Cluster Rank Demo Harness

Information Retrieval

Philip Robinson

https://github.com/probinso/IR-cluster-rank-demo

- Thanks for online resources
  - Brandon Rose
  - Dan Foreman

Simple Workflow

- ↓ Query
- ↓ Relevance
- ↓ Rank
- ↓ Display
- · Select/Restart

Problem

- Often conflates relevance and ranking
- Doesn't diversify results

Hypothesis

Introducing unsupervised clustering can reduce search time by diversifying results, especially with poor queries.

Proposal

- ↓ Query
- ↓ Relevance
- ↓ Cluster ← ⋆ ·
- ↓ Rank
- ↓ Zoom → ⋆ ↑
- · Select

Project

- Easily swap/test algorithmic components
- Flask server serving Json for module visualization

- Current Restrictions
  - Scale of numpy/scipy/gensim (ask if interested)
  - Doesn't reset well

Evaluation

- Select random documents
- Generate query sample from `tfidf distribution`
- Count steps to identifying document

Demo

Next Steps

- Scale to larger corpora
- Collaborate with `grimmscience` to flush out UI
- Abstract out feature extraction as component