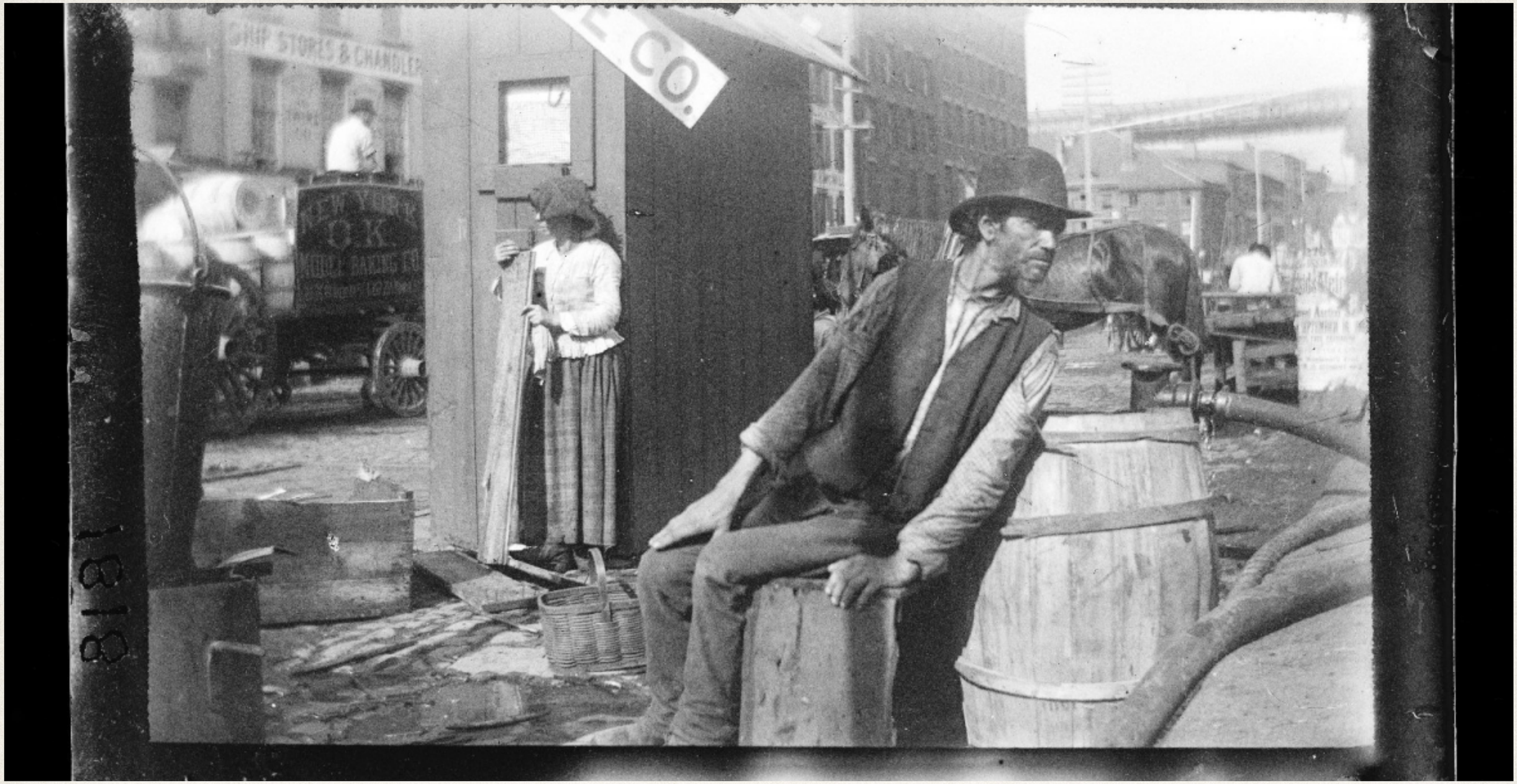

https://commons.wikimedia.org/wiki/File:Longshoreman,_Brooklyn,_ca._1872-1887._(5832938139).jpg

Repeatable Runs for Test Collection Documentation

Ian Soboroff, NIST

SIGIR 2019 OSIRRC Workshop, Paris, France

Data

TREC home

Versions of trec_eval

Ad hoc Test Collections

Web Test Collections

Blog Track

Chemical IR Track

Clinical Decision Support Track

Common Core Track

Confusion Track

Contextual Suggestion Track

Crowdsourcing Track

Dynamic Domain Track

Enterprise Track

Entity Track

Filtering Track

start

TRECipedia: the Digital Library of IR Test Collections

Browse test collections by…

Data type

- Newswire
- Web
- Tweets
- Blog
- Medical Cases
- … other …

Document set

- ClueWeb12
- KBA StreamCorpus
- TREC CDs 4 and 5 (minus CR)
- Tweets2011
- … other …

Conference and year

- TREC 2016
- 1999 (which was TREC 8)

Search task

- adhoc search
- filtering / routing
- question answering
- summarization

Popular test collections

- TREC 8 adhoc
- TREC 2012 web

Most recently updated collections:

- TREC 2004 Robust
- TREC 7 filtering
- trec_2018_core
- trec_2017_core
- trec_4_database_merging
- trec_1_routing
- trec_1_adhoc
- trec_2_routing
- trec_2_adhoc
- trec_2011_medical_records

Links:

- http://www.itl.nist.gov
- http://www.nist.gov
- http://www.commerce.gov

- http://trec.nist.gov
- http://ir.nist.gov

- Privacy policy, security notice, accessibility statement
- Disclaimer
- Freedom of Information Act

TREC 2004 Robust

Test Collection Metadata	
Conference:	trec_2004
Year:	2004
Document collection:	cd45_minus_cr
Task:	adhoc
Topic set:	robust04
Data type:	newswire
Topics file:	https://trec.nist.gov/data/robust/04.testset.gz
Relevance judgments:	https://trec.nist.gov/data/robust/qrels.robust2004.txt
Overview paper:	https://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf
Evaluation scores:	https://trec.nist.gov/pubs/trec13/appendices/robust.results.html
Top score:	0.333
Median score:	0.2561
Main measure:	map
main_condition:	automatic, title only, 249 topics
Baseline score:	0.2903
Baseline link:	https://github.com/osirrc/anserini-docker/tree/v0.1.1
Baseline notes:	BM25+RM3 model

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evalua- tion methodology emphasizes a system's least effective topics. The most promising approach to improving poorly performing topics is exploiting text collections other than the target collection such as the web.

The 2004 edition of the track used 250 topics and required systems to rank the topics by predicted difficulty. The 250 topics within the test set allowed the stability of evaluation measures that emphasize poorly performing topics to be investigated. A new measure, a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results, shows promise of giving appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

Overview paper:	https://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf
Evaluation scores:	https://trec.nist.gov/pubs/trec13/appendices/robust.results.html
Top score:	0.333
Median score:	0.2561
Main measure:	map
main_condition:	automatic, title only, 249 topics
Baseline score:	0.2903
Baseline link:	https://github.com/osirrc/anserini-docker/tree/v0.1.1
Baseline notes:	BM25+RM3 model

The robust retrieval track explores methods for improving the consistency of retrieval te
retrieval task in the track is a traditional ad hoc retrieval task where the evalua- tion met
most promising approach to improving poorly performing topics is exploiting text collecti

The 2004 edition of the track used 250 topics and required systems to rank the topics b
allowed the stability of evaluation measures that emphasize poorly performing topics to
MAP measure that uses a geometric mean rather than an arithmetic mean to average i
emphasis to poorly performing topics while being more stable at equal topic set sizes

My questions for OSIRRC…

- Can we do better at linking the image to the source code? What kind of reuse do we want to enable?
- Can we document what kind of resources are needed?
- How about pre-indexed collections in a standoff Docker for really big collections like ClueWeb?