Seeded Topic Modeling

Sonam Gupta

PhD Candidate at Harrisburg University

PhD Research Question

- Identify gender bias in doctor-patient communication
  - Is the bias more prominent towards one gender?
  - What do patients think of their physicians? Do they have any preconceived notions toward female physicians?
  - What is the language used to review the physicians?

ACL Conference Paper Submission

- Idea: Understand any gender bias from patients' perspectives for their physicians
- Data: ZocDoc.com patient reviews
- Tools used: Linguistic Inquiry and Word Count (LIWC) and R
- Conclusions (aligned with literature review):
  - More gendered language for female physicians
  - Female physicians are more personable and interpersonal with their patients
  - Numerical ratings don't suffice
  - Patients too should be more aware of such biases

Analytic
Tone
posemo
negemo
female
male
informal
social

Fig 1: LIWC variables

DocName	DocDegree	Location	PatientName	ReviewDate	OverallRating	BedsideMannerRating	WaitTimeRat	Review
A	MD	New York NY	a Verified Patient	April 7 2015	5	5	5	A is amazing ! She made me feel totally comfortable and answered all my questions. I was in and out. No wait time. The medical assista
A	MD	New York NY	a Verified Patient	April 6 2015	5	5	5	I was in a jam and couldn't get to my regular GP. Don't be put off by location. Dr. A was prompt friendly knowledgeable and thorough.
A	MD	New York NY	a Verified Patient	Less than a year ago	5	5	4	Dr. A was great - she took the time to explain my diagnosis and her associate (admin/nurse) was very friendly and helpful as well.
A	MD	New York NY	a Verified Patient	August 5 2014	5	5	4	Very knowledgeable and explained everything. Professional and acted courteously. Didn't feel rushed but was in and out quickly.
A	MD	New York NY	a Verified Patient	May 5 2014	5	5	5	A was very pleasant and easy to deal with. She got straight to the point and I was in and out with a solution quickly.
A	MD	New York NY	a Verified Patient	More than a year ag	1	4	5	First issue is they don't tell you that this is at the bottom of a Duane Reade so I spent 10 minutes trying to figure that out. Second issue

Fig 2: Online Patient Reviews Dataset

Data Pre-Processing

- One of the most challenging part of any ML or NLP pipeline especially with unstructured data
- Basic text pre-processing
  - Remove stop words
  - Tokenization
  - Remove punctuation and unnecessary numbers
  - Create a document term matrix (dtm) – “In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. Each ij cell, then, is the number of times word j occurs in document i. ” [Source: Wikipedia]
  - Document feature matrix, an instance of dtm, that also considers other text properties

Seeded vs Non-seeded Topic Models

- Detailed language analysis of patient reviews for male vs female physicians
- Traditional topic modeling using LDA
  - Gensim library in Python
  - Num of topics = 5
  - Some more parameter tuning
- Supervised topic modeling using LDA
  - Used seed words to guide the topics
  - guidedLDA in Python, seededLDA in R
  - Parameter tuning

Traditional LDA Results

- See the file in repository, lda_patientReviews_5_bigrams.html

Seeded Topic Modeling

- 4 seeded topics: appearance, communication, gendered, descriptive words for men
  - Physical appearance for female physicians
  - Communication skills
  - Use of gendered words
  - Typical words used to describe men

Results are not Perfect, yet!

- More data may help
- Patients' demographics are not clear, that can make the results partial
- More pre-processing to remove noise from the data

Technical Challenges

- Package installations for guidedLDA in Python, seededLDA in R
- Python package not entirely supported
- Model.fit() takes certain type of input ~ data preprocessing