Seminar 3
Introduction to topic modelling

Mikhail Kamrotov

Data Analysis in Politics and Journalism

Winter /Spring 2019

What is topic modelling

- Automatically identifying major themes in a text, usually by identifying informative words.
- Two main uses:
  - identifying major topics in unlabeled texts;
  - identify which words are important for text that is labeled for topic

B.S. 💩 Detector

- Chrome extension
- Searches all links on a given webpage for references to unreliable sources,
- Classification:
  - **Fake News:** Sources that fabricate stories out of whole cloth with the intent of pranking the public.
  - **Satire:** Sources that provide humorous commentary on current events in the form of fake news.
  - **Extreme Bias:** Sources that traffic in political propaganda and gross distortions of fact.
  - **Conspiracy Theory:** Sources that are well-known promoters of kooky conspiracy theories.
  - **Rumor Mill:** Sources that traffic in rumors, innuendo, and unverified claims.
  - **State News:** Sources in repressive states operating under government sanction.
  - **Junk Science:** Sources that promote pseudoscience, metaphysics, and other scientifically dubious claims.
  - **Hate Group:** Sources that actively promote racism, and other forms of discrimination.
  - **Clickbait:** Sources that are aimed at generating online advertising revenue and rely on sensationalist headlines or eye-catching pictures.
  - **Proceed With Caution:** Sources that may be reliable but whose contents require further verification.

Dataset

- 244 websites
- 12999 news
- Tagged as “bullshit” by B.S. Detector:
  - "bias"
  - "conspiracy”
  - "fake"
  - "bs" (unlabeled)
  - "satire"
  - "hate"
  - "junksci”
  - "state"

Unsupervised topic modeling

- Words are automatically grouped into topics
- Number of topics is defined manually
- An estimate of how much each topic contributes to each document
- An estimate of how much each word contributes to each topic