TOPIC MODELLING (Unsupervised ML) ON SCIENTIFIC NEWS ARTICLES

ZEHRA KEZER

INTRODUCTION

Popular Science (also known as PopSci) is an American digital magazine carrying popular science content, which refers to articles for the general reader on science and technology subjects.

POPULAR SCIENCE

1. **Goal**

   Analyzing scientific news articles pulled from the popsci website with topic modeling

2. **Tools**

   Analyzed with python using libraries sklearn, nltk, matplotlib, wordcloud, pyLDAvis, gensim, pandas, matplotlib. And store in MongoDB

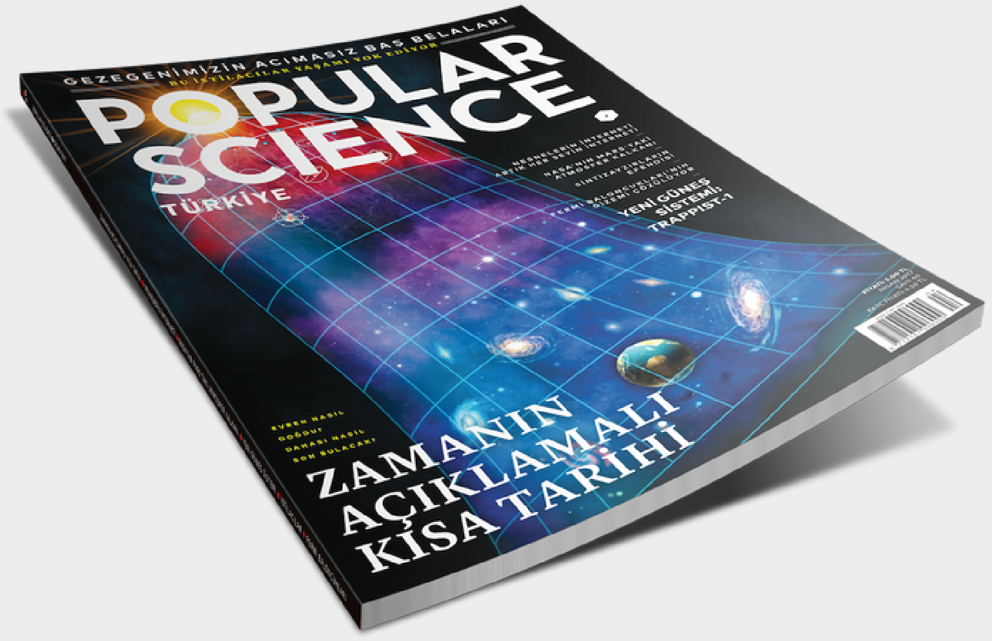
3. **Process**

   web scraping popsci then analyzing with Linear Discriminant Analysis (LDA)

Explore the Data

1 Extract data

	title	summary	bodytext	category
0	Utah teens will need parents' permission to us...	The new laws' broad language sets a curfew f...	Utah's governor signed two bills into law on T...	Technology
1	The first 3D printed rocket launch was both a ...	Relativity Space's Terran rocket failed to a...	Third time was unfortunately not the charm for...	Technology
2	This ATV-mounted, drone-killing laser burns wi...	The system was on display at a recent defens...	Earlier this month, Japan's Kawasaki Heavy Ind...	Technology
3	Don't plug in mysterious USB drives	From malware to more extreme scenarios, ther...	An Ecuadorian journalist has been injured by a...	Technology
4	The universe is getting a weigh-in thanks to AI	Step right up on the galactic scale, Alpha C...	Literally weighing the universe may sound like...	Technology




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9605 entries, 0 to 9604
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   title     9604 non-null   object
 1   summary   9594 non-null   object
 2   bodytext  9581 non-null   object
 3   category  9605 non-null   object
dtypes: object(4)
memory usage: 300.3+ KB
```

2 Data cleaning

Drop:

- Duplicates
- NaN values
- Stop words
- Punctuations

Stemming (Lemmatization)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9578 entries, 0 to 9604
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   category  9578 non-null   object
 1   article   9578 non-null   object
dtypes: object(2)
memory usage: 224.5+ KB
```

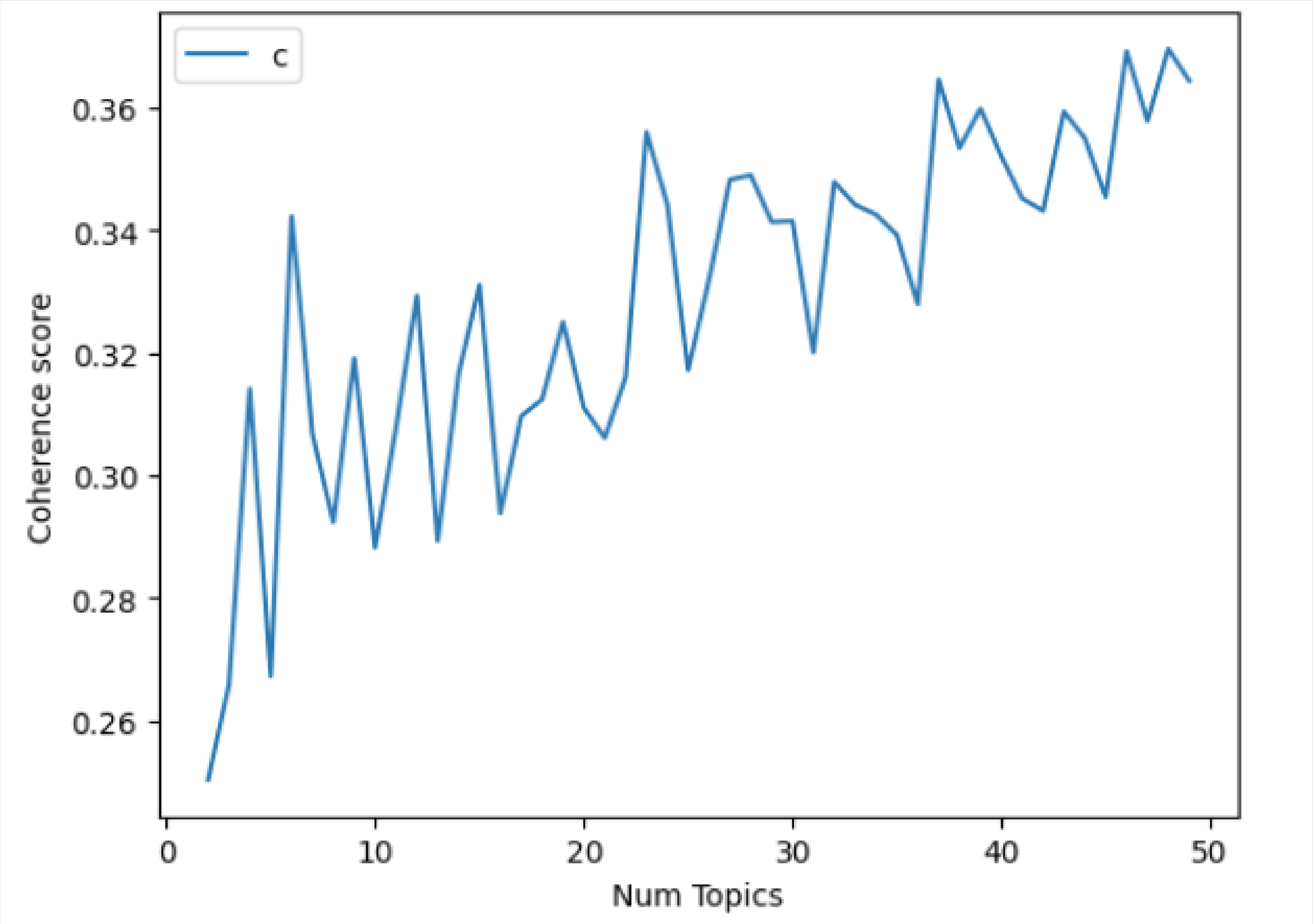
'atv mounted drone killing laser burns power one dishwasher system display recent defense conference needs kilowatts power work earlier month japan’s kawasaki heavy industries showed new tool fighting drones enclosed cabin top four wheel atv frame system mounts high energy laser back alongside power needed make work part growing arsenal counter drone weapons one fits expanded role arsenal japan’s modern military'

POS Tagging →

'drone laser burns power dishwasher system defense conference needs power work month japans industries tool drones wheel atv frame system mounts energy laser power make work part counter drone weapons role japans'

Model Development with LDA

Find best topic number



Model perplexity and topic coherence provide a convenient measure to judge how good a given topic model is.

Num Topics = 6

Perplexity: -7.880144192420245
Coherence score: 0.35466101147887574

The keywords for each topic and the importance of each keyword

Topic 1

'0.018*"game" +
0.017*"device" +
0.016*"time" +
0.014*"home" +
0.012*"computer" +
0.012*"phone" +
0.012*"way" +
0.011*"tv" + 0.010*"gift"
+ 0.010*"music"'

Topic 2

'0.021*"kid" +
0.013*"child" +
0.013*"conversation" +
0.009*"year" +
0.009*"bike" +
0.009*"article" +
0.008*"parent" +
0.008*"story" +
0.008*"people" +
0.007*"animal"'

Topic 3

'0.015*"year" +
0.013*"air" +
0.012*"car" +
0.009*"state" +
0.009*"time" +
0.008*"death" +
0.008*"season" +
0.008*"weather" +
0.008*"system" +
0.008*"vehicle"

Topic 4

'0.022*"vaccine" +
0.021*"people" +
0.021*"health" +
0.019*"week" +
0.015*"thing" +
0.012*"covid" +
0.011*"time" +
0.011*"science" +
0.011*"disease" +
0.011*"case"

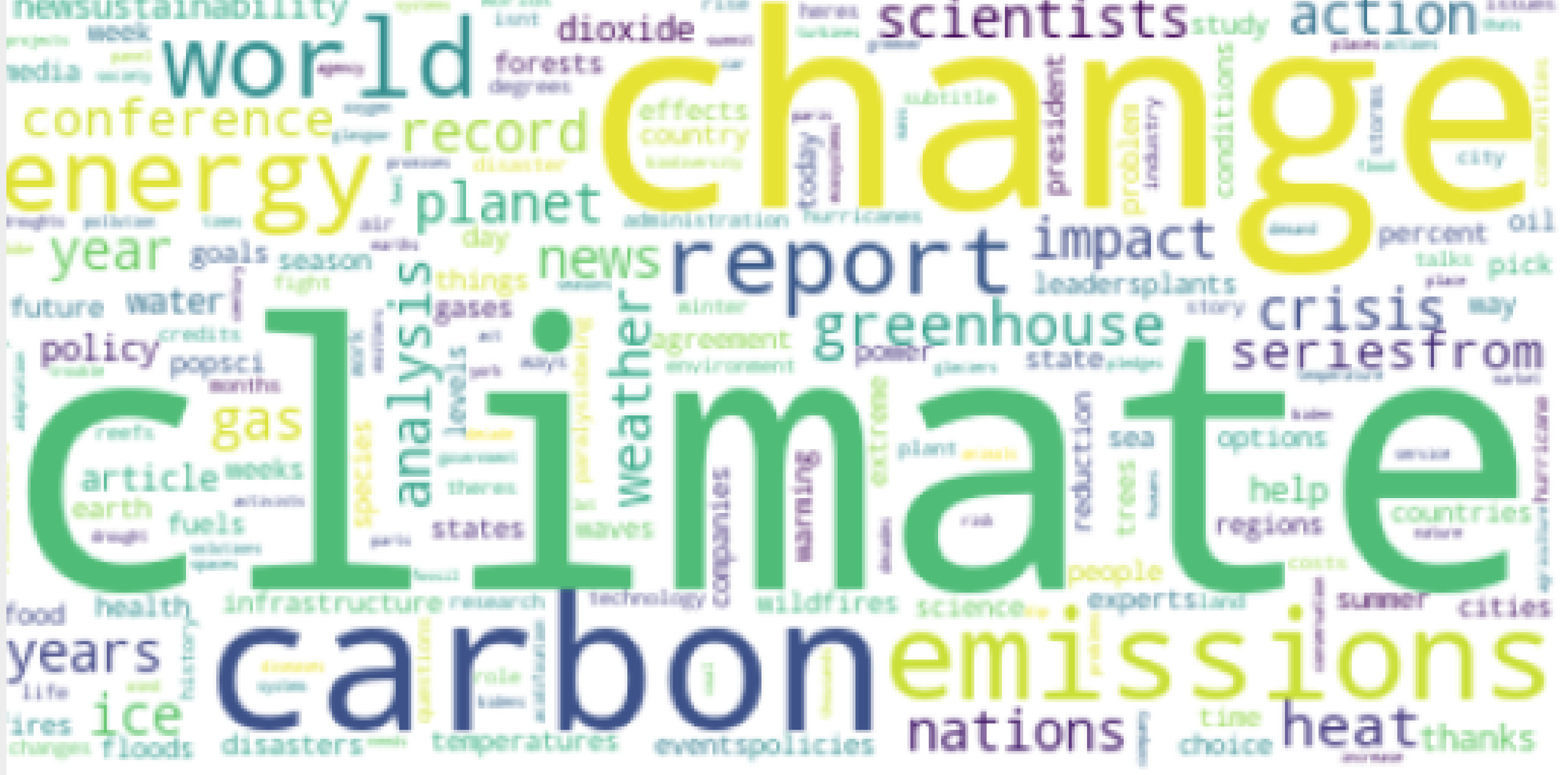
Topic 5

'0.018*"home" +
0.017*"dog" +
0.015*"food" +
0.013*"plant" +
0.012*"power" +
0.011*"way" +
0.008*"water" +
0.008*"market" +
0.008*"gas" +
0.007*"climate"'

Topic 6

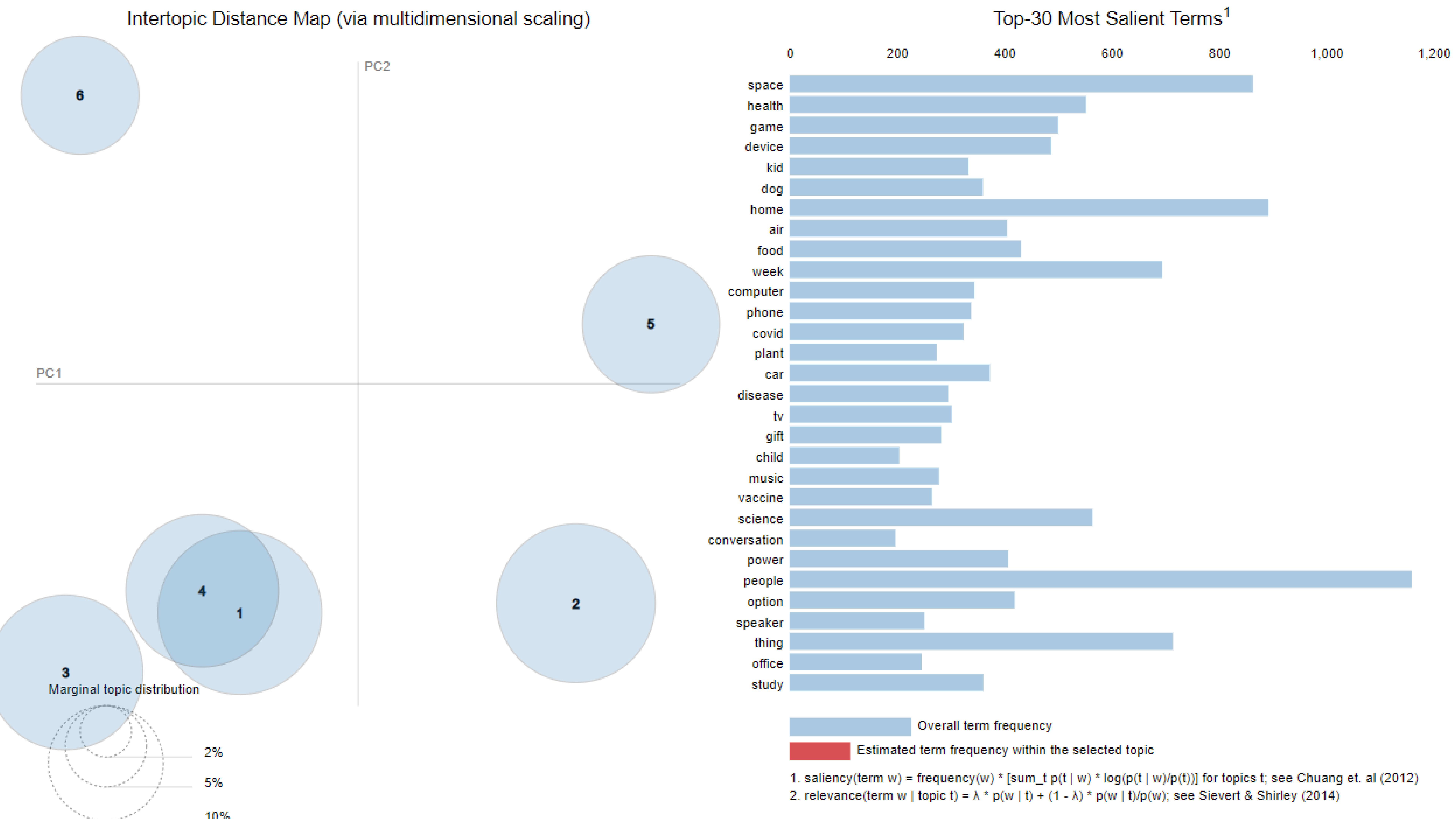
'0.033*"space" +
0.021*"year" +
0.014*"mission" +
0.011*"day" +
0.010*"way" +
0.010*"planet" +
0.009*"scientist" +
0.009*"night" +
0.009*"camera" +
0.008*"science"'

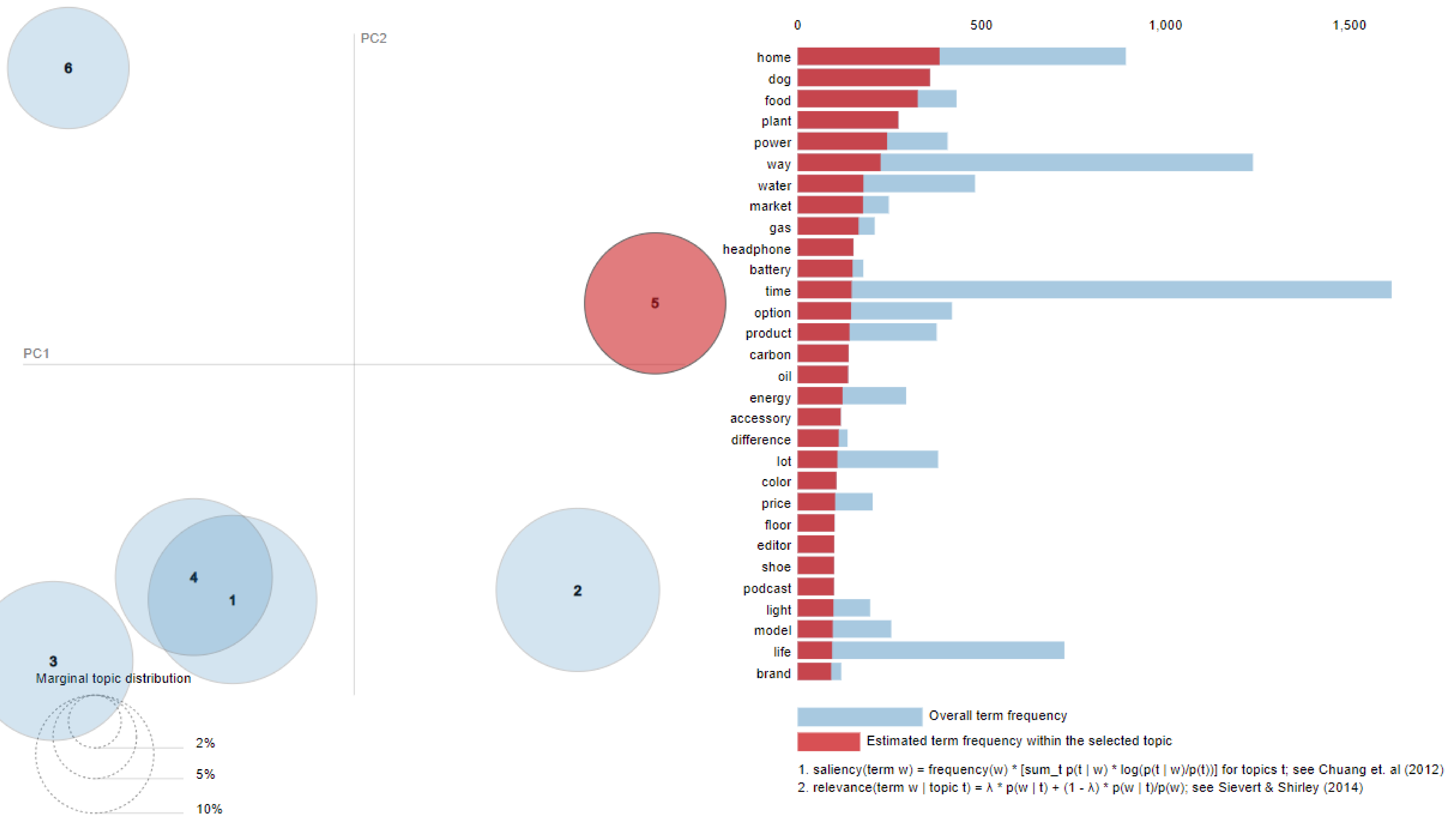
Energy



Space

Health, Covid


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
6
5
4
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
200
400
600
800
1,000
1,200
space
health
game
device
kid
dog
home
air
food
week
computer
phone
covid
plant
car
disease
tv
gift
child
music
vaccine
science
conversation
power
people
option
speaker
thing
office
study
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


PC2
PC1
6
5
4
1
2
3
Marginal topic distribution
2%
5%
10%
0
500
1,000
1,500
home
dog
food
plant
power
way
water
market
gas
headphone
battery
time
option
product
carbon
oil
energy
accessory
difference
lot
color
price
floor
editor
shoe
podcast
light
model
life
brand
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0	1	time, way, covid, computer, people, year, worl...	887.0	0.0923
1	5	time, space, day, home, option, hand, muscle, ...	1237.0	0.1288
2	6	year, gift, way, power, device, home, time, fo...	1369.0	0.1425
3	6	year, gift, way, power, device, home, time, fo...	1125.0	0.1171
4	6	year, gift, way, power, device, home, time, fo...	480.0	0.0500
5	3	people, year, virus, disease, researcher, baby...	1215.0	0.1265
6	2	year, speaker, people, product, home, way, hea...	1453.0	0.1513
7	4	plastic, bacteria, way, project, time, food, l...	998.0	0.1039
8	8	game, dog, thing, story, home, week, fact, sci...	841.0	0.0876

Conclusion

1. Scraping popular science articles and analyze it with LDA algorithm

2. Find the optimal number of topics using coherence scores

3. The best-suited value for the number of topics i.e. k comes out to be in the range of 5 and 6 for scientific news articles.

4. Visualize the topics using pyLDAvis and wordcloud

THANKS FOR LISTENING