataset. Also it preform better to distinguishes the topics in the AP wire Dataset.

- Maximum coherence value for AP Wire Dataset reached a maximum of 0.65 whereas it was 0.46 for State of the Union dataset.
- State of the Union dataset contained certain words which occurred in almost every document like 'government' , 'states' etc. Which confused the model to distinguish the topics.