Topic Modelling and Named Entity Recognition Results

Avinash Navlani

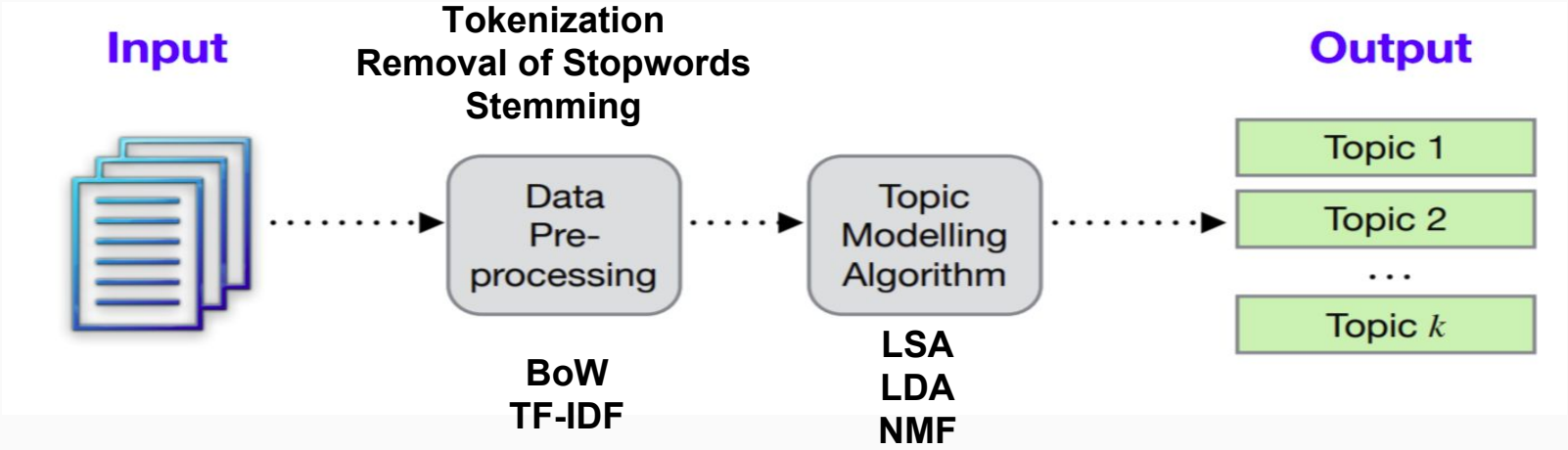
Session Highlights

1. About Problem
2. Topic modelling
3. Discovered Topics in Given Articles
4. Top Keywords of Discovered Topic
5. NER Extracted Entities
6. NER Grouped Extracted Entities
7. Summary

About Problem

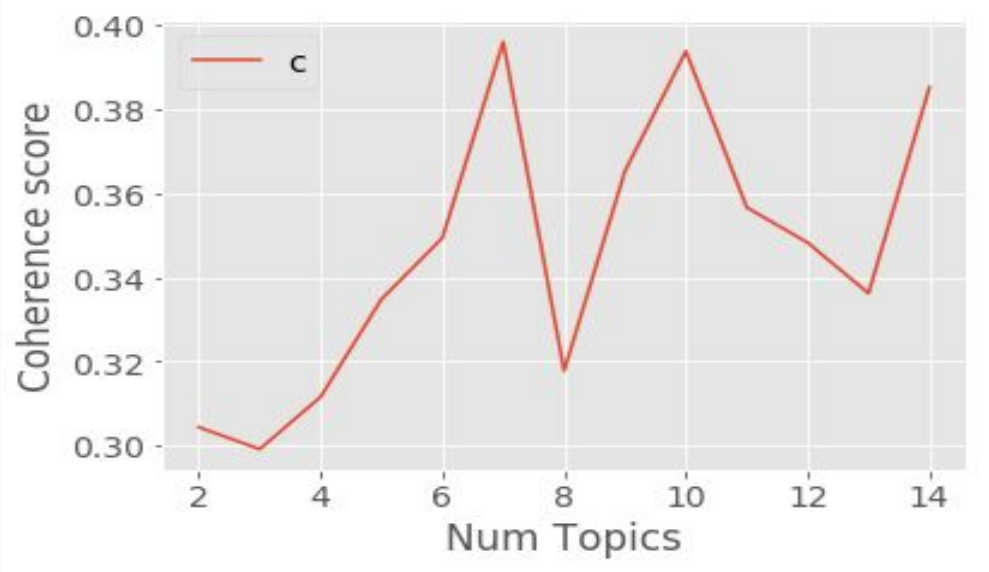
- Our dataset consist news articles of 2016.
- Dataset consists a text file of 4,551 news articles.

Approach for Topic Modelling



Topic Modelling

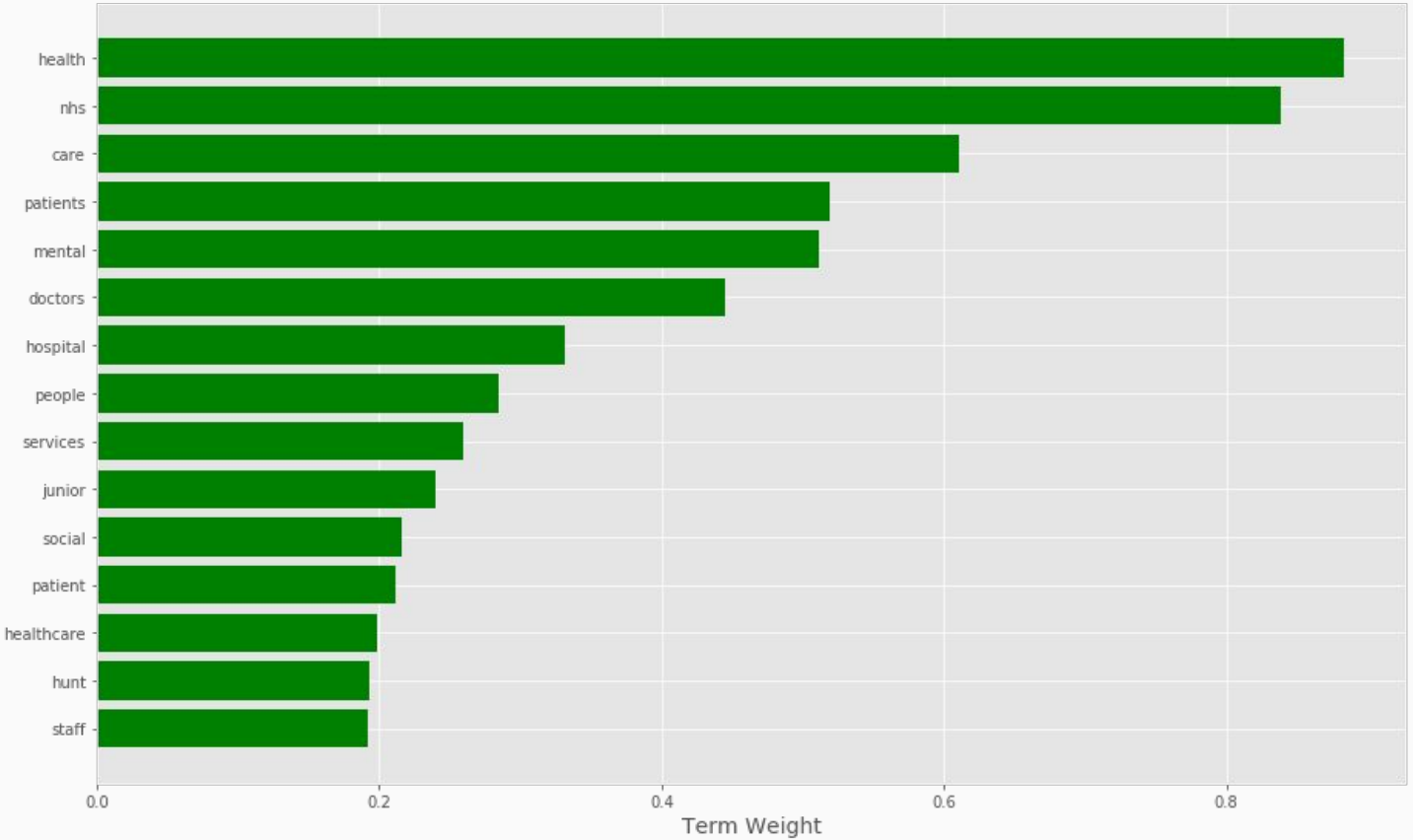
- Topic Modelling aims to automatically discover the hidden thematic structure in a large corpus of text documents.
- Here in our problem, first we found Number of optimal topics using Coherence Score, which is 7.



Discovered Topics in Given Articles

- **‘Topic-1':** eu, uk, brexit, britain, european, leave, europe, vote, referendum, trade (**Brexit**)
- **'Topic-2':** trump, clinton, donald, republican, campaign, president, hillary, cruz, sanders, presidential (**US Presidential Elections**)
- '**Topic-3':** film, films, movie, star, director, hollywood, actor, story, drama, cinema (**Movies**)
- **'Topic-4'**: league, season, leicester, goal, premier, united, city, liverpool, game, ball (**English Premier League**)
- **'Topic-5':** bank, banks, banking, financial, rbs, customers, shares, deutsche, barclays, lloyds (**BFSI**)
- '**Topic-6':** health, nhs, care, patients, mental, doctors, hospital, people, services, junior (**UK National Health Service**)
- **'Topic-7'**: album, music, band, song, pop, songs, rock, love, sound, bowie (**Music**)

Top Keywords of “UK National Health Service” Topics



Named Entity Recognition

"How big is Hillary Clinton's lead in the presidential race? It depends on the poll Democrat ic candidate Hillary Clinton now has an 11-percentage-point lead over her Republican opponen t Donald Trump, according to a poll released by PRRI and the Atlantic on Tuesday. If that we ren't already reason enough for Trump supporters to worry, a poll from NBC and the Wall Stre et Journal released on Monday put Clinton's lead at 14 percentage points. But why the differ ence in numbers? If you want to follow polls in the 28 remaining days before the US votes, I strongly recommend you ignore the date that the poll was published – and focus instead on th e dates that the poll was conducted. That PRRI/Atlantic poll was based on landline and cellp hone interviews that took place on 5-9 October while the data for the NBC/WSJ poll was gathe red on 8-9 October. Those dates are potentially significant given that on 8 October, a 2005 recording was released of Trump saying that, thanks to his fame, he was able to grab women "by the pussy". It's highly likely that a larger proportion of respondents were interviewed after the Trump recording was made public in the NBC/WSJ poll compared with the PRRI/Atlanti c poll. That could mean a 14-percentage-point lead is a more accurate indication of Clinto n's current position in the race. But the crucial question is whether Clinton's lead is temp orary or permanent. We'll need to keep an eye on numbers in the days ahead to understand tha t. In the meantime, though, it's worth looking beyond the horserace numbers that appear at t he top of the survey and digging a little further. In the PRRI/Atlantic poll, I was curious about a question that provided the statement: "These days society seems to punish men just f or acting like men" – 36% of respondents agreed. Another 41% agreed with the statement: "Soc iety as a whole has become too soft and feminine." Those attitudes could provide useful info rmation for understanding why voters might support their respective candidates."

News Article-US Presidential Elections

```
[("Hillary Clinton's", 'PERSON'),
 ('Democratic', 'NORP'),
 ('Hillary Clinton', 'PERSON'),
 ('Republican', 'NORP'),
 ('Donald Trump', 'PERSON'),
 ('PRRI', 'ORG'),
 ('Atlantic', 'LOC'),
 ('Tuesday', 'DATE'),
 ('Trump', 'NORP'),
 ('NBC', 'ORG'),
 ('the Wall Street Journal', 'ORG'),
 ('Monday', 'DATE'),
 ('Clinton', 'PERSON'),
 ('14', 'CARDINAL'),
 ('the 28 remaining days', 'DATE'),
 ('US', 'GPE'),
 ('PRRI/Atlantic', 'ORG'),
 ('5', 'CARDINAL'),
 ('October', 'DATE'),
 ('NBC', 'ORG'),
 ('WSJ', 'ORG'),
 ('8-9 October', 'DATE'),
 ('8 October', 'DATE'),
 ('2005', 'DATE'),
 ('Trump', 'ORG'),
 ('Trump', 'NORP'),
 ('NBC', 'ORG'),
 ('WSJ', 'ORG'),
 ('PRRI/Atlantic', 'ORG'),
```

Extracted Entities

Named Entity Recognition-Grouped Extracted Entities(Conti...)

```
[('GPE', ['US']),
 ('DATE',
  ['2005',
   '8-9 October',
   'the 28 remaining days',
   'Monday',
   'the days ahead',
   'These days',
   '8 October',
   'October',
   'Tuesday']),
 ('PERCENT', ['" – 36%', 'Another 41%']),
 ('ORG',
  ['PRRI/Atlantic', 'Trump', 'WSJ', 'the Wall Street Journal', 'NBC', 'PRRI']),
 ('CARDINAL', ['5', '14']),
 ('NORP', ['Republican', 'Democratic', 'Trump']),
 ('PERSON',
  ['Clinton', 'Donald Trump', 'Hillary Clinton', "Hillary Clinton's"]),
 ('LOC', ['Atlantic'])]
```

Grouped Extracted Entities

Summary

- Give dataset preprocessed and feature generated using BoW and TF-IDF.
- Topic Modelling modelling performed using LSA,LDA, and NMF.
- NMF typically more scalable than LDA, but running times can increase considerably as number of topics k increases.
- Topic Coherence is used for identify number of topics.
- Topic models reflect the structure of the data available. Best used carefully as an exploratory tool to aid human interpretation.
- Pretend entities were extracted from a US Presidential Elections article using spacy Entity Recognition.
- Combination of both Topic modelling will help us summarize document and generate tags about the document.
- Further exploration of text requires sentiment analysis of document, which gives us polarity of text/article/document.

Thank You


video
networking
social
media
process
conversations
content
share
authority
building
seo
production
instant
like
forum
collaboration
always
profile
research
editing
audience
marketing