40.016 The Analytics Edge

Text Analytics (Part 1)

Stefano Galelli

SUTD

Term 6, 2020

Outline

- Text Analytics
- Sentiment Analysis
- Sentiment Analysis with Twitter data
- Modelling process

Text Analytics

- Process of (automatically) deriving high-quality information from text
- Common tasks are (1) text categorization and summarization, and (2) sentiment analysis
- Text analytics build on several steps (e.g., processing the text, finding patterns, learning a classification model)

Sentiment Analysis

Tasks: it can be seen as a classification problem, where one wants to determine the

- *Polarity* (e.g., positive, negative, neutral), or
- *Emotional states* (e.g., angry, sad)

of a given text.

Sentiment Analysis

Typical applications:

- Political sentiment
- Opinion polling
- Recommendation systems

Sentiment Analysis with Twitter data

Where can we get data? Some options:

1. Twitter's API
2. Specialized websites, such as sentiment140
3. R package `TwitterR`

Sentiment Analysis with Twitter data

A **fundamental piece of information** we need are the labels (sentiments) associated to each tweet. Where do we get them? Some options:

1. Manual labelling
2. Centralized work places, such as Amazon Mechanical Turk
3. Leverage the information contained in emoticons

Sentiment Analysis with Twitter data

Challenges:

- Poor spelling and non-traditional grammar
  - Example: *"U say that iphone 5S didnt bring anything new 2?"*

Sentiment Analysis with Twitter data

Challenges:

- Poor spelling and non-traditional grammar
  - Example: *"U say that iphone 5S didnt bring anything new 2?"*
- Ambiguity of english language
  - Example: *"John and Mary took two trips around France. They were both wonderful."*
  - Example: *"Medicine helps dog bite victims."*

Modelling process

- This is a **classification problem**, where we want to predict the sentiment associated to a tweet *a*

Modelling process

- We first need to transform each tweet into a numerical representation in the form of a vector, known in this field as **Document-Term Matrix (DTM)**

Modelling process

- We first need to transform each tweet into a numerical representation in the form of a vector, known in this field as **Document-Term Matrix (DTM)**. Example:
- "John likes to watch movies. Mary likes movies too."
- "John also likes to watch football."

Modelling process

Modelling workflow

- Pre-processing
  1. Convert text to lower case
  2. Remove stopwords
  3. Remove punctuation
  4. Stemming
  5. Create DTM
  6. Removing sparse terms
- Preparing the DTM for model learning
- Train and test a classifier

References

- Teaching notes.