Machine Learning: Lab+PHW 3

Won Kim & Woong-Kee Loh

2021

Objective – Small Search Engine

- Using a vector space model, find the documents that best match the given query.
- Implement an *inverted index* for fast search.
  - Use tokens (words) in queries instead of documents.
  - Remove stop words, punctuation marks, frequent tokens, etc.
- Determine how many documents are returned by specifying
  - Number k (≥1) of documents, or
  - Similarity ϵ ($-1 \le \epsilon \le 1$) to the query.
- Compute precision and recall for each query.
  - Compare precision/recall for various parameter values such as IDF ratio, k, ϵ, etc.
  - Find the best parameters.

Dataset – Cranfield Collection (1/2)

- Four files
- cran.all.1400 – 1,400 documents
  - .I (id), .T (short description), .A (author), .B (book), .W (document content) fields
- cran.qry – 225 queries
  - .I (id), .W (query content) fields
- cranqrel – relevant documents for each query
  - Series of (query order, relevant doc id, relevance)
  - NOTE: first field (query order) is the order of appearance of the query in cran.qry, NOT the id of the query
- cranqrel.readme – description on cranqrel
- https://github.com/topics/cranfield-collection

Dataset – Cranfield Collection (2/2)

- cran.all.1400

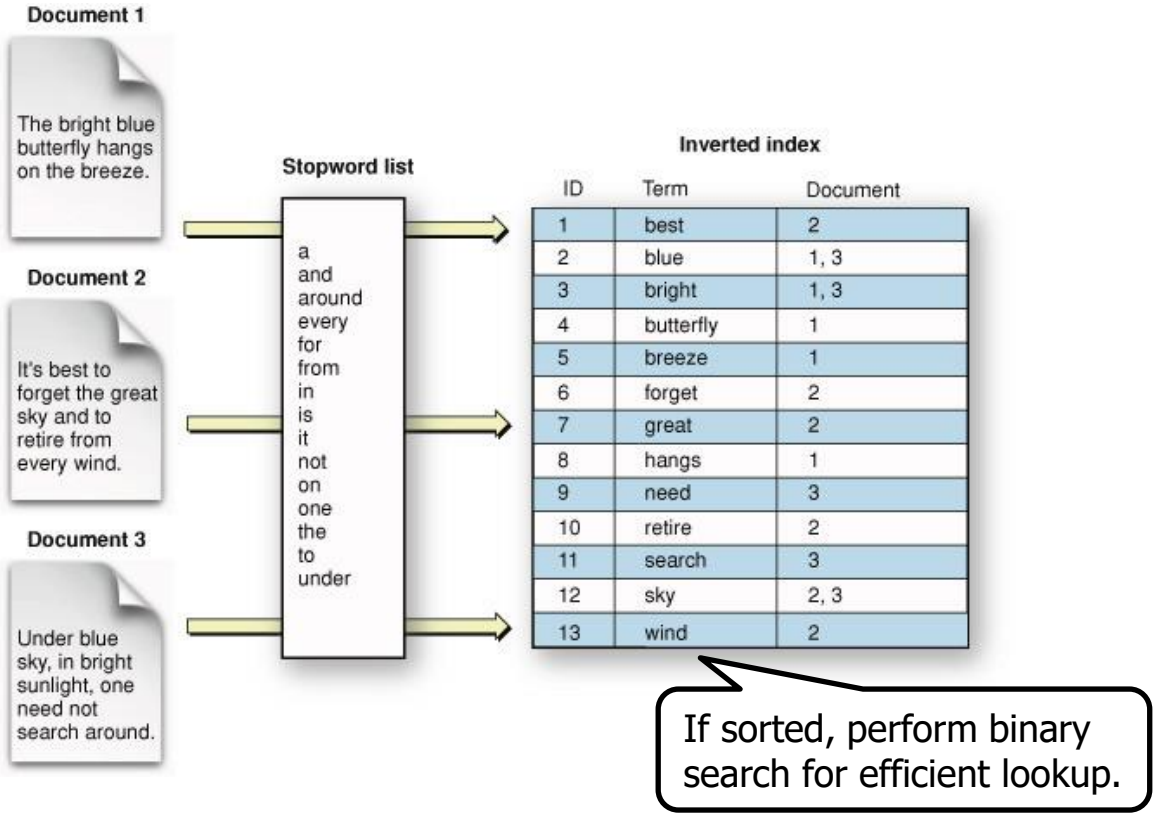
.I 1
.T
experimental investigation of the aerodynamics of a
wing in a slipstream .
.A
brenckman,m.
.B
j. ae. scs. 25, 1958, 324.
.W
experimental investigation of the aerodynamics of a
wing in a slipstream .
an experimental study of a wing in a propeller slipstre
made in order to determine the spanwise distribution
increase due to slipstream at different angles of attack
and at different free stream to slipstream velocity ratio
results were intended in part as an evaluation basis for
theoretical treatments of this problem .
the comparative span loading curves, together with
supporting evidence, showed that a substantial part of
produced by the slipstream was due to a /destalling/ o
boundary-layer-control effect . the integrated remainir
increment, after subtracting this destalling lift, was four

- cran.qry

.I 001
.W
what similarity laws must be obeyed when constructing
of heated high speed aircraft .
.I 002
.W
what are the structural and aeroelastic problems associ
of high speed aircraft .
.I 004 ← query order = 3
.W
what problems of heat conduction in composite slabs h
far .
.I 008 ← query order = 4
.W
can a criterion be developed to show empirically the va
solutions for chemically reacting gas mixtures based on
assumption of instantaneous local chemical equilibrium
.I 009 ← query order = 5
.W
what chemical kinetic system is applicable to hypersonic
problems .

Inverted Index (or *Inverted File*)

- An index that maps from content, such as words or numbers, to its locations (documents)
- For a given query, the index is used to quickly find the documents containing the query tokens.

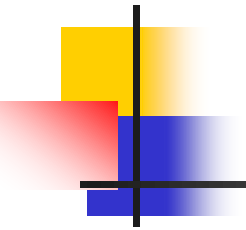


In-class Lab (25pts)

- Time limit: 90 minutes
- Download and examine the dataset.
- Write up a plan for the project.
  - Dataset preprocessing
  - Tokenization, stop word removal, etc.
  - Construction of inverted index
- Write up (in WORD/HWP) the structure of the program to implement the project.
  - Adequate details necessary
- Submit it to CyberCampus.

PHW (100pts)

- This is a team project.
  - The detailed project planning and analysis of result are natural opportunities for all team members to contribute ideas.
- What to submit:
  - Python source codes (.py files ZIPPED)
  - Results – code description, all outputs for selected queries, precision/recall, best parameters (one WORD/HWP file)
- Due: 24:00 Nov. 9 (Wed class) or 10 (Thur class)
- Post any questions on CyberCampus.



End of Lab+PHW Guide