Importance of Unigram Models for Text Retrieval and Analysis

- Words are meaningful units designed by humans and often sufficient for retrieval and analysis tasks
- Difficulty in moving toward more complex models
  - They involve more parameters, so need more data to estimate (A doc is an extremely small sample)
  - They increase the computational complexity significantly, both in time and space
- Capturing word order or structure may not add so much value for “topical inference”, though using more sophisticated models can still be expected to improve performance
- It’s often easy to extend a method using a unigram LM to using an n-gram LM