Summarize It

Goal: Generate a summary of text documents which are as close to natural language as possible

Two Stages: Feature Extraction + Feature Abstraction

What we couldn’t do?

- Feature Abstraction: Generate Natural Language summaries

What we have done?

- Feature Generation: Extract a bunch of keywords which hint about the text

Which methods we made work? (a little)

- Un-supervised Machine Learning:
  - Latent Dirichlet Allocation – Extract some representative words from the text
  - TextRank Algorithm – Rank those representative words
  - Result: 30 % - 40% match between the bag-of-words for provided catch-phrases and our summary (Not good ☹)

Which methods we couldn’t work-out?

- Supervised Machine Learning:
  - Support Vector Machine

How can we Improve?

- Try tuning the parameters for LDA
- Try some grid search features for classification algorithms.

What we learned?

- Topic modeling
- Latent Dirichlet Allocation algorithm
- TextRank algorithm

Thank You!