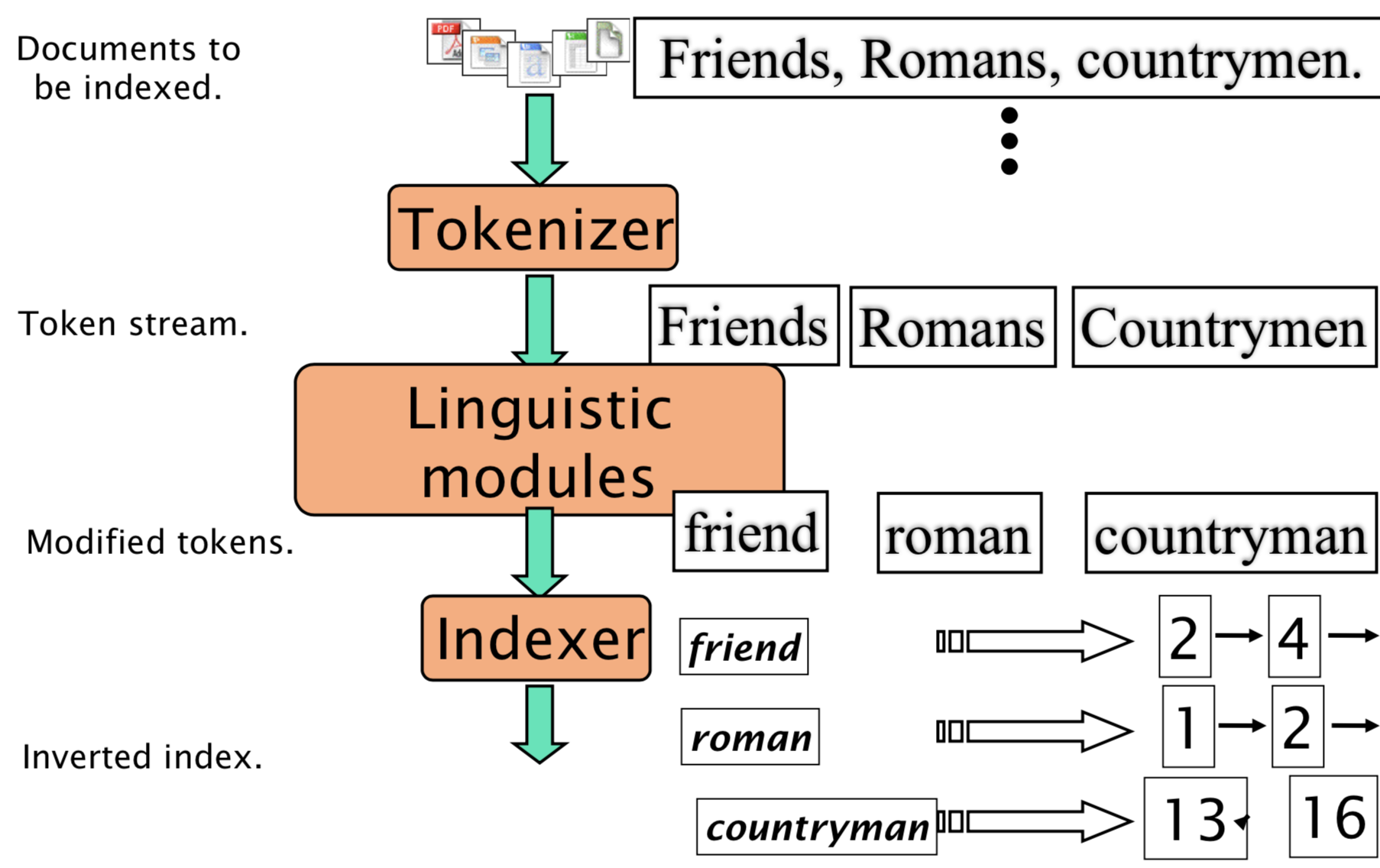
* Inverted Index:
  + Why inverted?
  + Why sorted?
  + How?
* Parsing, tokenizing text, stemming, normalization, ...
  + Parsing: Token -> Tree
  + Tokenizing
    - How many tokens exit in a sentence/document?
    - Why it is hard? I visited San Francisco
  + Normalization:
    - U.S.A. == USA?
    - Windows = window?
  + Lemmatization
    - am, are, is → be
  + Stemming
    - automate, automatic, automation → automat
    - Algorithm: Porter stemmer
* Basic Indexing Pipeline:



* Data structure for looking up a term
  + Hashes vs. trees: Criteria? Pros and Cons?
  + Phrase queries
    - Bigram index: false positive?
    - Positional index:
      * size?
      * Proximity index as well
      * Rules of thumb
* Heap’s Law
  + M = KTb, what does it mean?
  + M is decreased by?
* Zipf’s Law
  + How to explain it generally? A few words occur very often, and many words are infrequent.
  + Why do we observe Zipf’s law?
* Vector Space Ranking
  + Cosine similarity/distance
  + TF-IDF
    - Why log? Why plus 1?
* Tiered Indexed
  + Why?
  + How?
  + Title vs. rest of document
* BM25
  + Formula
  + What is b used for?



* Anchor Text
  + Why it is important?

*“As long as backlinks are an important SEO ranking factor, anchor text is going to play a significant role. It's added for the benefit of search engines, so that they can determine what the “linked-to page” is about. Before the 2012 Penguin update, Google used anchor texts to understand whether or not a particular website was relevant. Nowadays, anchor texts serve as a great way for the search engine to penalize websites for spam and over-optimization.”*

* + How you can update your index to include it?
* PageRank
  + Calculation
  + Teleporting (why? How?)
  + Topic-Sensetive PageRank
* Hubs-and-Authorities
* Topic-Sensetive PageRank Evaluation -- precision, recall, f, precision@k, ndcg
  + Calculation