https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

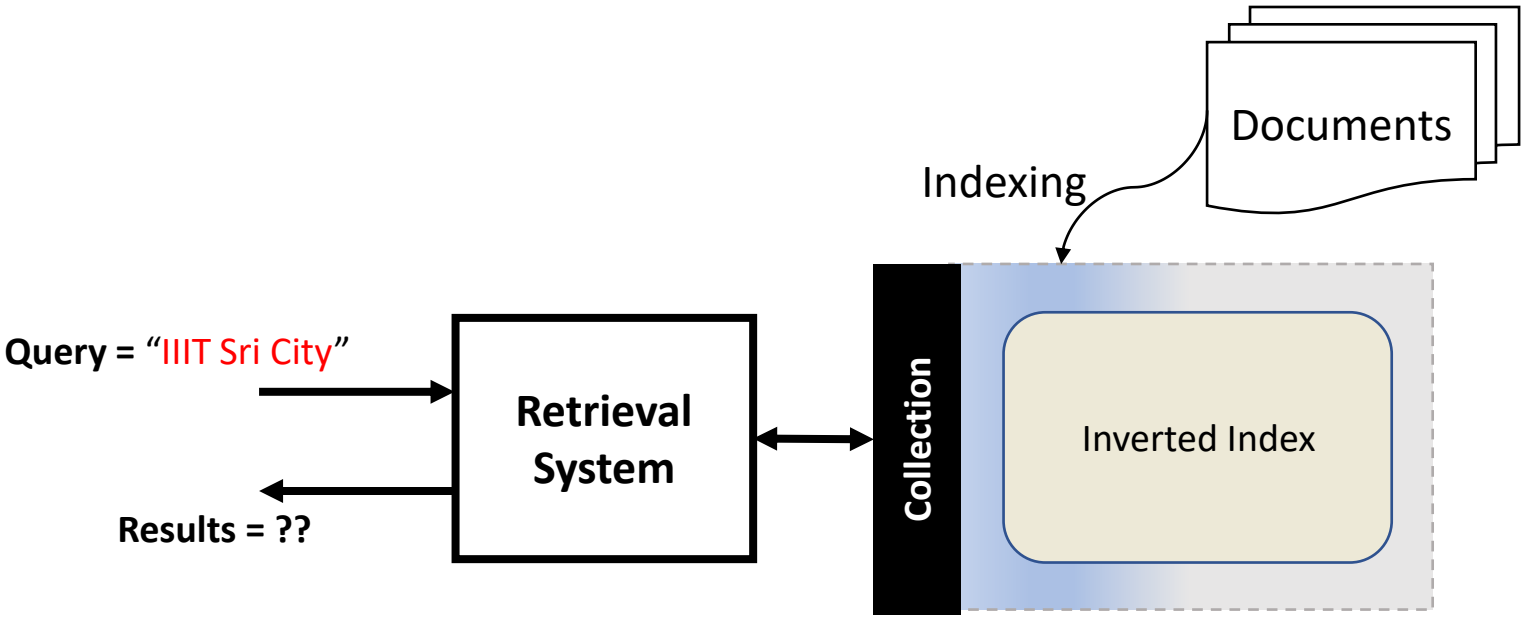
Term: Aug – Sep, 2019
Chennai Mathematical Institute

அட பாடல் போல தேடல் கூட ஒரு சுகமே
Search, like a song, is also a joy.

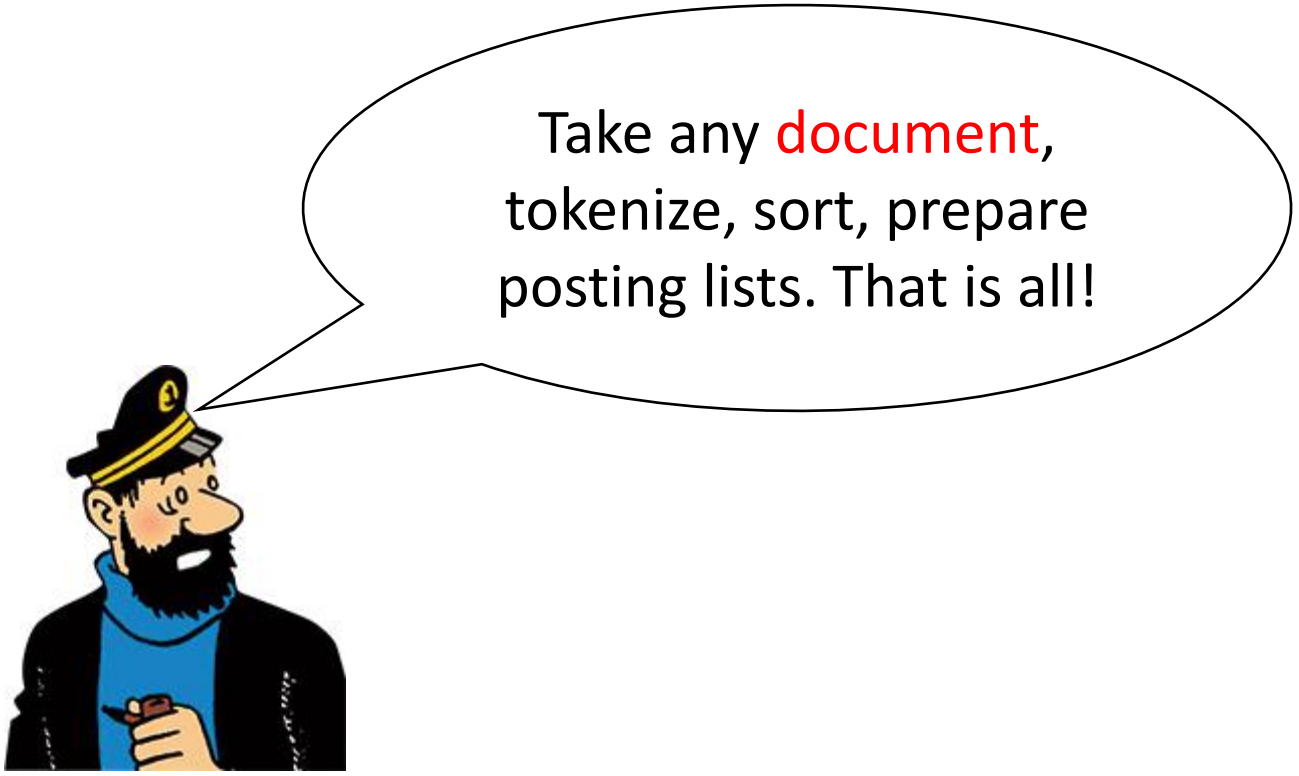
- From the movie, Thulladha Manamum Thullum. Lyrics by Vaali.

Indexing

The Big Picture



How to Index?



Captain Haddock

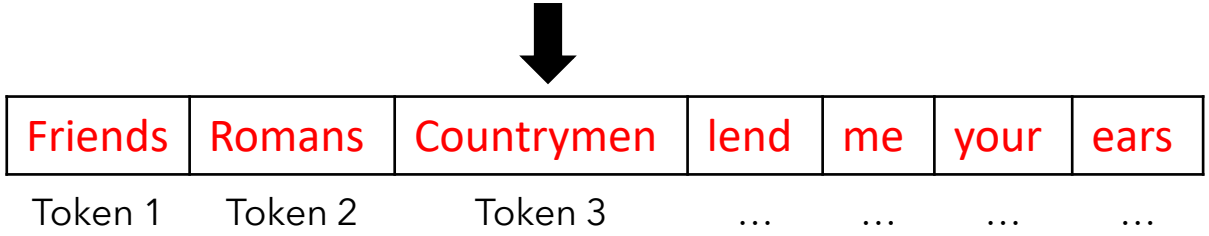
What is a Document?

- Some systems store a single email in multiple files. Is each file a document?
- Some files can contain multiple documents (as in XML, Zip).

Tokens Vs. Terms

- Tokens
  - A ***Token*** is a sequence of characters that make a semantic unit

Friends, Romans, Countrymen, lend me your ears.



- Terms
  - Indexed by the IR system
  - Throws away “less important” tokens that we do not expect in the query

Quiz

- Tokenize O’Neil Can’t study.

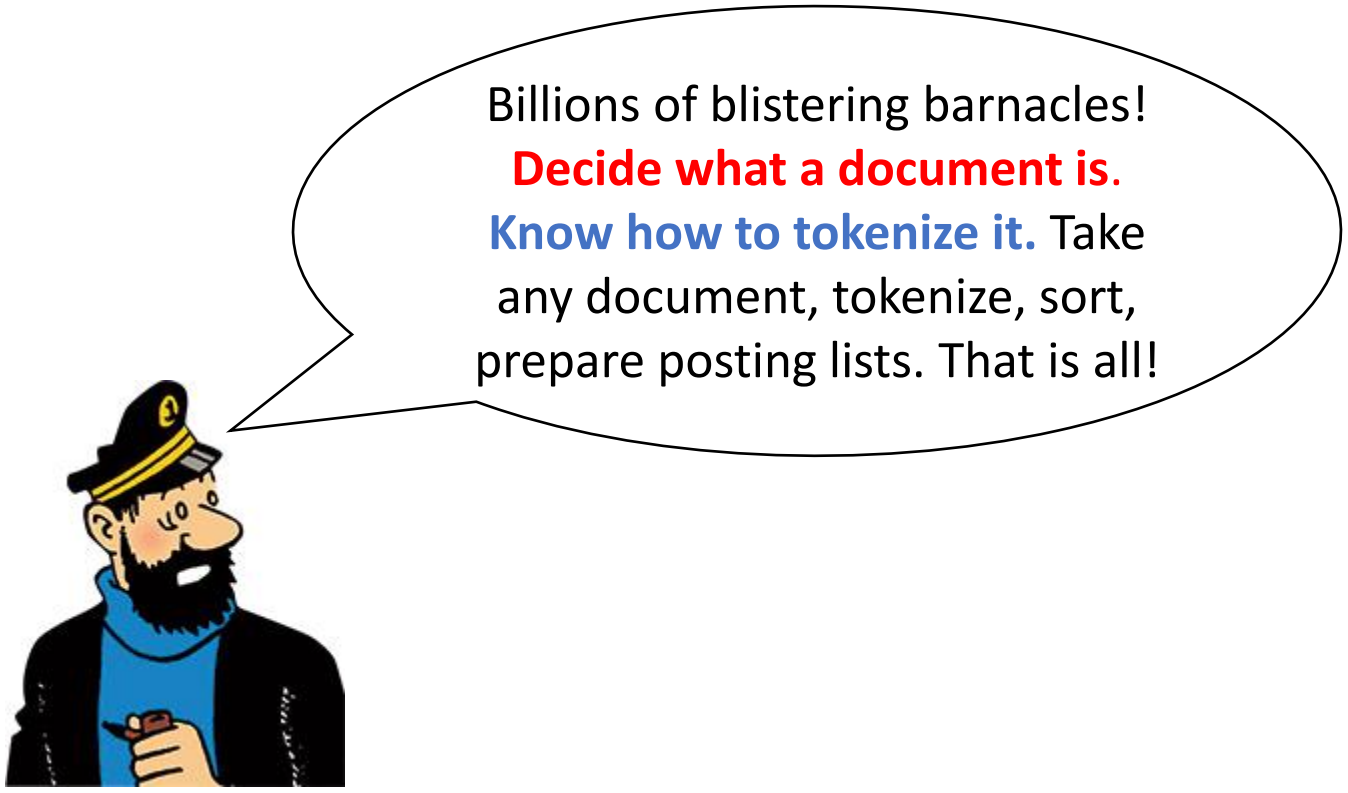
O’Neil	Can’t	study

What if we tokenize based on ‘ ?

O	Neil	Can	t	study

O	Neil	Can’t	study

How to Index?


Billions of blistering barnacles!
Decide what a document is.
Know how to tokenize it. Take any document, tokenize, sort, prepare posting lists. That is all!

Which Tokens to Index?

- Which tokens are interesting?

It is difficult to imagine living without search engines

Stop Word Removal

difficult imagine living search engines

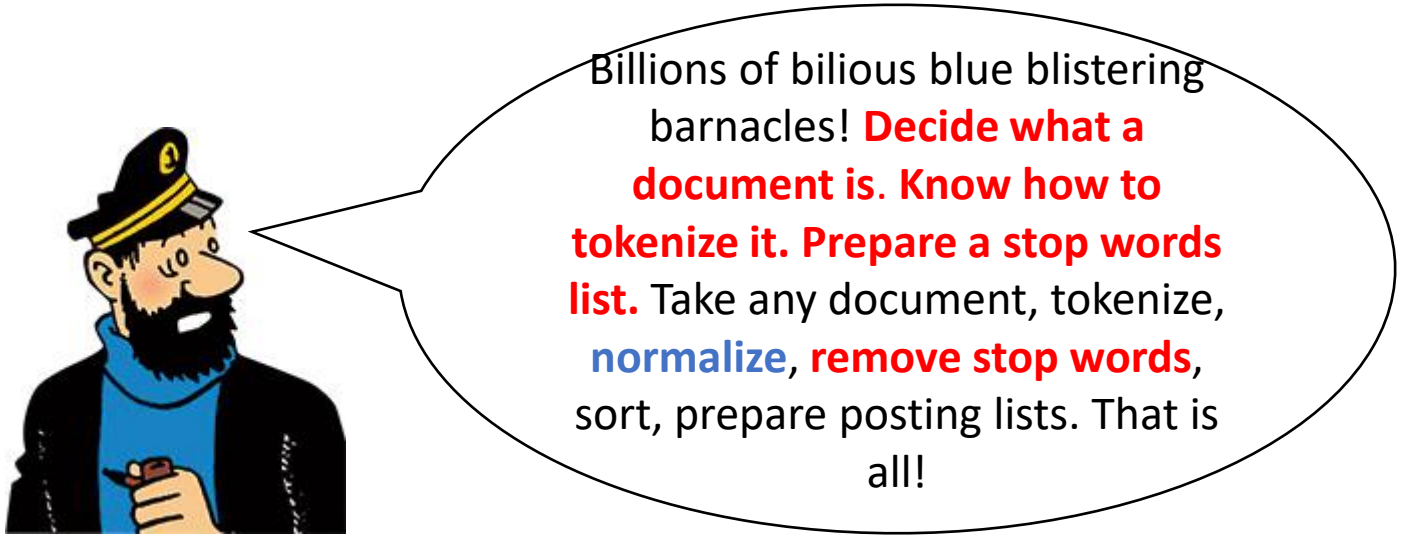
- it, is, to, without are “Stop Words” for us here.

How to Index?


Billions of blue blistering barnacles! **Decide what a document is**. **Know how to tokenize it.** **Prepare a stop words list.** Take any document, tokenize, **remove stop words**, sort, prepare posting lists. That is all!

Token Normalization

- Equivalence Classes
  - (case folding) window, windows, Windows, Window → window
  - anti-theft, antitheft, anti theft → antitheft
  - color, colour → color



Normalization Challenges

- We lose the meaning if we normalize incorrectly:
  - C.A.T is not cat
  - Bush may be a person name. Need to be careful with proper nouns.
- Is TrueCasing a potential solution?
  - TrueCasing
    - Convert words at beginning of a sentence to lowercase.
    - Leave the rest capitalized.
- Usually, we lowercase everything.

Stemming and Lemmatization

- Stemming (chop the ends)
  - going → go, analysis → analys (Need not result in a dictionary word)
- Lemmatization
  - Return the dictionary form of the root word (lemma)
    - saw → see.
- More Examples
  - am, are, is → be
  - car, cars, car's, cars' → car
  - democrat, democratic, democracy, democratization → democrat

Porter Stemmer

- Multiple phases of rule-based refinement

word measure

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat
(m > 1) EMENT →	replacement → replac (does not apply to **cement**)

Stemming Examples

Stemmer	Text
Porter	Such an analysis can reveal features that are not easil visible from the variations in the individual genes and can lead to a picture of expression that is more
Lovins	such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor
Paice	such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor

Issues in Stemming

- Stemmers are not perfect!
- Overstemming
  - Too many characters are cut off from the word
  - Example: university, universal → univers
- Understemming
  - Example: data → dat, datum → datu. Ideally, we would like the result to be the same for both.

How to Index?

Billions of bilious blue blistering barnacles! **Decide what a document is. Know how to tokenize it. Prepare a stop words list.** Take any document, tokenize, **normalize**, **remove stop words**, **stem/lemmatize**, sort, prepare posting lists. That is all!

Quiz

- Can you tokenize the following?
  - **반갑습니다**
    - (Korean for "*Nice to meet you*")
  - **Bundesausbildungsförderungsgesetz**
    - A German compound word for "*Federal Education and Training Act*")
- Can you think of a case where splitting with white space is bad?
  - Los Angeles, New Delhi, IT Park

Thank You