**Articles : 2500**

**Topics : 5**

Concept 0:

romney

gingrich

santorum

obama

people

campaign

state

Concept 1:

romney

gingrich

santorum

iowa

hampshire

south carolina

campaign

Concept 2:

game

season

points

team

wizards

games

play

Concept 3:

iran

obama

iranian

government

military

nuclear

taliban

Concept 4:

google

sopa

sites

piracy

wikipedia

internet

online

**Checking the results obtained (for first 15 documents):**

Using k means (first 15 documents cluster numbering):

|  |
| --- |
| 4 |
| 3 |
| 0 |
| 4 |
| 3 |
| 1 |
| 0 |
| 3 |
| 4 |
| 4 |
| 3 |
| 3 |
| 4 |
| 1 |
| 3 |

**Groups** :

{1, 4, 9, 10, 13}

{2, 5, 8, 11, 12, 15}

{3, 7}

{6, 14}

**Using lsa (first 15 documents plotting (x,y) points):**

x axis [ -8.087902 -2.891658 6.205677 -11.604637 -0.3203161 5.635933

6.1271524 -0.7888022 -8.563436 -10.857158 -3.4420178 -7.7822437

-11.039599 3.6988852 -3.5251436]

y axis [-3.3307571 4.5389795 -9.443952 -6.2534056 2.5448472 0.05497414

-6.760261 4.6975994 -4.809026 -5.5664964 1.4661603 -0.5303667

-5.217061 -2.335261 3.485784 ]

G**roups** :

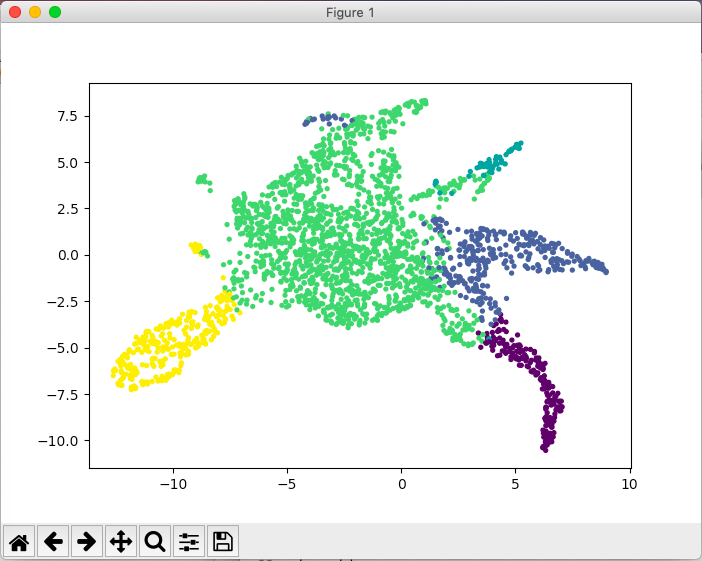
{yellow : 1, 4, 9, 10, 13}

{ green : 2, 5, 8, 11, 12, 15}

{violet : 3, 7}

{blue : 6, 14}

**Checking actual docs manually for topics:**

1. Baseball game, teams, nfl etc

2. Planet Venus, stargazing, outer space

3. Political campaigns news

4. Baseball game news

5. Housing, Taxes, finance, loan building

6. Soldier, islam militants, al-quadea, America, attacks, government

7. Election, republican, voting, Obama, political campaigns

8. homicide, crime, killing, police, victims, violent, arrest, robberies

9. nfl, coach, American football, playoffs, games

10. frustrated player, nfl, games, goal, season, playoff

11. temperature, warmest day, national weather, warm winters

12. nfl playoffs, playoff, home field, raiders team

13. football, injuries, defence, team, play, game, Virginia, practice

14. gulf oil, congressman, politics, tv, presidential elections, oil spill

15. catholic church, married priest, followers, catholicism, us, movement, liberal ,gay marriage, protests, rights, religion.

**Interpreting manual check results:**

The LSA classifier classifies sports and government related news pretty good.

1, 4, 9, 10 and 13 are completely related to sports. But it wrongly classifies 12 as non-sports.

Similarly, document 3 and 7 are related to election campaigns which is classified accurately

Document 6 and 14 are related to international affairs between governments and wars etc which is put in a separate group and classified accurately.

Document 2, 5, 8, 11 and 15 do not belong to a specific group. They are about outer space, weather, crime etc which is put in same group because total number of topics were 5.

Overall, lsa gives a good generalisation and classification of documents.