

21 Nov 2021

# CATEGORIZE THE ARTICLE TOPICS

Hussain Alhadab

Ahmed Alonaizi

Omar Alhadi

Feras Alyahya

Mohammed Alhamoud



A close-up photograph of a person's hand holding a smartphone. The screen displays a news website with a blue header and various news articles. The background is dark, and the phone is held at an angle.

# Table of Contents

## Part 1: Introduction

.....

## Part 2: Methodology

.....

## Part 3: Models

.....

## Part 4: Conclusion

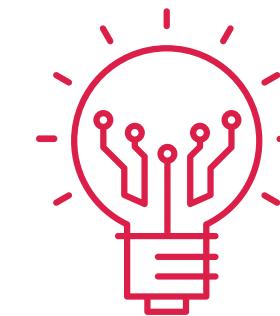
.....

# Introduction



## Problem:

- With a large number of articles and daily news and the increase in the length of articles on websites, Site administrators face difficulty in categorizing the subject of articles, the goal of categorizing articles is to help the administrators of the site to arrange articles in the related sections, and to suggest articles to users by their interests and previous readings.



## Solutation

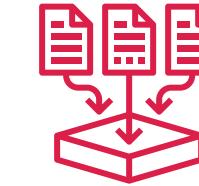
- Use Natural Language Processing (NLP) models to categorize the article topics



## Objectives

- Create a models to categorize the article topics
- Build Recommendations system

# Methodology :



## Data Sources:

- The dataset contains a variety of articles in different Fields, and we collect the data from BBC News it contains 2126 articles.



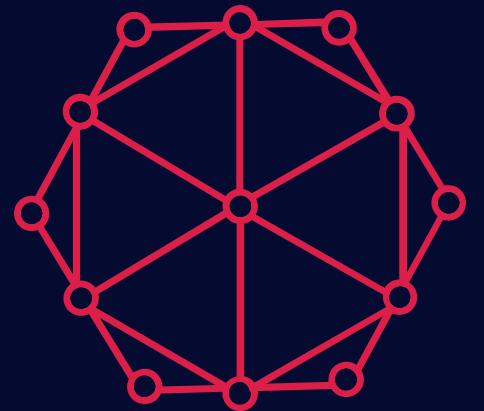
## Tools:



## Preprocess:

- Remove all stop words
- Remove emojis
- Remove all punctuations
- Remove URLs

# Models:



# K-Mean

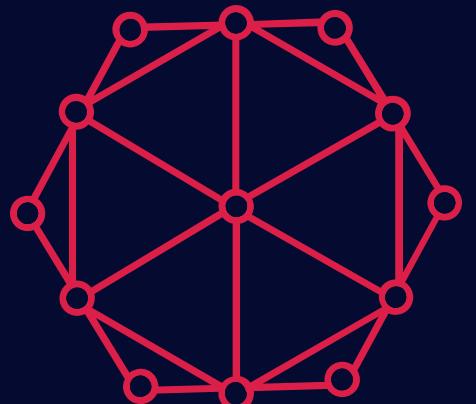
- **Topic 1**
  - music, game, album, band, chart, gadget, song, single, player, will, said, video, Sony, number, console, sale, digital, Nintendo, gaming, device
- **Topic 2**
  - labor, election, party, Blair, tory, said, brown, government, Howard, minister, lord, chancellor, prime, will, plan, conservative, people, leader, campaign, Kennedy
- **Topic 3**
  - phone, mobile, user, people, software, said, service, technology, computer, broadband, site, will, network, system, program, Microsoft, search, firm, virus, email
- **Topic 4**
  - award, best, film, oscar, nomination, actor, prize, ceremony, category, actress, aviator, winner, named, year, nominated, star, director, foxx, comedy, nominee
  -
- **Topic 5**
  - economy, growth, rate, economic, price, dollar, said, bank, year, market, rise, figure, deficit, spending, export, economist, quarter, month, consumer, china



## (K-Mean) Results:

- Cluster 1: Music industry
- Cluster 2: Governments and Politics
- Cluster 3: Technology
- Cluster 4: Film industry
- Cluster 5: Economy
- Cluster 6: Companies & Businesses
- Cluster 7: Sports
- Cluster 8: Sports
- Cluster 9: Companies & Businesses
- Cluster 10: Film industry

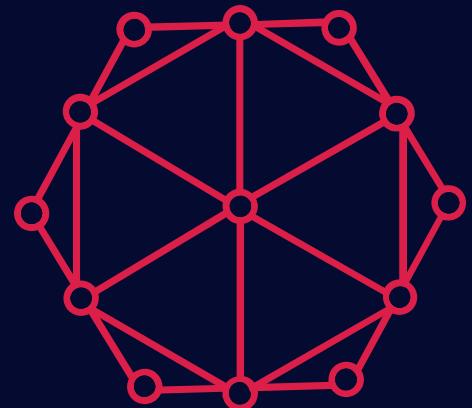
# Models:



# K-Mean

		text	cleaned	Topic_KMeans
1401	brazil plays down varig rescue the brazilian g...	[brazil, play, varig, rescue, brazilian, gover...	Companies & Businesses	
115	byrds producer melcher dies at 62 record produ...	[byrd, producer, melcher, dy, record, producer...	Companies & Businesses	
1834	school sport is back says pm tony blair has...	[school, sport, back, say, tony, blair, promis...	Companies & Businesses	
664	no uk premiere for rings musical the producers...	[premiere, ring, musical, producer, behind, lo...		Film industry
2010	fast moving phone bugs appear security firms a...	[fast, moving, phone, bug, appear, security, f...		Technology
2120	ferguson rues failure to cut gap boss sir alex...	[ferguson, rue, failure, bos, alex, ferguson, ...		Sports
2198	asian banks halt dollar s slide the dollar reg...	[asian, bank, halt, dollar, slide, dollar, reg...		Economy
1993	parmalat bank barred from suing bank of americ...	[parmalat, bank, barred, suing, bank, america,...	Companies & Businesses	
1421	roddick splits from coach gilbert andy roddick...	[roddick, split, coach, gilbert, andy, roddick...		Sports
1027	uk coal plunges into deeper loss shares in uk ...	[coal, plunge, deeper, loss, share, coal, fall...	Companies & Businesses	

# Models:

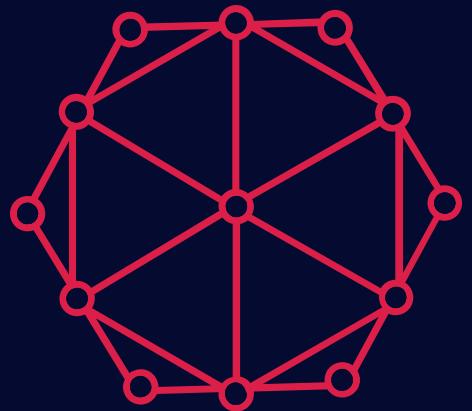


# Latent Semantic Analysis (LSA)

- Topic 1  
said, will, year, people, labor, game, election, government, film, party, Blair, brown, tory, time, world
- Topic 2
  - software, microsoft, program, virus, search, user, email, site, security, computer, spyware, window, england, spam, film
- Topic 3  
film, award, best, england, labour, oscar, blair, game, actor, election, star, party, wale, brown, tory
- Topic 4  
economy, growth, rate, bank, economic, price, dollar, year, rise, china, quarter, film, sale, economist, export
- Topic 5
  - kenteris, olympic, athens, iaaf, thanou, greek, phone, mobile, test, race, drug, athlete, athletics, indoor, champion

## (LSA) Results:

- |   |   |
|---|---|
| <ul style="list-style-type: none"><li>• Topic 1: Politics</li><li>• Topic 2: Technology</li><li>• Topic 3: Film industry</li><li>• Topic 4: Economy</li><li>• Topic 5: Sports</li></ul> | <ul style="list-style-type: none"><li>Topic 6: Legal</li><li>Topic 7: Technology</li><li>Topic 8: Technology</li><li>Topic 9: Sports</li><li>Topic 10: Music industry</li></ul> |
|---|---|

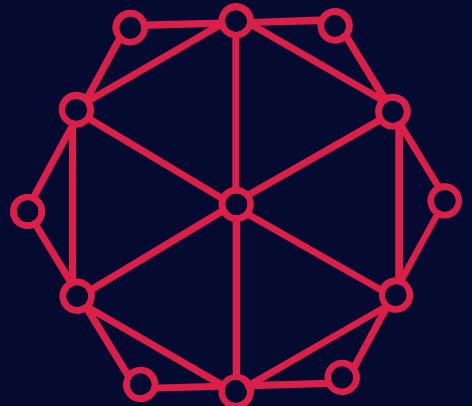


# Models:

# Latent Semantic Analysis (LSA)

	Politics	Politics	Film industry	Film industry	Economy	Legal	Technology	Technology	Sports	Music industry
text										
buyers snap up jet airways shares investors have snapped up shares in jet airways india s biggest airline following the launch of its much anticipated initial public offer (ipo). the ipo for 17.3 million shares was fully sold within 10 minutes of opening on friday. analysts expect jet to raise at least 16.4bn rupees (\$375m; £198m) from the offering. interest in jet s ipo has been fuelled by hopes for robust growth in india s air travel market. the share offer representing about 20% of jet s equity was oversubscribed news agency reuters reported. jet which was founded by london-based travel agent naresh goyal plans to use the cash to buy new planes and cut its debt. the company has grown rapidly since it launched operations in 1993 overtaking state-owned flag carrier indian airlines. however it faces stiff competition from rivals and low-cost carriers. jet s ipo is the first in a series of expected share offers from indian companies this year as they move to raise funds to help them do business in a rapidly-growing economy.	0.142	-0.023	-0.115	-0.006	0.112	0.045	-0.057	0.022	-0.059	0.039
wenger shock at newcastle dip arsenal manager Arsene Wenger has admitted he is at a loss to explain why Newcastle are languishing in the bottom half of the table. the gunners travel to st james park on wednesday with Newcastle 14th in the premiership after a troubled season. and wenger said: at the beginning of the season you would expect them to be fighting for the top four. i don t know how they got to be where they are. it looks to me from the outside that they have many injuries. Arsenal go into the game on the back of a 2-0 victory over fulham on sunday. and wenger added: the best way to prepare for a game is to win the previous one. we will go to newcastle in good shape. fatigue won t play too big a part in the next few weeks as we have players coming back so i can rotate a bit more. we do not play a season with 11 players and i believe that all of our squad deserve a chance in the team. striker thierry henry along with robert pires scored against fulham. and henry afterwards described the display as beautiful to watch . he said: what matters is winning and the three points of course. that is the only thing that really matters. but it is more enjoyable when you play like we did against fulham. we are playing as a team and that is important because there were some games when we maybe were not there as a team and suffered for that. those were games we lost.	0.155	-0.109	0.125	-0.181	-0.037	-0.032	-0.003	-0.026	-0.073	0.159

# Models:



# Non-Negative Matrix Factorization (NMF)

## Topic 1

- mobile, phone, music, technology, digital, people, gadget, service, device, will, video, broadband, game, camera, handset

## Topic 2

- labour, election, blair, party, tory, brown, said, howard, minister, government, chancellor, prime, will, plan, conservative

## Topic 3

- game, england, player, will, wale, match, said, team, ireland, play, side, injury, rugby, club, champion

## Topic 4

- film, award, best, oscar, actor, star, actress, festival, nomination, director, aviator, prize, year, comedy, movie

## Topic 5

- growth, economy, rate, price, economic, year, bank, said, market, rise, dollar, figure, quarter, sale, china

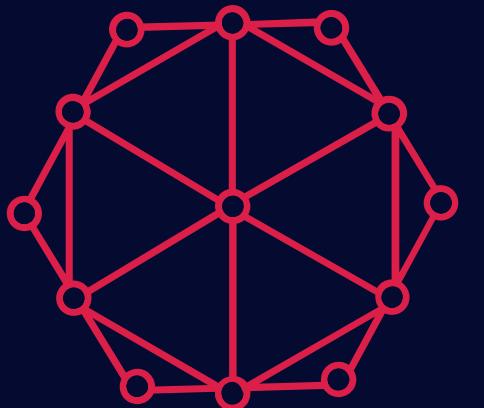
## (NMF) Results:

- Topic 1: Technology
- Topic 2: Politics
- Topic 3: Sports
- Topic 4: Film industry
- Topic 5: Economy

Topic 6: Legal

Topic 7: Technology

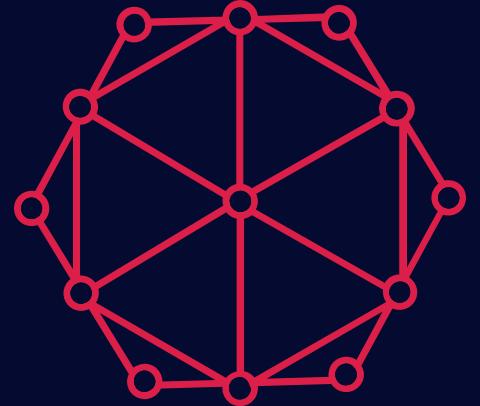
# Non-Negative Matrix Factorization (NMF)



## Models:

	Technology	Politics	Sports	Film industry	Economy	Legal	Technology
text	0.034	0.00	0.0	0.000	0.000	0.024	0.064
us hacker breaks into t-mobile a man is facing charges of hacking into computers at the us arm of mobile phone firm t-mobile. the californian man nicholas lee jacobsen was arrested in october. mr jacobsen tried at least twice to hack t-mobile's network and took names and social security numbers of 400 customers said a company spokesman. the arrest came a year after t-mobile uncovered the unauthorised access. the us secret service has been investigating the case. t-mobile has stringent procedures in place where we monitor for suspicious activity so that limited his activities and we were able to take corrective action immediately peter dobrow a t-mobile spokesperson said. it is thought that mr jacobsen's hacking campaign took place over at least seven months during which time he read e-mails and personal computer files according to court records. although mr jacobsen 21 managed to get hold of some data it is thought he failed to get customer credit card numbers which are stored on a separate computer system said mr dobrow. t-mobile confirmed that the us secret service was also looking into whether the hacker accessed photos that t-mobile subscribers had taken with their camera phones. the associated press agency reported that mr jacobsen also read personal files on the secret service agent who was apparently investigating the case. a los angeles grand jury indicted mr jacobsen with intentionally accessing a computer system without authorisation and with the unauthorised impairment of a protected computer between march and october 2004. he is currently on bail. t-mobile is a subsidiary company of deutsche telekom and has about 16.3 million subscribers in the us.	0.000	0.00	0.0	0.000	0.174	0.000	0.000
house prices suffer festive fall uk house prices fell 0.7% in december according to figures from the office of the deputy prime minister. nationally house prices rose at an annual rate of 10.7% in december less than the 13.7% rise the previous month. the average uk house price fell from £180 126 in november to £178 906 reflecting recent land registry figures confirming a slowdown in late 2004. all major uk regions apart from northern ireland experienced a fall in annual growth during december. december is traditionally a quiet month for the housing market because of christmas celebrations. however recent figures from the land registry - showing a big drop in sales between the last quarter of 2004 and the previous year - suggested the slowdown could be more than a seasonal blip. the volume of sales between october and december dropped by nearly a quarter from the same period in 2003 the land registry said. although both the office of the deputy prime minister (odpm) and the land registry figures point to a slowdown in the market the most recent surveys from nationwide and halifax have indicated the market may be undergoing a revival. after registering falls at the back-end of 2004 halifax said house prices rose by 0.8% in january and nationwide reported a rise of 0.4% in the first month of the year.	0.161	0.01	0.0	0.005	0.000	0.000	0.007

# Models:



# Latent Dirichlet Allocation (LDA)

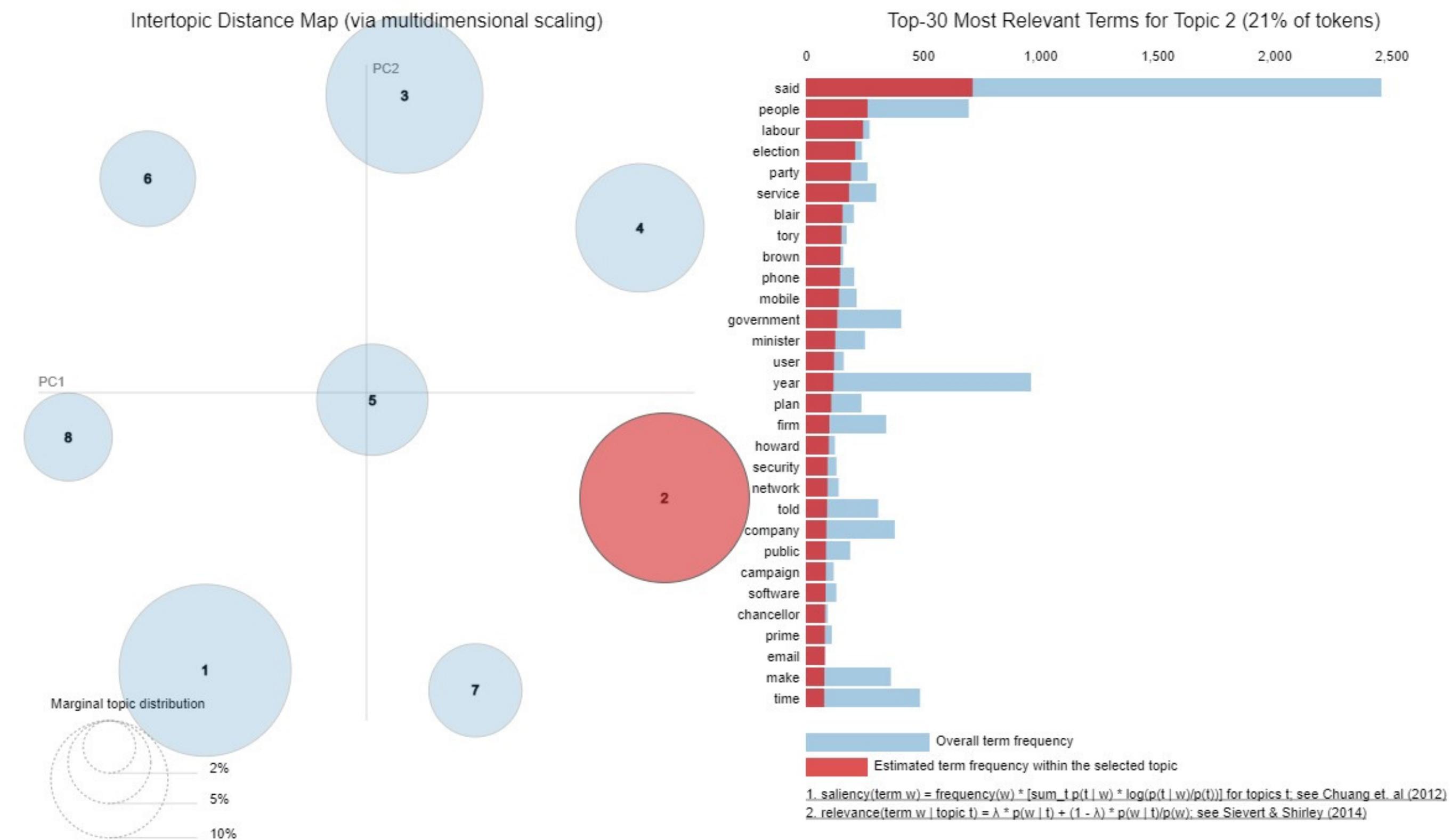
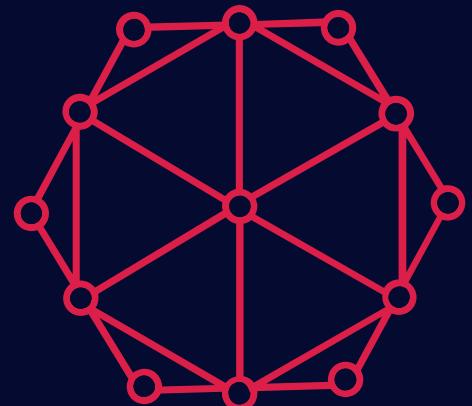
- **Topic 1**  
film, said, best, award, actor, director, year, star, party, oscar, british, actress, london, role, ukip
- **Topic 2**
  - said, open, match, year, game, final, play, test, world, roddick, australian, injury, second, champion, number
- **Topic 3**  
music, said, song, band, year, album, people, number, search, site, single, chart, record, radio, artist
- **Topic 4**
  - said, people, game, technology, government, year, make, time, lord, digital, work, want, right, video, home
- **Topic 5**  
game, said, england, player, club, wale, ireland, team, time, rugby, play, chelsea, minute, united, good

## (LDA) Results:

- **Topic 1: Film industry**
- **Topic 2: Sports**
- **Topic 3: Music**
- **Topic 4: Technology**
- **Topic 5: Sports**
- Topic 6: Economy**
- Topic 7: Film industry**
- Topic 8: Economy**

# Latent Dirichlet Allocation (LDA)

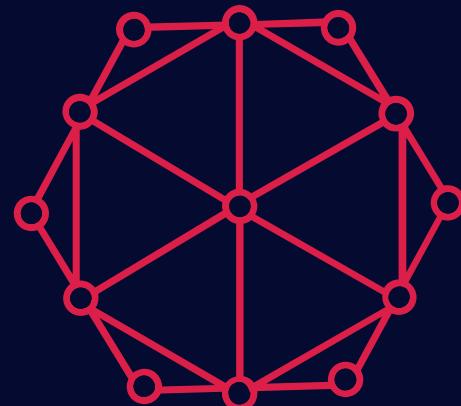
## Models:



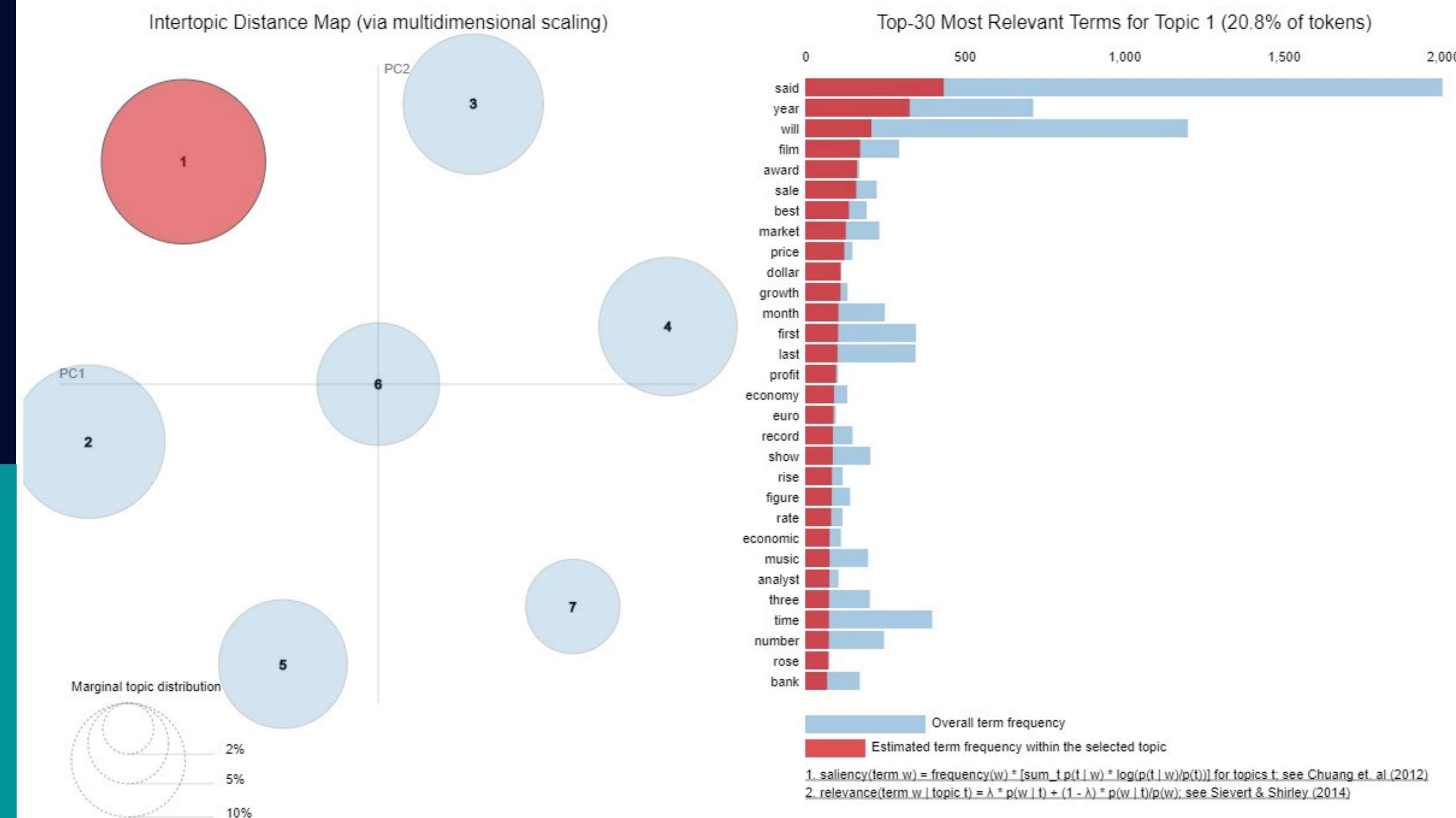
## Gensim LDA Result

- Topic 1: Sports
- Topic 2: Sports
- Topic 3: Film industry
- Topic 4: Technology
- Topic 5: Businesses
- Topic 6: Politics
- Topic 7: Technology

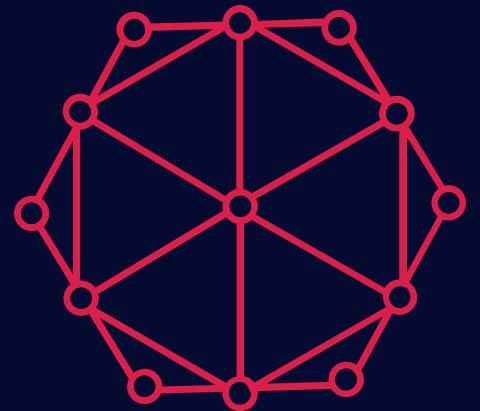
## Models:



# Gensim LDA



# Models:



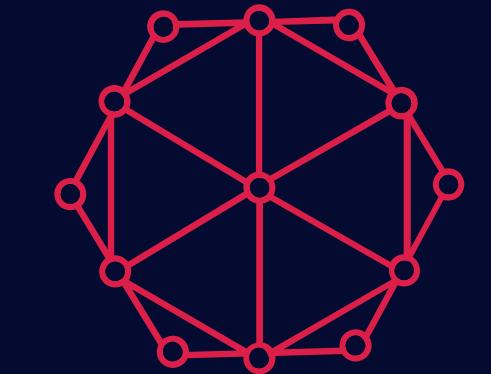
# CorEx

- **Topic 1**  
star, game, coach, champion, yearold, match, award, injury, play, season
- **Topic 2**
  - user, technology, computer, digital, online, software, device, website, video, internet
- **Topic 3**
  - tory, labour, conservative, election, party, blair, liberal, tony, leader, democrat
- **Topic 4**
  - firm, market, company, analyst, share, growth, consumer, business, price, product
- **Topic 5**
  - england, rugby, wale, ball, ireland, thing, think, chance, know, robinson

## CorEx Results:

- |  |   |
|--|---|
| <ul style="list-style-type: none"><li>• <b>Topic 1: Sports</b></li><li>• <b>Topic 2: Technology</b></li><li>• <b>Topic 3: Politics</b></li><li>• <b>Topic 4: Businesses</b></li><li>• <b>Topic 5: Sports</b></li></ul> | <b>Topic 6: Politics</b><br><b>Topic 7: Technology</b><br><b>Topic 8: Economy</b> |
|--|---|

# Recommendations



## Models:

===== Queried article details =====

headline : labour targets hardcore truants a fresh crackdown on persistent truants in england has been launched by education secretary ruth kelly. serial truants make up one in 13 pupils. previous initiatives brought 40 000 pupils back to school since 1997 according to official statistics. parenting contracts penalty notices and fast track prosecution systems have been used to tackle what has been a stubborn problem. it is thought that nearly half a million children skip school each day. tories say labour s previous success regarding the issue came because it tackled the easy part of the problem by reducing authorised absence where parents are permitted to take children out of school. such absences are often due to family holidays. however serial truants avoid the classroom despite government schemes costing £885m. those missing classes are more likely to become involved in crime as well as failing academically. measures such as parenting contracts and penalty notices were adopted by most local education authorities last term and come into force in the remainder this term. in one local education authority alone 800 parents were warned they would receive a penalty notice unless their child s attendance improved. the tough stance paid off with just 24 issued while attendance improved in 776 cases. truancy has been reduced by 5% at the 128 worst hit schools through the government s behaviour improvement programme. this is the equivalent of 200 pupils back in classes since september. the new measures come on top of national truancy sweeps - the sixth of which will take place on monday. police and education welfare officers patrol problem hotspots picking up truants and returning them to school. he ld twice each year in addition to routine local patrols previous country-wide sweeps have apprehended 31 000 pupils dodging school. in almost 14 000 of those cases the youngsters were accompanied by their parents. a department for education and skills source said: every day in school counts. it is clear form these figures that schools and local education authorities are now seizing the tools we have given them to improve school attendance and crack down hard on the very small numbers of pupils which account for almost half of the nation s truancy.

===== Recommended articles : =====

C:\Users\oalha\anaconda3\lib\site-packages\ipykernel\ipkern...: DeprecationWarning: `should\_run\_async` will not call `transform\_cell` automatically in the future. Please pass the result to `transformed\_cell` argument and any exception that happen during the transform in `preprocessing\_exc\_tuple` in IPython 7.17 and above.  
and should\_run\_async(code)

	headline	Euclidean similarity with the queried article
1	blair joins school sailing trip the prime minister has donned a life jacket and joined school ch...	1.177945
2	kelly trails new discipline power teachers could get more powers to remove unruly pupils from cl...	1.223273
3	school sport is back says pm tony blair has promised that sport is back as a priority for s...	1.243559
4	schools to take part in mock poll record numbers of schools across the uk are to take part in a ...	1.269675
5	schools to take part in mock poll record numbers of schools across the uk are to take part in a ...	1.269675
6	tories pledge free sports lessons children would be offered two hours free sports training a we...	1.281005
7	pupils to get anti-piracy lessons lessons on music piracy and copyright issues are to be taught ...	1.283174
8	research fears over kelly s views scientists have expressed concerns that new education secretar...	1.288923
9	faith schools citizenship warning schools must improve the quality of citizenship lessons - or s...	1.293649
10	student inequality exposed teenagers from well-off backgrounds are six times more likely to go...	1.298162

# Most 100 Common Words

# Conclusion:

In this project, We applied several types of NLP models in unsupervised dataset to identify news topics in BBC News dataset, Also we have noticed the diversity of topics from one to another, finally we developed a recommendations system to suggest similar topics to BBC News users.



# Thank you!