INST728E - MODULE 9

TOPIC MODELING

CODY BUNTAIN
@CODYBUNTAIN
CBUNTAIN@CS.UMD.EDU

MODULE 9
TOPIC MODELING

```
for (i, tokenList) in ldaTopics:
    print "Topic %d:" % i, ' '.join([pair[0] for pair in tokenList])
```

```
Topic 16: #brussels brussels metro airport station explosion #prayforbrus
sels eu amp reports
Topic 3: brussels #brussels airport explosions belgium #brusselsairport #
brusselsattack safe facebook world
Topic 2: brussels #brussels amp airport explosions thoughts people explos
ion belgian concerned
Topic 10: #brussels brussels people explosions airport attacks #prayforbe
lgium metro thoughts #prayforbrussels
Topic 11: #brussels airport explosions #bruxelles brussels explosiones br
uselas metro gt aeropuerto
Topic 15: #brussels brussels airport attacks metro explosions explosion #
zaventem bbcbreaking dead
Topic 13: airport brussels metro #brussels news anyone explosiones sad lo
ve going
Topic 6: brussels #brussels airport explosion belgium dead breaking metro
reports attack
Topic 7: #brussels brussels explosions airport explosion metro bruxelles
belgian pas breaking
Topic 17: #brussels metro explosion brussels #bruxelles belgium dans thou
ghts bruxelles une
```

```
print "Most common from analyzer:"
for x in fd.most_common(20):
    print x[0], x[1]
```

```
Most common from analyzer:
#brussels 1538
brussels 1411
airport 1155
explosions 854
metro 612
explosion 431
people 304
attacks 298
station 280
belgium 268
explosiones 244
#zaventem 241
#bruxelles 230
belgian 219
breaking 204
attack 203
bruselas 200
thoughts 184
injured 184
#brusselsattack 176
```

WHAT IS A TOPIC?

- A category/secroty in a newspaper or Netflix
- A hashtag
- A collection of related words
  - “star wars”, “barack obama”, “donald trump”, etc.

FINDING TOPICS

- Human-selected categories or keywords
  - A human annotator assigns messages to categories (or “topics”)
- Or automated algorithms, for example:
  - Latent Dirichlet Allocation (LDA)
  - Author-Topic Model (ATM)

LDA OVERVIEW

- Topics are hidden, or “latent,” within content
- A topic is a mixture of words
  - E.g., “star” is in both the Star Wars and Astronomy topics
- A message is a weighted collection of topics
  - E.g., a tweet may be about the topic “Immigration” and “Donald Trump”

LDA OVERVIEW

- Called a “generative model” as it can be used to create documents

- Method:
  - Select a random set of topics
  - For each topic, select a random set of words
  - String these words together to create a document

LDA OVERVIEW

- LDA has three main parameters
  - k - number of topics
  - alpha - document concentration, how many topics appear in a given document
    - Lower means fewer topics, higher means more
  - beta - topic-word concentration, how many words appear in a given topic

ATM OVERVIEW

- Very similar to LDA
- Includes authors as an additional indicator
- An author tends to tweet about just a few topics

MODULE 9 HOMEWORK

FILL OUT AND SUBMIT MODULE NOTEBOOK

- Run LDA and ATM on your relevant tweet data for three values of K
    - Compare the topics returned by LDA and ATM
- Select a number of topics (k) for your relevant data, and justify your choice

JUPYTER NOTEBOOK

EXAMPLE